From Conceptualization to Metaphysical Reasoning: Frameworks and Benchmarks Towards Generalizable Reasoning

by

Weiqi Wang

A Thesis Submitted to
The Hong Kong University of Science and Technology
in Partial Fulfilment of the Requirements for
the Degree of Doctor of Philosophy
in Computer Science and Engineering

April 2026, Hong Kong SAR, China

From Conceptualization to Metaphysical Reasoning: Frameworks and Benchmarks Towards Generalizable Reasoning

by Weiqi Wang

Department of Computer Science and Engineering

The Hong Kong University of Science and Technology

Abstract

Large language models (LLMs) show strong reasoning across many tasks, yet their reliability can vary when assumptions change, inputs shift distribution, or familiar knowledge must be recomposed in novel ways. This thesis argues that a key organizing principle for improving such generalization is conceptualization: the ability to abstract concrete events and entities into reusable concepts, and to instantiate those concepts in new situations. We develop a unified paradigm in which conceptualization structures how commonsense knowledge is represented, acquired, scaled, modified, and evaluated beyond surface competence.

We first systematize concept-centric methods for LLMs and formalize a conceptualization–instantiation cycle over commonsense knowledge bases (CSKBs) as a lens spanning generation, question answering, and knowledge manipulation. Building on this lens, we introduce approaches that construct and exploit concept-structured event and entity knowledge to improve generative commonsense inference and zero-shot commonsense question answering, showing that concept-level structure can strengthen reasoning without relying solely on model scale. To address limited CSKB coverage, we propose a scalable distillation framework that extracts large volumes of concept-structured knowledge from strong LLMs and uses critic-style filtering to retain plausible, useful knowledge, expanding coverage while preserving quality.

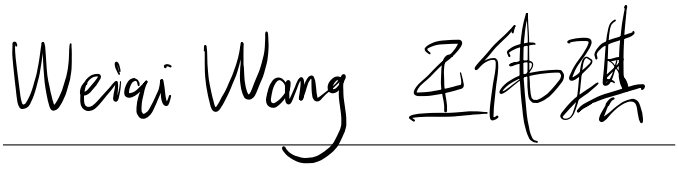
Beyond acquisition, we study controlled knowledge modification via a concept-level editing framework that couples automated plausibility verification with concept-aware rewriting, improving both factuality and downstream utility. Finally, we introduce metaphysical reasoning as a concept-driven stress test: reasoning about improbable or counterfactual changes to conceptualized events. We provide a benchmark that decomposes this challenge into discriminating event, inference, and transition validity under controlled distribution shifts, revealing persistent gaps between apparent competence and deeper conceptual understanding. Together, these contributions advance frameworks, resources, and evaluations that push LLMs toward more robust, generalizable reasoning grounded in concept-level structure.

Authorization

I hereby declare that I am the sole author of the thesis.

I authorize the Hong Kong University of Science and Technology to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize the Hong Kong University of Science and Technology to reproduce the thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.



Weiqi Wang

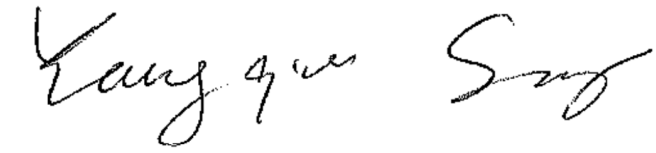
1 April 2026

From Conceptualization to Metaphysical Reasoning: Frameworks and Benchmarks Towards Generalizable Reasoning

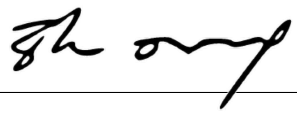
by

Weiqi Wang

This is to certify that I have examined the above PhD thesis
and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by
the thesis examination committee have been made.



Prof. Yangqiu Song, Thesis Supervisor



Prof. Xiaofang Zhou, Head of Department

Department of Computer Science and Engineering

1 April 2026

ACKNOWLEDGEMENTS

Writing this thesis marks the end of a long journey, one that I once doubted I would be able to complete. Looking back to where it all began, I can hardly believe how far I have come. There were moments when the road ahead felt uncertain, and times when I questioned whether I had the strength and clarity to keep going. Yet, step by step, through setbacks and small breakthroughs, this journey gradually became something real, something I could hold on to. Now, as it comes to an end, I feel deeply grateful for having reached this point, proud of how much I have grown, and quietly fulfilled by all that this journey has come to mean. This will always be a period of my life that I will treasure deeply. I feel truly fortunate to have spent these years working on ideas I care deeply about, and I am profoundly thankful to the many people who supported me, challenged me, and believed in me throughout this process.

Along the way, I was also fortunate to spend two wonderful years in Hong Kong and one and a half years in the United States. These years were among the most cheerful and fruitful chapters of my life, filled with discovery, growth, and unforgettable memories. The people and experiences in both places shaped not only this thesis, but also the person I became throughout the journey.

First and foremost, I would like to thank my supervisor, Professor Yangqiu Song, for his guidance and mentorship. We have been collaborating for six years, since my undergraduate days, and he has always been a steady source of encouragement, inspiration, and honest feedback. He consistently supported my research ideas, encouraged me whenever I wanted to explore new directions, and provided the resources and freedom that allowed those ideas to take shape. His patience, high standards, and trust have shaped not only this thesis, but also the way I think about research and about my future as a scientist. I feel extremely fortunate to have grown under his supervision.

I must also acknowledge Professor Daniel Khashabi and Professor Benjamin Van Durme at Johns Hopkins University for their supervision during my six-month visit at JHU. I learned a great deal during this period, both academically and personally, about how to conduct research and, just as importantly, how to enjoy it. It was a truly pleasant and enriching experience that broadened my perspective beyond my studies at HKUST.

I also gratefully acknowledge the support of Dr. Limeng Cui, Dr. Chen Luo, Dr. Qingyu Yin, and Dr. Xin Liu for their mentorship during my two internships with the Amazon Search

Experience Science and Amazon Stores Foundational AI teams. They provided invaluable guidance and helped me grow substantially, broadening my understanding of how AI technologies are reshaping industry and how I might apply my skills to contribute to that progress.

I would also like to thank my thesis committee, Professor Long Chen, Professor Sherry Lixue Cheng, Professor Maggie Wenjie Li, and Professor Raymond Chi Wing Wong, for their time, participation, and valuable feedback on my thesis.

I am grateful to all of my co-authors: Antoine Bosselut, Baixuan Xu, Bin He, Bo Li, Changlong Yu, Chen Luo, Ching Ming Samuel Lau, Chunkit Chan, Chun Yi Louis Bo, Chunyang Li, Dan Yu, Daniel Khashabi, Dongyu Ru, Feihong Lu, Ginny Y. Wong, Hao Liu, Haochen Shi, Haoran Li, Huiru Xiao, Hong Ting Tsang, Hongming Zhang, Hui Liu, Huihao Jing, Jiefu Ou, Jiayang Cheng, Jiayao Zhang, Jiaxin Bai, Jianxin Li, Junfei Cheng, Junxian He, Kelvin Kiu-Wai Tam, Lei Chen, Limeng Cui, Lirong Zhang, Liyu Zhang, Long Chen, Mutian He, Newt Hue-Nam K. Nguyen, Qiang Yang, Qi He, Qi Hu, Qian Li, Qihan Lin, Qing Zong, Qingyun Sun, Qingyu Yin, Quyet V. Do, Rui Wang, Sehyun Choi, Sheikh Muhammad Sarwar, Shibo Hao, Shiqi Gao, Shuiwang Ji, Simon See, Sreyashi Nag, Sze Heng Douglas Kwok, Tao Fan, Tianqing Fang, Tianshi Zheng, Tianyu Cao, Tong Zhang, Tsz Ho Chan, Wei Fan, Weihan Li, Weijia Zhang, Weixing Shen, Wenju Xu, Wenxuan Ding, Xin Liu, Yauwai Yim, Yanming Zhu, Yang Li, Yangyifei Luo, Yangqiu Song, Yifan Gao, Ying Su, Yiran Hu, Yi Wu, Yuqi Yang, Yuxi Sun, Yuxin Jiang, Yue Zhang, Zheng Li, Zheng Zhang, Zerui Huang, Zhaobo Wang, Zhaowei Wang, Zheye Deng, Zihao Wang, and Ziqin Zhu. Thank you all for your hard work, creativity, and support in making so many meaningful projects possible and in bringing these publications to life. The names above are listed in alphabetical order.

I also thank all of my lab mates in the HKUST KnowComp group, who made this journey both productive and joyful through generous discussions, thoughtful feedback, and everyday companionship. Your support, collaboration, and shared enthusiasm for research created an environment in which I could learn quickly, stay motivated, and grow with confidence.

In particular, I wish to express my sincere gratitude to Dr. Tianqing Fang and Dr. Xin Liu. Tianqing was my mentor during my early undergraduate research experiences and introduced me to the fascinating world of NLP. I have learned so much from him since the very beginning of my research journey, and the independence and scientific mindset he helped me develop have been invaluable. Xin has been a long-standing collaborator, involved in nearly all of the major research projects I have led, and has offered countless insightful suggestions. He also opened my path toward becoming an industrial researcher, where many exciting ideas and projects are underway to build large language models that benefit people's lives. I sincerely appreciate all of

his help and support throughout this journey.

Finally, I express my deepest gratitude to my parents, Yakun Song and Wei Wang. No matter what happens, your care and support have been the light in my darkest nights, giving me the courage to pursue my dreams far from home, in Hong Kong, in the United States, and wherever my path may lead. None of my academic achievements or personal growth would have been possible without you. I will always be grateful for your sacrifices and for everything you have done to support me.

Last but not least, I would like to thank the friends who have meant so much to me during my PhD journey. I first thank Yichen Chen for her companionship and emotional support during some of the most difficult times. It has been a wonderful and unexpected friendship, and we share many common characteristics, experiences, and perspectives on life. The joyful moments we have shared, and everything we have gone through together, form an important part of the last two years of my doctoral journey. I hope our friendship continues to grow, and that our paths will cross again soon.

I must also thank Wen Chen, with whom I have shared many joyful moments, who has accompanied me through these years, and who has helped me think through and plan for my future. It has been a genuine pleasure to have you as a close friend during my PhD journey.

I also thank Dr. Yueqi Zhang, a talented researcher from Zhejiang University, for standing with me through difficult moments and for sharing the pressure and responsibility along the way. I am truly grateful for your trust, courage, and support.

Cheng Chen has also been a very dear friend of mine since high school. During my PhD years, we spent a great deal of time together and shared many joyful moments, including countless hours playing games. I am deeply grateful for that long-lasting friendship.

There are many other friends who mean a great deal to me, and although I cannot mention everyone by name, please know that your kindness, companionship, and encouragement have made this journey far richer and far more bearable.

Finally, I would like to thank myself. I am grateful that I kept going, even in moments of doubt, exhaustion, and uncertainty. I am grateful that I did not give up on the questions that mattered to me, and that I found the courage to continue, even when the way forward was unclear. I am also grateful for the failures and detours, because they taught me humility, patience, and how to ask better questions. Whatever this thesis may mean to others, to me it will always stand as proof that I endured, learned, and grew.

And so, with gratitude, I close this chapter. When I began this journey, I thought a thesis was measured by results and pages. Now I know that it is measured by people, by conversations that

change you, and by the quiet encouragement that keeps you moving on days when nothing seems to work. If these years have shaped anything in me, it is a deeper respect for careful thinking, honest uncertainty, and the patience it takes to turn small steps into something real. I will carry these lessons, and the kindness I received, into whatever comes next.

With that, I turn from the people behind this story to the work itself.

Here, this thesis begins.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

To my family,
for the love that carried me,
and the home I always return to.

CHAPTER 1

INTRODUCTION

This thesis studies how to make language models reason more reliably about everyday situations. Modern Large Language Models (LLMs) can generate fluent explanations and achieve strong performance on many benchmarks, including a wide range of commonsense evaluations. Yet, in deployed and open-ended settings, reliability can still be fragile when inputs shift distribution [1], when background assumptions change, or when knowledge must be recomposed in a new context. In this thesis, we study these reliability challenges through the lens of *commonsense reasoning*, the ability to draw plausible inferences about actions, intentions, consequences, and constraints in the world [2]. Commonsense reasoning is not chosen here because it is uniformly hard for today's models, but because it is a precise stress test for three requirements that matter in practice: (i) reusing knowledge across many surface forms, (ii) preserving validity under contextual constraints, and (iii) behaving consistently across closely related situations [3].

A running example: recomposition under context. To ground the discussion, consider a kitchen assistant asked to help with a short plan: *"I have leftover soup. How should I reheat it safely?"* In isolation, the model may correctly answer many sub-questions: *"Microwaves heat liquids," "some containers deform under heat,"* and *"avoid spills."* The difficulty emerges when these pieces must be combined under small contextual changes. For instance, the plan *"pour the soup into a container and microwave it"* may be fine when the container is *ceramic*, but becomes unsafe if the container is described as *thin plastic*. A second minimal change can flip plausibility again: the same user might add, *"I am in a hotel room with no microwave,"* where a previously reasonable step becomes infeasible. A third change can shift the relevant constraints: *"My younger sibling is nearby, so prioritize safety over speed,"* which affects whether the assistant should recommend moving hot liquid across a room, using boiling water, or leaving items unattended. These are not failures of recalling isolated facts. They are failures of *generalization under context*, where the model must preserve what should remain invariant (the reusable regularity) while adapting what must change (the context-specific realization and constraints) [4].

A central claim of this thesis is that a key missing organizing principle behind robust commonsense reasoning is *conceptualization*: the ability to lift concrete instances into reusable con-

cepts, and to ground concepts back into concrete, context-appropriate instances. We develop conceptualization as a unifying lens that connects how commonsense knowledge is represented and acquired, how robustness should be evaluated, how models can be maintained and corrected over time, and how models can reason when situations change.

The introduction is organized into two parts. The first part provides background and motivations: what commonsense knowledge is, why it matters for NLP systems and LLM-era applications, and why conceptualization is a natural interface between knowledge and generalization. The second part describes key gaps and challenges that motivate the technical chapters of this thesis, and summarizes how each contribution addresses one of these gaps.

1.1 Commonsense Knowledge and Commonsense Reasoning

1.1.1 What is Commonsense Knowledge?

Commonsense knowledge refers to information about everyday entities, events, and their typical properties and relations that humans routinely rely on, often without explicitly stating it [5]. It includes entity-centric facts (typical properties, affordances, and uses of objects), as well as event-centric inferential knowledge (what tends to cause what, what people likely intend, what consequences follow, and what actions enable or prevent outcomes). This thesis focuses primarily on *event-level* commonsense: knowledge about actions and situations and the plausible inferences they support [6].

To make this concrete, consider the event *"someone pours a hot liquid into a container."* Commonsense inferences include: (i) likely *preconditions* (a container is available; the liquid is already heated), (ii) likely *effects* (the container becomes hot; spills are possible), (iii) plausible *constraints* (some materials deform under heat; some settings prioritize safety over speed), and (iv) *alternatives* (letting it cool first; transferring with a ladle; using a heat-safe mug). Event-level commonsense is therefore relational and conditional: plausibility depends on who acts, what is acted upon, and under which constraints.

Event-level commonsense is particularly important for two reasons [7]. First, it captures structured regularities about human behavior and the physical and social world, such as prerequisites, side effects, goals, and constraints. Second, it aligns naturally with how LLMs are trained and used today, since many downstream applications require reasoning about sequences of actions and their consequences, rather than retrieving isolated properties of objects. Accordingly, the central objects studied in this thesis are event descriptions and the inferential relations

that connect them.

1.1.2 Why Commonsense Matters for Modern NLP Systems

Commonsense knowledge is a prerequisite for language understanding in realistic settings [8]. In interactive systems, users rarely provide complete information; instead, they expect systems to fill gaps with plausible assumptions. In question answering and dialogue, commonsense supports interpreting intent, resolving ambiguity, and producing responses that align with social norms and physical constraints.

In agentic settings, commonsense becomes even more central. Tool-using assistants, web agents, and long-horizon planners must anticipate consequences, avoid implausible actions, and adapt when the environment changes [9, 10]. A representative failure pattern is brittle recomposition: the system knows many relevant pieces in isolation, but does not reliably integrate them into a coherent, context-valid decision. For example, a model may correctly state *"check whether a store is open"* and *"avoid non-refundable purchases if uncertain,"* yet still propose a plan that commits to a fixed booking before verifying the schedule under a changed constraint (such as a holiday closure or a last-minute transit disruption). As LLMs are increasingly deployed as persistent assistants, commonsense is no longer only a benchmark target; it becomes a safety- and reliability-critical component of model behavior.

1.1.3 Commonsense Knowledge Bases as a Computational Substrate

A common approach to computational commonsense is to canonicalize world knowledge into machine-usable resources. Commonsense Knowledge Bases (CSKBs) represent commonsense knowledge in structured or semi-structured forms, often as tuples or text-based assertions with relation labels [11]. Entity-centric CSKBs focus on properties and associations of objects, while event-centric CSKBs focus on causal, intentional, and effect-related inferences. Event-level CSKBs are especially compatible with LLM-era pipelines: nodes can be expressed in natural language, relations can be verbalized, and the resulting knowledge supports inference patterns that resemble how models are prompted and evaluated.

At the same time, CSKBs expose a core tension that motivates this thesis. On one hand, they provide an explicit substrate for representing and transferring commonsense. On the other hand, their coverage is necessarily incomplete, their assertions are context-dependent, and their instance-level representations can make generalization brittle [12].

A simple toy example illustrates the brittleness. A CSKB might contain an assertion like

"put ice cream in a freezer → ice cream stays cold." The statement is useful at the instance level, but it does not specify which substitutions preserve validity (for example, *medicine*, *soup*, or *electronics* behave differently), nor which contextual constraints matter (short-term cooling versus long-term storage, sealed containers versus open bowls, a working freezer versus a power outage). In modern applications, these unmodeled conditions are often exactly where systems fail: the assistant produces a step that sounds reasonable but violates a constraint that becomes apparent only after grounding the situation.

1.2 From Commonsense Knowledge to Generalizable Structure

The limitations above indicate a missing intermediate representation for commonsense reasoning. Most CSKB-style resources, and many LLM prompting pipelines, treat events as instance-level descriptions: a node corresponds to a particular phrasing of an action or state, and knowledge is attached to that phrasing. This makes coverage inefficient, since semantically similar situations must be stored and retrieved separately. It also makes robustness fragile, since validity often depends on relational structure (who does what to whom, under what constraint), while surface forms vary widely across contexts.

This suggests a natural requirement for reliable commonsense reasoning: models need a way to represent reusable regularities that transfer across paraphrases and substitutions, while preserving the contextual conditions under which an inference remains valid. In other words, we need a representation that can compress families of situations without discarding the structure that inference depends on. Conceptualization provides such an interface. It separates what should generalize, the underlying concept-level regularity, from what must remain specific, the concrete contextual realization that determines whether an inference is appropriate.

Despite its intuitive appeal, conceptualization has been comparatively underdeveloped in practical commonsense pipelines for three reasons. First, abstractions are context-sensitive: an abstraction that is linguistically plausible may be inferentially invalid in a particular setting. Second, supervision is scarce: datasets rarely annotate which abstractions are appropriate and which are not. Third, evaluation is subtle: the quality of conceptualization should be measured by its downstream inferential utility and robustness, not only by surface similarity or ontological neatness. This thesis addresses these challenges by making conceptualization operational as a learned, contextualized process that can support robust inference, scalable acquisition, mainte-

nance, and reasoning under change.

1.3 Conceptualization as the Interface for Generalizable Commonsense

1.3.1 The Lift-and-Ground Loop

At a high level, conceptualization is the process of *lifting* a concrete instance into one or more abstract concepts. Instantiation is the complementary process of *grounding* a concept back into a concrete, context-appropriate instance. Together they form a lift-and-ground loop: lift instances into abstractions to expose reusable regularities, then ground abstractions into new contexts to produce plausible instance-level knowledge and inferences.

Returning to the kitchen example, from *"Dana pours hot soup into a ceramic bowl"* a lift step can abstract the event into something like *"an agent transfers a hot liquid into a container."* This abstraction is reusable across paraphrases and substitutions. A ground step can then instantiate it into a new context, such as *"transfer hot coffee into a heat-safe mug"* or *"pour hot broth into a metal pot."* Crucially, not every grounding is valid: grounding into *thin plastic* under *high heat* may break plausibility, and grounding into *an open cup while walking across a crowded room* may violate a safety constraint. The lift-and-ground loop therefore provides a concrete way to express the thesis-level goal: generalize by sharing concept structure, but remain faithful by grounding under context.

This loop is appealing as an organizing principle because it connects several needs that are otherwise treated separately. First, it provides a mechanism for *generalization*, since concept-level structure can be shared across many surface forms. Second, it offers a path for *scalability*, since families of situations can be represented compactly. Third, it supports *robustness*, since reasoning can operate on stable regularities rather than accidental correlations in specific phrasings. Fourth, it provides a handle for *maintenance*, since updating knowledge at the concept level can propagate systematically to many related instances.

1.3.2 Why Conceptualization is Hard in Practice

Although conceptualization is simple to state, it is not a purely syntactic rewriting step. The validity of an abstraction depends on *context*. For event-level commonsense, this context includes the roles and relations within an event (who does what to whom, and under what constraints),

as well as the intended downstream use (which inference relations the abstraction is expected to support).

A compact example shows the difficulty. The abstraction *"heat food to make it safe to eat"* is often reasonable, but whether it supports a particular inference depends on details such as the type of food, the available tools, and the constraint being optimized. If the goal is *safety*, then reheating may be recommended. If the goal is *preserving texture*, then a different method may be preferable. If no heating tool exists, then the same abstract intent must be grounded into an alternative plan. Thus, conceptualization must satisfy two requirements that are often in tension. It must be *general*, capturing reusable regularities across many instances. It must also be *contextualized*, preserving the conditions under which a particular inference remains plausible. This thesis treats the tension between reusability and contextual validity as a central technical theme, and develops learning frameworks that make conceptualization and instantiation operational under limited supervision, while evaluating them by their effect on downstream reasoning robustness.

1.4 Gaps and Challenges in the LLM Era

The motivation for this thesis comes from several gaps that remain even as language models scale. We organize these challenges around a progression from acquiring commonsense, to applying it robustly, to maintaining it over time, and finally to reasoning when the situation itself changes.

1.4.1 Coverage, Cost, and Bias in Commonsense Acquisition

Commonsense knowledge is vast, long-tailed, and often left implicit in raw text. High-quality human annotation is expensive and slow, and it inevitably covers only a small portion of the space. Information extraction can scale, but it is shaped by reporting bias: what people choose to mention is not proportional to what is typical or likely in the world [13]. LLM-generated knowledge can be cheap and broad, but it can reflect selection biases, inherit blind spots, and produce plausible-sounding yet incorrect assertions without careful quality control.

A more detailed example illustrates why event-level commonsense is especially exposed to these issues. Consider the inference *"if someone reheats soup, the container may get hot."* This is obvious to humans, but often absent from text because it is rarely worth stating. By contrast, text is more likely to mention unusual failures, such as spills or accidents. As a result, an

extraction pipeline may over-represent atypical outcomes and under-represent routine prerequisites and constraints. A second failure mode comes from underspecified context. Text might say *"microwave the soup"* without stating that the bowl must be microwave-safe, or that metal containers are inappropriate, or that the lid should be vented. When such implicit prerequisites are missing, downstream systems may generate plans that are locally coherent but globally unsafe. This thesis addresses the gap by proposing a scalable distillation perspective: use strong LLMs as teachers to propose conceptualized and instantiated knowledge, then apply principled filtering and learning to retain plausibility while expanding coverage.

1.4.2 Knowing is Not Reasoning

Coverage alone is not enough. Even when a model appears to "know" relevant commonsense, it may still fail to apply it correctly in a concrete reasoning context. This failure is often triggered by distribution shift: changing a surface form, swapping entities for atypical ones, or placing an event in a different context can cause the model to rely on spurious associations rather than the intended inference pattern. In other words, commonsense reasoning is not only a knowledge storage problem; it is also a *generalization and robustness* problem [14, 15].

A concrete example is recomposition under a minimal change. Suppose a model can answer: *"Is it safe to microwave soup in a ceramic bowl?"* and *"Is thin plastic safe under high heat?"* Yet when asked for a plan, it may still produce: *"pour soup into the available container and microwave for two minutes"* even after the user specifies *"the only container is thin plastic."* This is a failure to integrate a constraint into the plan, not a failure to state the constraint in isolation. A second example involves paraphrase robustness. A model might correctly reject *"heat soup in a plastic takeout box"* but accept *"warm it in the to-go container"* if it overfits to surface tokens rather than structure. A third example involves goal shift. If the user changes the objective from *speed* to *safety-first*, the assistant should adapt its recommendations (for instance, recommending lower heat, smaller transfers, or additional checks), yet models often produce inconsistent behavior across such closely related prompts. Conceptualization provides a mechanism to address this gap: lift instance-level situations into concept-level structure to encourage stable regularities, then ground those regularities back into context-specific instances to recover the specificity needed for concrete inference. This thesis builds and evaluates this idea in settings where small contextual changes should not produce large behavioral inconsistencies.

1.4.3 Maintenance and Updates for Long-Lived Models

Robustness failures become more consequential in deployment, where models are long-lived and must be corrected repeatedly. As LLMs become larger and more widely deployed, a new requirement becomes central: models must be *maintainable*. They should be correctable when they produce implausible or unsafe commonsense, and updatable when their knowledge becomes stale or incomplete.

Knowledge Editing (KE) provides computationally efficient tools for modifying model behavior without retraining from scratch [16–18]. Yet, commonsense editing is harder than editing isolated facts because commonsense is context-sensitive, the same regularity can manifest in many surface forms, and edits can have cascading effects across related knowledge.

A detailed example highlights the challenge. Suppose a deployed assistant repeatedly suggests microwaving food in an inappropriate container type, and an engineer wants to correct this behavior. A single edit targeted at one phrasing, such as *"do not microwave thin plastic containers"*, may not generalize to paraphrases (*"to-go box"*, *"disposable container"*) or to nearby situations that share the same constraint (*"high heat"* in an oven or air fryer, or *"container deforms"*). Worse, repeated patches can interact. One edit might encourage safe containers for microwaves, while another edit about preventing spills might inadvertently encourage sealing lids tightly, which can conflict with the safety requirement to vent heated liquids. Iterative updates therefore risk knowledge drift, where successive edits subtly conflict with or overwrite one another. This thesis addresses these challenges by integrating conceptualization and instantiation into the editing loop, so that edits are semantically enriched and can generalize across contexts rather than remaining brittle single-surface patches.

1.4.4 Reasoning Under Distributional Change

Even with effective updates, real environments remain non-stationary. Conditions change, actions trigger new constraints, and agents must recombine existing knowledge to anticipate consequences. Standard commonsense benchmarks often evaluate one-step inference under a fixed snapshot of conditions, leaving reasoning under distributional change under-specified and undermeasured [19].

A concrete example shows why non-stationarity is qualitatively different from ordinary robustness. Consider an assistant planning the reheating task under one condition: *"there is a microwave."* Now introduce a change: *"the microwave is broken."* The assistant must judge feasibility (is reheating still possible), consequence plausibility (does the alternative method

preserve the goal and constraints), and transition planning (what additional change would restore plausibility, such as obtaining a pot, using a stove, or switching to a no-heat option). Some changes are feasible but shift trade-offs, while others make the original goal unattainable. Critically, infeasible changes can still look linguistically ordinary. For example, *"heat soup without any heat source"* is syntactically similar to a normal instruction but violates a physical constraint. Models often struggle to discriminate such cases, especially when they rely on pattern completion rather than structured feasibility reasoning.

This thesis introduces a benchmarked formulation of such reasoning. We frame reasoning under distributional change as a multi-step discriminative process: models must judge whether a change is feasible, whether the resulting inference is plausible, and what additional change could restore plausibility if the inference becomes implausible. We refer to this capability as *metaphysical reasoning*, emphasizing that the model must distinguish realistic transitions from those that only exist in highly abstract or improbable variations. Conceptualization remains central here: abstract changes provide a compact way to represent a large space of shifts, and concept structure provides scaffolding for reasoning across them.

1.5 Thesis Scope and Approach

This thesis develops a coherent progression from representation, to robust inference, to scalable acquisition, to maintenance, and finally to reasoning under change. Across chapters, the lift-and-ground loop acts as the connective tissue: we study how to learn contextualized conceptualization and instantiation, how to use them to improve downstream robustness, how to acquire them at scale with limited supervision, how to integrate them into knowledge editing pipelines, and how to evaluate reasoning when the world itself changes.

1.5.1 Thesis Organization

This thesis is organized as follows. Chapter 2 reviews the related literature from three perspectives: commonsense reasoning tasks and benchmarks, commonsense knowledge acquisition, and methods for acquiring entity- and event-level conceptualizations (including extraction-, retrieval-, and generation-based paradigms). Chapter 4 introduces *CAT*, which formulates event-level conceptualization and instantiation as a paired abstraction and grounding process, and studies how such concept-structured representations can be learned and used to support more generalizable commonsense inference. Chapter 5 presents *CAR*, which connects conceptualization to downstream commonsense reasoning by showing how concept-level structure can

improve robustness and reduce false negatives under distribution shift in realistic QA settings. Chapter 6 develops *CANDLE*, a scalable distillation framework that leverages large language models to acquire conceptualizations and abstract inferential knowledge at scale, and integrates quality control to retain plausibility while expanding coverage. Chapter 7 proposes *CONKE*, a conceptualization-guided knowledge editing framework that combines automated verification, semantic enrichment via conceptualization and instantiation, and targeted editing to update commonsense knowledge in language models with improved generalization. Chapter 8 introduces *MARS* and the task of *metaphysical reasoning*, reframing reasoning under distributional change as a multi-step discriminative process and providing a large-scale benchmark to evaluate how models assess feasibility, consequences, and transitions under situational non-stationarity. Finally, Chapter 9 concludes the thesis by summarizing key findings, reflecting on the strengths and limitations of conceptualization as a central organizing principle, and outlining promising directions for future research.

CHAPTER 2

BACKGROUND AND RELATED WORK

This thesis studies how large language models can acquire, organize, and reliably apply commonsense knowledge beyond surface pattern matching. The core thread is *conceptualization*: abstracting concrete events and entities into reusable concepts, and grounding those concepts back into context to support inference. The main chapters develop this thread progressively—from introducing event-level conceptualization and its use in reasoning (CAT, CAR), to scaling conceptualization acquisition through distillation and quality control (CANDLE), and later to updating and stress-testing knowledge under non-stationarity (e.g., conceptualization-guided editing and metaphysical reasoning). This background chapter situates these contributions within the broader landscape, highlighting the problems they are designed to answer and the methodological choices they build upon.

A recurring gap in prior work is the mismatch between what we *evaluate* and what we *need*. Benchmarks such as commonsense QA have become the dominant lens for progress, yet they often conflate knowledge availability with reasoning and may reward shallow correlations. In response, the community has developed two complementary lines: richer task formulations and more scalable pipelines for acquiring commonsense resources. These two lines are tightly coupled: task designs implicitly define what "commonsense" looks like, while knowledge acquisition methods determine what information models can bring to those tasks. Understanding this coupling is essential for motivating why our later chapters emphasize concept-level structure, semantic coverage, and quality-controlled scaling.

At the same time, the emergence of increasingly large LLMs reshapes the practical constraints of commonsense reasoning research. Earlier paradigms were frequently validated on relatively small pre-trained language models, where performance gains could be attributed to carefully engineered supervision, structured KB injection, or task-specific fine-tuning. In the LLM era, however, models already contain broad but uneven commonsense; the central challenge shifts toward *reliability, generalization, and controllability*: how to extract and distill usable knowledge at scale, how to represent it in forms that transfer across contexts, and how to maintain coherence when knowledge must be updated or recomposed. This motivates the thesis-wide emphasis on conceptualization as a unifying representation that connects evalua-

tion, acquisition, and later editing and distribution-shift reasoning.

With this framing, we organize related work along two axes. First, we review commonsense reasoning tasks and benchmarks, emphasizing how evaluation protocols shape what models learn and how progress is measured. Second, we review commonsense knowledge acquisition, focusing on scalable mechanisms—from generative inference resources to LLM-based distillation—that supply the raw material for reasoning. We then narrow to *conceptualization acquisition methods* specifically, categorizing extraction, retrieval, and generative approaches, which together form the methodological foundation that the main chapters extend and integrate.

2.1 Commonsense Reasoning

Commonsense reasoning aims to equip NLP systems with the ability to make plausible inferences about everyday situations, a long-standing goal that remains challenging in practice [20]. A key theme in the literature is that progress is shaped by two tightly coupled aspects: how we *evaluate* commonsense reasoning through tasks and benchmarks, and how we *acquire* commonsense knowledge that models can use. Below we reorganize related work along these two axes.

2.1.1 Commonsense Reasoning Tasks

A variety of benchmarks have been proposed to test commonsense reasoning under different task formulations and sources of supervision [21–24]. Among them, commonsense question answering has become a standard evaluation setting because it directly measures whether models can select plausible answers under implicit world knowledge.

Zero-shot commonsense QA in particular emphasizes generalization, as models must solve unseen QA instances without using labeled supervision from the corresponding annotated training data. Existing approaches largely follow two paradigms. One paradigm keeps model parameters fixed and uses pretrained language models with prompting or inference-time reasoning procedures. Representative mechanisms include direct prompting and language modeling formulations [25, 26], self-talk style decomposition [27], cloze-based transformations [28], and dynamically constructed reasoning graphs combined with graph reasoning [29]. These methods are often instantiated with strong pretrained backbones or auxiliary generators such as ALBERT [30], COMET [31], and GPT-3 [32], yielding unsupervised QA pipelines that improve answer selection through structured inference [27–29, 33].

A second paradigm introduces explicit supervision via external commonsense knowledge bases (CSKBs), typically by converting knowledge triples into synthetic QA instances for fine-tuning [34–36]. A common construction maps the head entity and relation into a question, uses the tail as the correct answer, and samples alternative tails as distractors. This recipe has been extended to support cross-domain adaptation by incorporating CSKBs from different sources [37, 38], and to exploit multi-hop structure with graph-based modeling and learning objectives [39]. By training against plausible and implausible candidates, these methods encourage models to better separate valid commonsense from confounders in QA contexts.

2.1.2 Commonsense Knowledge Acquisition

In parallel with task design, a substantial body of work focuses on acquiring commonsense knowledge in forms that are easy to integrate into reasoning systems. A representative direction learns to generate *if-then* style inferences, where COMET is trained to produce structured commonsense knowledge that can be used by downstream models [21, 31]. While generation provides broad coverage, approaches that rely mainly on distributional regularities can be brittle when applied to situations that depart from the patterns observed during training, motivating further work on supervision and distillation.

Recent progress in large language models has led to renewed interest in knowledge distillation as a scalable route to acquire commonsense resources [40, 41]. Symbolic knowledge distillation uses human-designed prompts to elicit targeted knowledge from LLMs, then trains student models on the resulting outputs [42–45]. Other work studies transferring distilled signals across components, for example distilling from a ranker into a retriever to improve robustness of subsequent generation [46]. There is also a dialogue-oriented thread that distills conversational responses and rationales from LLMs to enhance dialogue agents with commonsense grounded behaviors [47, 48].

2.2 Conceptualization Acquisition Methods

Next, we review related methods for performing or collecting entity and event-level conceptualizations. We categorize them into three paradigms: extraction, retrieval, and generative-based methods, which are briefly demonstrated in Figure 2.1.

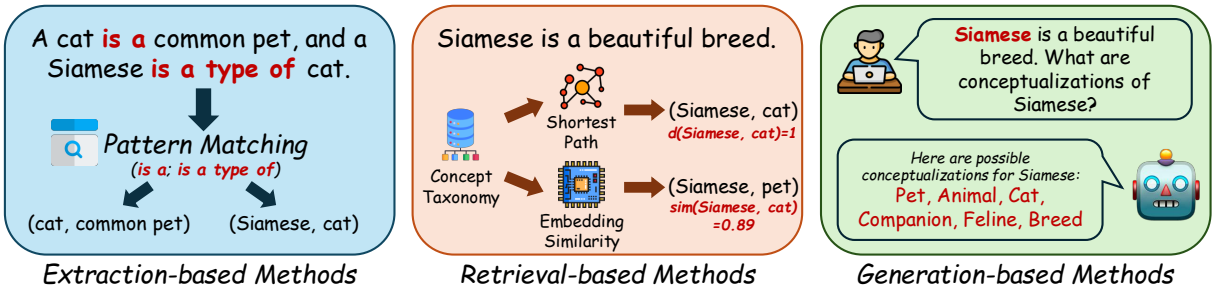


Figure 2.1: Conceptual demonstration of different types of methods in performing or collecting entity and event level conceptualizations. Instance and conceptualization pairs can be obtained at the end of each type of method.

2.2.1 Extraction-Based Methods

Extracting concepts from text is the earliest paradigm for systematically collecting conceptualizations [49, 50]. It typically involves first extracting all possible concepts from the text, followed with identifying the relationships between these concepts. In this process, concepts are recognized either by looking for the most frequent words or by matching against a predefined list of patterns, such as "is a,", "is a type of", etc. Instances are then matched by looking for the subject of these patterns in the text, which forms instance-conceptualization pairs. The main advantages of extraction-based methods [51–56] are easy implementation, high processing speed, and free of training data. This has facilitated the development of many large-scale concept taxonomies and knowledge bases, such as WordNet [57], ConceptNet [58–61], Probase [62, 63], and DBpedia [64, 65]. However, these methods, while successful in extracting conceptual relationships from text, are limited by text quality, reliance on predefined concepts, lack of semantic understanding, difficulty handling ambiguous words, and poor generalization to new domains or unseen concepts.

2.2.2 Retrieval-Based Methods

Semantic-Based Retrieval

Semantic-based retrieval methods aim to obtain conceptualizations by looking at the semantic similarity between the input instance and the concepts in a pre-defined concept taxonomy. It typically involves representing both the instance and a set of concepts into a shared semantic space and calculating the similarity between them. One representative approach is to use WordNet [57], a large lexical database of English words, to calculate semantic similarity between two words as their shortest path in the WordNet hierarchy [66]. Other methods [67–73] also share similar aspirations and define their own way of calculating such similarities. However, these

methods are usually limited by the need for comprehensive and accurate knowledge bases, high computational costs, the inability to handle unseen concepts, and the loss of important semantic context, prompting the development of neural-based retrieval methods.

Neural-Based Retrieval

Neural-based retrieval methods overcome previous limitations by leveraging neural networks (or language models) to learn the semantic representations of the input instance and the concepts in the knowledge base or concept taxonomy. Then, the similarity between the input instance and the concepts can be calculated based on the learned representation embeddings. This approach can be benefitted by the advancement in language modeling, such as BERT [74], RoBERTa [75], and DeBERTa [76, 77]. The most representative work in neural-based concept retrieval is AbstractATOMIC [78]. It uses GlossBERT [79] to encode concepts (from WordNet and Probase) and instances (extracted from events in ATOMIC [80]) into embeddings and leverage cosine similarity and human annotations to collect conceptualizations in a large scale manner. Other methods [81–87] similarly adopt different strategies in leveraging LMs as encoders, expanding the coverage of instances,training retrieval models. Despite their promising results, these methods are limited by their need for extensive labeled data, reliance on the completeness and accuracy of the knowledge base, and inability to retrieve new concepts that are out of training data.

2.2.3 Generative-Based Methods

Fine-Tuning-Based Generative Methods

Fine-tuning-based generative methods aim to take an entity or event as input and generate the concept directly via a fine-tuned generative language model. This approach allows the model to generate conceptualizations for new instances and offers maximum flexibility of the input. Several methods [78, 81, 85, 88–90] have adopted this paradigm in training generative conceptualizers, based on models such as GPT2 [91], BART [92], and T5 [93], for automated conceptualization acquisition. These methods typically train LMs on human-annotated or pre-existing conceptualization resources and yield outstanding results. However, fine-tuning-based generative methods are limited by their high computational cost, time-consuming and resource-intensive data collection, uncertain performance across diverse domains, and relatively low quality of novel concepts compared to human annotations [94]. While these are common limitations associated with fine-tuned generative models, zero-shot generative methods using powerful LLMs

and advanced prompting techniques potentially address these issues.

Zero-Shot Generative Methods

Finally, zero-shot generative-based methods leverage powerful LLMs [32, 40, 41, 95–97] to generate the concept directly from an input instance. They rely on the vast amount of internal knowledge within the model and human-crafted prompts to efficiently distill conceptualizations and abstract knowledge from the models. This is particularly useful when training data is scarce or when the domain is new and there are no existing training data available. Existing methods [81, 98–100] all share similar aspirations in collecting conceptualizations. The benefits are significant, as these methods can collect conceptualizations efficiently and at low cost without specific fine-tuning. The resulting conceptualization knowledge base are thus scalable and downstream models trained on them typically have improved generalization ability to new instances and domains. However, to ensure high-quality generated conceptualizations, it is recommended to implement quality control mechanisms such as human evaluation or discriminators as post-filters. Recent studies [98, 101] have shown that commonsense plausibility estimators [102] are effective for such quality control.

CHAPTER 3

CONCEPTUALIZATION AS A UNIFYING LENS FOR GENERALIZABLE REASONING

"*Concepts are the glue that holds our mental world together.*"– [103]

Chapters 1–2 motivated the problem of generalizable reasoning in language models and surveyed the surrounding literature. This chapter begins the thesis's conceptualization-focused development by establishing a unified vocabulary for abstraction and transfer. Before introducing new models or algorithms, the objective here is to make the object of study precise: what it means to lift from instances to concepts, how those concepts should be grounded back into context, and which forms of conceptualization are genuinely comparable versus easily conflated.

Concretely, this chapter formalizes conceptualization and instantiation as a "lift-and-ground" loop and shows how the same loop can be realized at multiple semantic levels. It then introduces a four-level view—entity, event, document, and system—to organize both prior work and the technical developments that follow. Finally, it delineates the scope of the thesis by centering entity- and event-level conceptualization as the main interface between language and reasoning. These definitions provide connective tissue for the remaining chapters: they shape what is learned, what is evaluated, and where generalization is expected to arise.

3.1 What is Conceptualization and Instantiation

Conceptualization is widely regarded as a central component of human intelligence, with deep roots in psychology [104–106] and close connections to computational linguistics and machine learning [107–109]. In the era of deep learning, conceptualization has also become a recurring theme in work that aims to improve the generalization of (Large) Language Models (LLMs; [40, 41, 95–97, 110]) across settings such as commonsense reasoning [85, 98, 111], causal reasoning [112, 113], and physical reasoning [114–116].

At a high level, the motivation is straightforward: when the surface form of a situation changes, models should still be able to recognize what kind of thing it is (an entity category, an event pattern, a document intent, or a task type) and reuse the relevant knowledge accord-

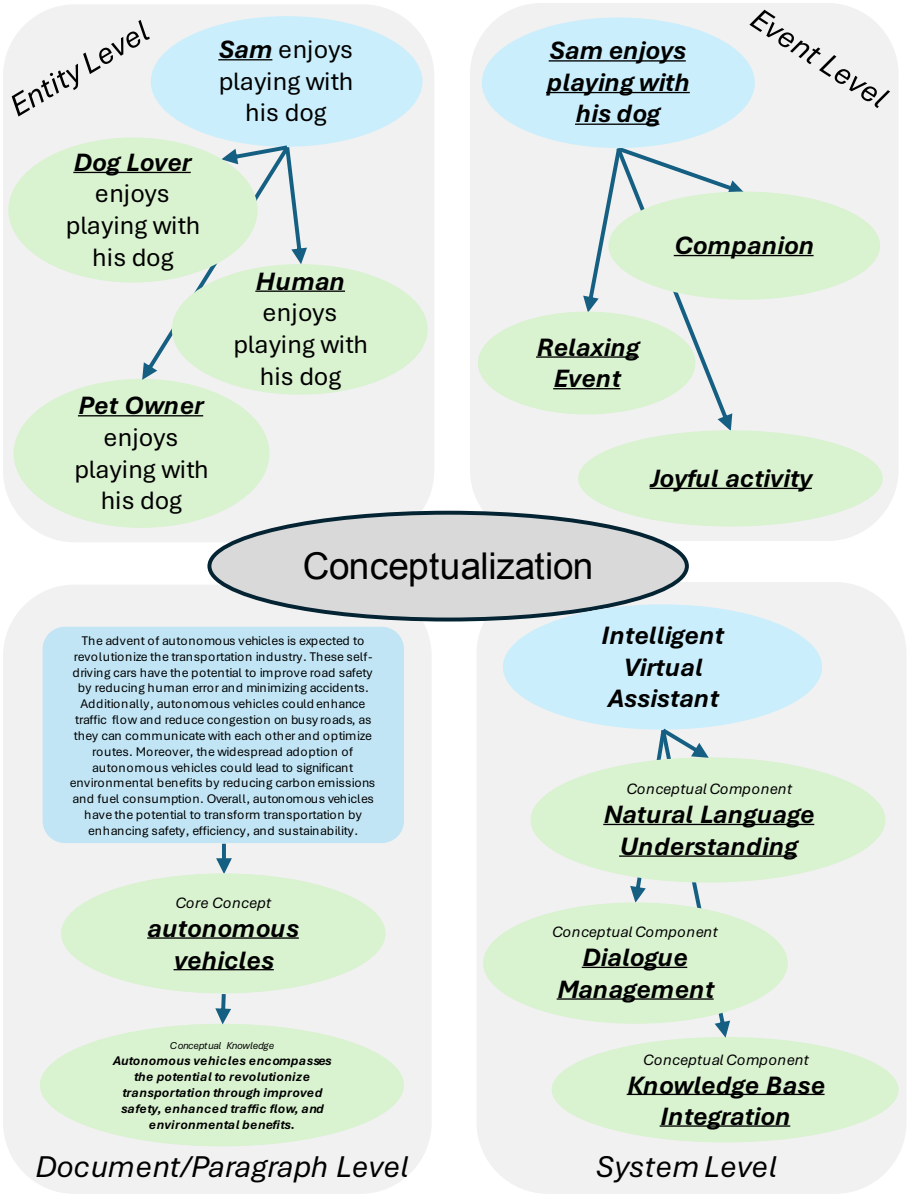


Figure 3.1: Examples of conceptualization at different semantic levels.

ingly. This chapter introduces conceptualization as a unifying lens for thinking about that ability and fixes terminology that will be used throughout the thesis.

3.1.1 Conceptualization

In this thesis, *conceptualization* refers to the process of *lifting* an instance into one or more abstract concepts that support reusable knowledge. Throughout, a *concept* is an abstract representation that indexes a set of instances and serves as an anchor for invariances that hold beyond any single example.

Definition (Conceptualization). Let $\mathcal{X}$ denote a space of instances (e.g., entities or events) and $\mathcal{C}$ denote a space of concepts. A conceptualization operator is a mapping

$$\phi : \mathcal{X} \to 2^{\mathcal{C}},$$

which assigns to an instance $x \in \mathcal{X}$ a (possibly singleton) set of concepts $\phi(x)$. Each $c \in \phi(x)$ is intended to *cover* x and to summarize properties shared by a broader cluster of instances. Equivalently, each concept c can be associated with an *extension* $\mathrm{Ext}(c) \subseteq \mathcal{X}$, and conceptualization selects concepts such that $x \in \mathrm{Ext}(c)$.

3.1.2 Instantiation

Conceptualization is most useful when it enables transfer: knowledge represented at the concept level must be brought back down to concrete situations. In this thesis, *instantiation* refers to this complementary grounding step, producing instance-level realizations or instance-level predictions from concepts, conditioned on a context.

Definition (Instantiation). Let $\mathcal{K}$ denote a space of contexts (e.g., partial descriptions, constraints, or downstream task inputs). An instantiation operator is a mapping

$$\psi : \mathcal{C} \times \mathcal{K} \to \mathcal{X},$$

which takes a concept $c \in \mathcal{C}$ and a context $\kappa \in \mathcal{K}$ and returns an instance $\hat{x} = \psi(c, \kappa)$ that is consistent with c under κ. More generally, instantiation may output an instance-level *claim* or *prediction* grounded in c (e.g., properties, likely consequences, or plausible next steps), which can then be consumed by a reasoning procedure.

Together, conceptualization and instantiation form a "lift-and-ground" loop: instances are lifted into concepts to obtain stable abstractions, and those abstractions are then grounded back into new contexts to support generalization [108]. This loop can be realized at multiple semantic levels, as illustrated in Figure 3.1. The next section introduces a four-level view that will serve as an organizing vocabulary for the rest of the thesis.

3.2 Four Levels of Conceptualization

Conceptualization can be categorized according to the type of instances being abstracted. In this thesis, four semantic levels are distinguished—entity, event, document, and system—each with different inputs, outputs, and evaluation concerns (Figure 3.1). The purpose of these definitions is not to exhaust the space of abstractions, but to provide a practical vocabulary that avoids collapsing fundamentally different operations under the same label.

3.2.1 Entity-Level Conceptualization

Entity-level conceptualization groups multiple entities under a shared concept [88, 117]. It is among the most common forms of conceptual organization in human cognition and is frequently used for knowledge acquisition [103, 118]. For example, entities such as "apple," "pear," and "grape" can be linked to the concept "fruit." Once such links are established, knowledge can be represented at the concept level (e.g., "fruit is delicious") while remaining grounded in instance-level evidence (e.g., "apple is delicious"). When encountering a novel entity that can be categorized as a fruit, the concept provides a shortcut for transferring expectations (e.g., likely taste, nutrition, or typical usage) to the new instance.

3.2.2 Event-Level Conceptualization

Event-level conceptualization extends the same idea from entities to events [78, 81, 98]. Here, a concept represents a higher-level pattern over actions, intentions, states, and outcomes, with the aim of preserving the original semantics as much as possible while enabling reuse across surface variations. For instance, "Sam playing with his dog," "Alex dancing in the club," and "Bob doing yoga" can be conceptualized as "relaxing events." This supports concept-level regularities such as: "If someone engages in relaxing events, they feel happy and relaxed." When a new event is observed (e.g., "Charlie paints the sunset"), event-level conceptualization enables recognizing it as a likely relaxing event and transferring the corresponding inferences to that new situation.

3.2.3 Document-Level Conceptualization

Document-level conceptualization broadens the scope from entities and events to spans of text such as paragraphs or full documents. It aims to construct a higher-level representation that captures the main ideas and essential information of the original text while preserving overall meaning and context. This objective closely overlaps with abstractive summarization [119, 120],

and it has been systematically reviewed in prior surveys [121–123]. Document-level conceptualization is included here to complete the semantic picture and to clarify terminology, but it is not treated as a primary focus in later chapters to avoid duplicating well-established lines of work.

3.2.4 System-Level Conceptualization

System-level conceptualization abstracts not a single text instance but a complex system—its behavior, functionality, or task space—into a higher-level representation. The idea is related to how complex systems are simplified for understanding and design [124], but it remains comparatively under-studied in NLP. A representative direction is to conceptualize NLP tasks themselves by grouping them according to objectives and characteristics while abstracting away low-level input/output formats and dataset-specific details [125]. Because the space of established methods and evaluations at this level is currently limited, system-level conceptualization is treated as contextual background rather than a central technical theme of this thesis.

3.2.5 Scope of This Thesis: Why Entity and Event Levels

Although conceptualization can be discussed at multiple semantic levels, this thesis primarily adopts *entity-* and *event-level* conceptualization as its main lens. These two levels sit at a particularly useful interface between language and reasoning: they are abstract enough to support transfer, yet concrete enough to remain tightly grounded in instance-level inference.

First, entities and events are the basic units from which much of everyday reasoning is composed. Entity concepts support category-based generalization and knowledge reuse [103, 118], while event concepts organize regularities over actions, intentions, and consequences that are central to commonsense, causal, and physical reasoning [78, 81, 98].

Second, entity and event levels naturally instantiate the lift-and-ground loop introduced in Section 3.1.1–3.1.2. They make it possible to study when and how concept-level abstractions remain stable under distribution shift, and how grounding those abstractions can enable reliable inference in novel situations [108].

Finally, the other two levels play different roles in this thesis. Document-level abstraction substantially overlaps with summarization and has been extensively surveyed elsewhere [121–123]. System-level conceptualization is promising but remains comparatively under-explored in NLP, with limited methodology and evaluation conventions [125]. For these reasons, they serve primarily as background context, while entity and event levels provide the main viewpoint for the technical developments in later chapters.

CHAPTER 4

SEMI-SUPERVISED LEARNING FOR EVENT CONCEPTUALIZATION AND INSTANTIATION

Chapter 3 introduced conceptualization and instantiation as a "lift-and-ground" loop: lifting concrete instances into reusable concepts, and grounding those concepts back into new contexts to support generalization. This chapter takes the first step from that conceptual lens to a concrete methodological question: *how can we learn such a loop automatically and at scale*, especially in domains where the space of possible abstractions is combinatorial and high-quality supervision is scarce.

Our focus is *knowledge-level event* conceptualization and instantiation in commonsense knowledge bases (CSKBs). Here, the goal is not merely to rewrite an event into a more abstract phrase, but to acquire *abstract commonsense knowledge*—regularities that hold across many surface realizations—and to use those regularities to deduce new, concrete commonsense facts in novel situations. This problem sits at the core of the thesis theme of generalization: instance-level knowledge alone is brittle under distribution shift, whereas concept-level knowledge offers a mechanism for reuse, but only if it remains grounded enough to support reliable inference.

This setting also exposes a practical obstacle that will recur throughout the thesis: conceptualization cannot be treated as a purely decontextualized rewriting step. In CSKBs, an abstraction that looks plausible in isolation may become invalid when placed back into a relational context (e.g., when paired with a particular relation and consequence), meaning that the correctness of abstraction is inseparable from how it will be *used*. At the same time, fully annotating context-sensitive conceptualizations and their downstream implications is infeasible at scale, which makes the learning problem fundamentally weakly supervised.

To address these challenges, we develop a semi-supervised framework that couples event conceptualization with instantiation in a single learning pipeline, using the two directions to regularize and bootstrap each other from abundant unlabeled data. Concretely, we introduce CAT (**C**ontextualized Conceptu**A**lization and Ins**T**antiation), which unifies event conceptualization and triple-level verification into a cycle that (i) filters context-incompatible abstractions, (ii) leverages instantiation to provide additional grounding signals, and (iii) scales via teacher–

student pseudo-labeling with a bootstrapping mechanism over alternative concepts and instances. In doing so, this chapter operationalizes the lift-and-ground loop from Chapter 3 in a realistic knowledge acquisition setting and demonstrates how learned abstract commonsense knowledge can translate into improvements in downstream commonsense reasoning.

The rest of the chapter formalizes the tasks and datasets, analyzes why existing conceptualization-only pipelines break under contextual constraints, and then presents CAT in detail (its training procedure, components, and ablations) before evaluating both the quality of the acquired abstractions and their utility for commonsense inference modeling.

4.1 Preliminaries

Commonsense reasoning is a crucial ability for machines to make situational presumptions and draw inferences from the knowledge that reflects our humans' understanding of situations and common facts [126, 127]. It has gained increasing popularity in the Natural Language Processing (NLP) community with the emergence of CommonSense Knowledge Bases (CSKB) [58, 128, 129] and large language models [31, 36, 130–132]. However, when encountering situations beyond the data given, more abstract background knowledge must be acquired and generalized to assist the reasoning [108], and language models trained with an autoregressive language modeling objective do not explicitly leverage such abstract knowledge during inference.

Instead, humans rely on conceptual induction and deduction [103] to make inferences on novel situations without the need to memorize all special cases. As shown in Figure 4.1, humans can derive conceptualizations based on the assertion that "PersonX watches a football game, as a result, he feels relaxed" to infer that "relaxing events can make someone feel relaxed," where the acquired abstract commonsense knowledge can be further used as general knowledge to perform reasoning on similar or associated situations. A new commonsense knowledge "PersonX plays with his dog, as a result, he feels happy and relaxed" can be deduced by instantiating *relaxing events* to *playing with his dog*. That is to say, when humans encounter an unknown situation, we first perform conceptual induction to derive plausible conceptualizations [78, 133] of similar situations and retrieve their associated abstract commonsense knowledge [108], which is indirectly induced and summarized from known commonsense knowledge. Plausible abstract commonsense knowledge within the context is then instantiated [134, 135] to deduce concrete commonsense knowledge concerning the situation and aid commonsense reasoning. As the cornerstone of generalizable commonsense reasoning, such a process is extremely challenging for machines to replicate due to the absence of contextualized conceptualizations and abstract com-

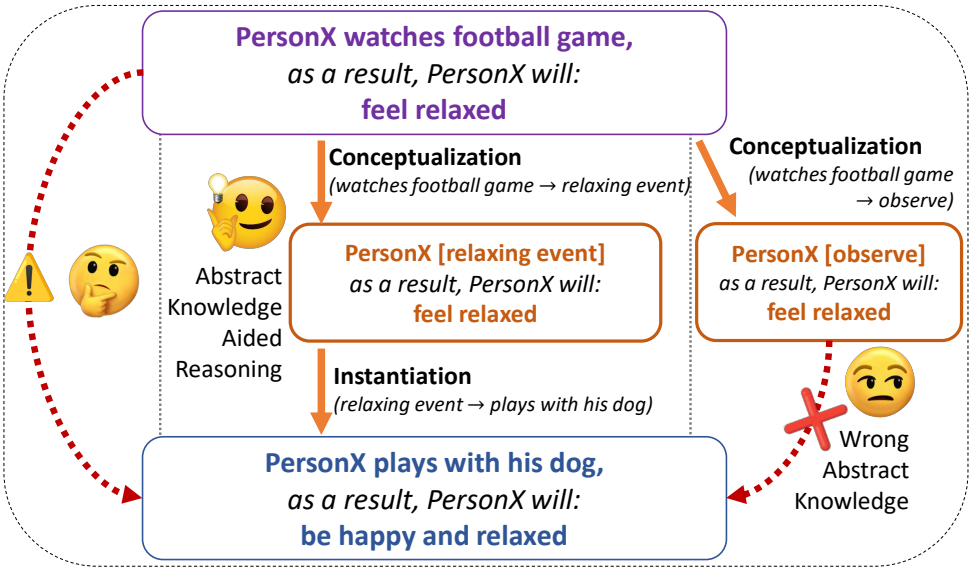


Figure 4.1: A demonstration of commonsense reasoning on an unknown situation, *PersonX plays with his dog*, with the aid of abstract commonsense knowledge. Decontextualized conceptualization, such as *observe*, may yield wrong abstract commonsense knowledge that cannot be instantiated within the corresponding context.

monsense knowledge in CSKBs and a lack of relevant methodologies.

4.1.1 Limitations of Conceptualization and Instantiation Pipeline

Yet, existing works address the process of induction and deduction separately via conceptualization and instantiation. Several methods performing conceptualization are proposed with a specific focus on entity-level [68, 69, 88, 136–138] and event-level [78, 139] semantics. Instantiation [135], as the process that simulates conceptual deduction, is tackled separately and not leveraged by these methods. Though abstract commonsense knowledge can be derived by using existing conceptualization methods to abstract a certain instance from factual commonsense knowledge, several limitations still exist.

First, the plausibility of abstract commonsense knowledge banks on both the correctness of *conceptualization* and proper *contextualization* under specific assertions. The latter one, which is an essential step for the deduction of abstract knowledge, is missing from current methodologies. Take Figure 4.1 as an example, the concept *observe* will not necessarily lead to the result of "feeling relaxed", as *observe* omits the entertaining property of the original instance as a cost of abstraction. Second, instantiating abstract commonsense knowledge can yield much more and diverse concrete commonsense knowledge that can serve as an augmentation of the training dataset, while current methods undervalue such a process and only focus on conceptualization. Finally, the complex *contextualization* and *conceptualization* of commonsense knowledge and

complicated hierarchy of conceptualizations and instantiations can easily bring more than two orders of magnitude of data on top of the original dataset. This makes current labeled data scarce and infeasible for practitioners to annotate all of them, leaving a large amount of unlabeled data.

To fill in these research gaps, in this chapter, we propose CAT (**C**ontextualized Conceptu**A**lization and Ins**T**antiation), a semi-supervised learning framework that unites event conceptualization and instantiation in cascade to conceptualize CSKBs and acquire abstract commonsense knowledge to aid commonsense reasoning. Inspired by how humans learn with concepts [140], we design a novel bootstrapping[1] method to enhance conceptualizations and abstract commonsense knowledge verification with the help of similar conceptualizations and instantiations as a reference. A mixture of pseudo-labeled and annotated conceptualizations are used to train a neural event conceptualization generator as an automatic acquisition. We demonstrate the effectiveness of CAT by using the acquired abstract commonsense knowledge to train COMET [31], a commonsense inference language model that generates if-then commonsense knowledge, and showing that our derived abstract commonsense knowledge can significantly improve commonsense inference modeling.

This chapter makes three contributions: (1) It formulates a semi-supervised learning framework (CAT) that couples event conceptualization and contextual triple verification through an instantiation-enabled cycle. (2) It shows that CAT can acquire higher-quality abstract commonsense knowledge at scale, suitable for training downstream commonsense inference models. (3) It demonstrates empirical gains on CSKB conceptualization benchmarks and improved commonsense inference modeling when the acquired abstract knowledge is used as additional training signal.

4.1.2 Task Definitions

Conceptualizing an event-centric CSKB to derive abstract commonsense knowledge comprises two steps [78]: event conceptualization and triple conceptualization.

Denote the triples in the original CSKB as $D_o = \{(h_o, r, t) | h_o \in H_o, r \in R, t \in T\}$, where H_o, R, and T are the set of heads, relations, and tails in the original CSKB. The first step only operates on head events without considering the context in r and t. The goal of event conceptualization is to produce conceptualized head event h_a from the original head h_o to represent an abstraction of h_o. In the second step, the task is to verify whether the conceptualized head h_a still makes sense in the context of r and t, as r and t will further restrict the level of abstractness

[1] Bootstrapping refers to the linguistics term in language acquisition that humans learn new knowledge by recognizing its semantic elements and connecting them with known knowledge [141].

in h_a. As shown in Figure 4.1, conceptualizing *watch football game* to *observe* is wrong within the context of having *feel relaxed* as a result. Plausible (h_a, r, t) triples will be considered as valid abstract commonsense knowledge.

Specifically, in the first step, there are two ways of conceptualizing head events alone: a *retrieval-based discriminative* way and a *generative* way. The retrieval-based discriminative paradigm identifies and links a component i in h_o to a concept c in a concept taxonomy C to form a conceptualization h_a by replacing i with c. The model needs to verify whether h_a is a valid conceptualization of h_o. The generative paradigm aims to generate a h_a directly given h_o and the designated component i in h_o.

Formally, denote the annotated dataset in the first step, event conceptualization, as $D_h^l = \{(h_o, h_a, y) | h_o \in H_o, h_a \in H_a, y \in \{0, 1\}\}$, where h_o is an original head event without conceptualization, h_a is a corresponding conceptualization of h_o, and y is the human-annotated label indicating whether such a conceptualization is plausible or not. The labeled dataset in the second step, triple conceptualization, is denoted as $D_t^l = \{(h, r, t, y) | h \in H_a, r \in R, t \in T, y \in \{0, 1\}\}$, where h is a conceptualized head event from the first step, r and t are a relation and a tail from the original CSKB accompanied with the corresponding original head h_o, and y is the human-annotated label indicating whether such abstract commonsense knowledge, in the form of a conceptualized triple, is plausible or not. Besides labeled datasets, unlabeled datasets are defined similarly as D_h^u and D_t^u only with the difference that labels y are missing. Thus, the task objective for discriminative event conceptualization is to determine whether a h_o can be properly abstracted using h_a, where h_a is derived by replacing a component $i \subset h_o$ with its linked concept c from a concept taxonomy C. The task objective for generative event conceptualization is to generate h_a directly from h_o with text generation models. For the triple conceptualization task, the objective is to distinguish whether a conceptualized triple (h_a, r, t), representing abstract commonsense knowledge, is plausible or not.

On the suitability of abstraction level. A natural question is how to determine whether a conceptualization is at a *suitable* level of abstraction, since the same event can often be lifted to multiple concepts of different granularity. In this thesis, we do not treat suitability as a fixed ontological level independent of context. Instead, a suitable abstraction is defined operationally by two coupled requirements. First, it must preserve enough semantic content from the original event to support the intended inferential relation. Second, it must be general enough to transfer beyond the original surface realization and support reuse across related instances. Abstractions that are too specific fail to improve generalization, while abstractions that are too coarse may

Data	Type	Train	Dev	Test
D^l	#event	107,384	12,117	11,503
	#triple	65,386	8,403	7,408
D^u	#event	304,983	36,023	31,578
	#triple	4,851,272	499,523	570,400

Table 4.1: Statistics of labeled data D^l and unlabeled data D^u in AbstractATOMIC.

erase the latent property that makes the original triple plausible.

This view is consistent with the motivating example in Figure 4.1. For the event *PersonX watches football game*, an abstraction such as *relaxing event* remains suitable because it preserves the aspect of the original event that explains the downstream inference *PersonX will feel relaxed*. By contrast, an abstraction such as *observe* is linguistically valid but inferentially too weak, because it removes the entertaining and restorative aspect that supports the tail event. In other words, the correctness of an abstraction cannot be judged solely by lexical or taxonomic relatedness; it must be evaluated relative to what inferential role the abstracted event is expected to play.

Accordingly, CAT determines abstraction quality through contextual survivability rather than through an externally fixed abstraction depth. A conceptualized head event is not only required to be plausible as an abstraction of the original head, but also required to remain plausible when re-inserted into the relation–tail context. Moreover, the instantiation step provides a complementary signal: if an abstraction can be grounded into multiple concrete, context-compatible realizations, this is evidence that the abstraction has retained the right reusable structure rather than merely becoming vague. The suitable abstraction level is therefore the level at which an event still supports the original inferential regularity while enabling transfer to new instances. This criterion is especially important for commonsense reasoning, where overly aggressive abstraction can easily produce conceptually neat but functionally unusable knowledge.

4.1.3 Datasets

To study conceptualization over CSKBs, we use the AbstractATOMIC dataset [78] as the benchmark to investigate the effect of having (r, t) as contextualization. In AbstractATOMIC, ATOMIC is used as the original CSKB [128]. Different from traditional conceptual knowledge benchmarks, the presence of commonsense relations and inferences enforces conceptualization and instantiations to be contextualized, and the event conceptualization adopts a *discriminative* way, where a syntactic parsing schema is defined to identify the components i in h_o to be heuristi-

cally linked to concept taxonomies Probase [62] and WordNet [57] to form conceptualized h_a. Such a heuristic can produce over 32 times more candidate conceptualized head events and over 10 times more conceptualized triples compared with the original ATOMIC, as the number of retrieved concepts from the concept taxonomy C can be manually controlled to acquire a large number of conceptualizations. Triple conceptualization is defined as predicting the plausibility of the triples whose head is conceptualized. Only 131K (26%) conceptualizations of 7K (45%) ATOMIC head events and 81K (1.3%) conceptualized triples are manually annotated as D_h^l and D_t^l, while others remain unlabeled D_h^u and D_t^u. The *trn/dev/tst* partition follows the same split as in the original ATOMIC. Statistics of AbstractATOMIC are shown in Table 4.1.

4.2 Related Works

4.2.1 Conceptualization and Instantiation.

Existing studies have explored conceptualization and instantiation largely as separate problems. Early work derived more general knowledge by abstracting over large collections of factoids obtained from WordNet synsets [57, 136]. Subsequent approaches mapped instances in sentences to higher-level concepts using weight matching with Probase [62, 68, 69, 137]. More recently, taxonomy-guided induction has been proposed to mine verb-oriented commonsense knowledge from verb phrases [142]. In parallel, benchmarks have been developed to probe whether language models possess conceptual knowledge, including zero-shot probing suites for conceptual reasoning [88]. While much of the above focuses on entity conceptualization, event-level conceptualization has also been studied, e.g., via an ATOMIC-based benchmark constructed with syntactic parsing, heuristic semantic matching, and human annotation [78, 128]. Relatedly, ultra-fine entity typing aims to assign free-form type phrases to named entities, nominals, and pronouns, sharing a similar goal of mapping mentions to semantically meaningful abstractions [143–145]. On the instantiation side, controllable generation has been used to automatically probe valid instantiations of abstract knowledge [135]. Although existing evidence suggests that pretrained language models still lack robust conceptual knowledge [88, 146], prior work has not explicitly integrated conceptualization and instantiation into a unified framework for deriving abstract knowledge that is both context-sensitive and generalizable.

4.2.2 Commonsense Reasoning.

Endowing NLP systems with commonsense reasoning remains a central yet challenging goal in artificial intelligence [20]. Accordingly, a variety of benchmarks have been proposed to evaluate commonsense reasoning from different perspectives [21–24]. A representative line of work learns to generate *if-then* commonsense knowledge; for example, COMET is trained to produce structured commonsense inferences that can support downstream reasoning tasks [21, 31]. However, purely generative approaches largely rely on distributional co-occurrence patterns, which can limit generalization beyond what is observed in training data.

4.2.3 Semi-Supervised Learning.

Semi-supervised learning (SSL) leverages unlabeled data to improve generalization, typically by augmenting training with pseudo-labeled examples [147–149]. This paradigm has been widely adopted across domains, including image classification [150, 151], text classification [152–154], commonsense knowledge base population [155], and named entity recognition [156, 157].

4.3 The CAT Framework

Then, we introduce our proposed Contextualized ConceptualizAtion and InsTantiation (CAT) framework for conceptualizing commonsense knowledge bases and acquiring abstract commonsense knowledge. An overview is presented in Figure 4.2. Our motivation is two-fold: first, adding instantiation after conceptualization to form a cycle can strongly benefit two conceptualization tasks simultaneously. On the one hand, instantiating conceptualized triple relies on the correctness of event conceptualization. On the other hand, properly conceptualized triples can benefit event conceptualization via instantiation by providing more context brought by (r, t). Second, to address the lack of annotations, we resort to pseudo labeling, a typical semi-supervised learning approach to automatically assign pseudo labels to the vast majority of unlabeled data using a teacher model.

Following prior work [78], we study the retrieval-based discriminative paradigm of event conceptualization and leave the generative paradigm as an intrinsic evaluation. In CAT, we unify event conceptualization and triple conceptualization into one cycle and make them mutually benefit each other through instantiation and conceptualization. Our framework can be summarized into four steps:

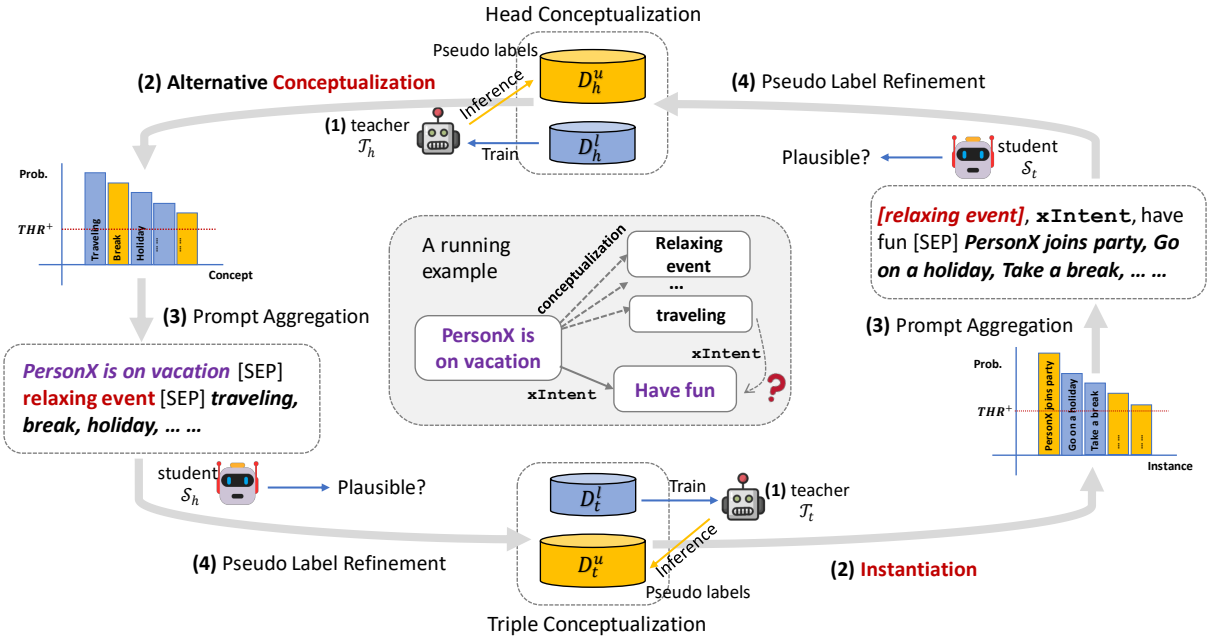


Figure 4.2: Overview of our CAT framework. A running example that conceptualizes the triple (PersonX is on vacation, `xIntent`, have fun) is presented in the figure, where the head is conceptualized first, and the model needs to determine whether the conceptualized triple still holds after the event conceptualization.

(1) Train teacher models for both event conceptualization and triple conceptualization on the labeled dataset D_h^l and D_t^l, respectively. Use the two teachers to assign pseudo labels to unlabeled datasets.

(2) Conduct alternative conceptualization or instantiation on labeled and pseudo-labeled data.

(3) Bootstrap (aggregate) the alternative concepts and instances in the second step using natural language prompt templates and train student models on both labeled and pseudo-labeled data.

(4) Use the student models to refine the pseudo labels and then re-train the student models.

The resulting models of CAT can derive abstract commonsense knowledge (h_a, r, t) from original CSKB triples (h_o, r, t), and we use that knowledge later for commonsense inference modeling.

4.3.1 Teacher Model Training

Two teacher models on both event and triple conceptualization tasks are trained separately on the labeled dataset D_h^l and D_t^l. As both tasks are inherently text/triple classification, we adopt KG-BERT [158] as the skeleton of our models. The event conceptualization model determines whether h_a is a valid conceptualization of h_o, and the triple conceptualization model determines whether a conceptualized triple (h_a, r, t) is plausible or not. Both verifiers take the prompted sentences in Section 4.3.3 as inputs. The two models θ are trained on annotated examples x_i with

a cross-entropy loss (Eq. 4.1) and used to provide pseudo labels to instances from the unlabeled datasets D_h^u and D_t^u. Two thresholds, T^+ and T^-, are set to determine the pseudo labels of unlabeled examples with high confidence and quality. Examples with a pseudo-labeled score higher than T^+ will be labeled $y_i = 1$, and those lower than T^- will be labeled $y_i = 0$. The rest will be discarded. The reason for setting two thresholds is to select high-confident pseudo examples to ensure quality.

$$L(x_i, \theta) = -\sum_{i=1}^{|x|} y_i \log\big(\theta(x_i)\big) \tag{4.1}$$

4.3.2 Alternative Conceptualization and Instantiation

According to prior work [103], when humans learn a new concept, we pre-extract similar known concepts in our minds and infer possibly equivalent unknown concepts on the fly. Inspired by this theory, we retrieve additional abstract concepts or instantiated events to help discriminate conceptualizations and abstract commonsense knowledge. For event conceptualization, we retrieve some alternative possible conceptualizations of h_o to accompany the learning of h_a. Additional conceptualizations of h_o from both labeled and pseudo-labeled examples are predicted again by the teacher model and ranked according to their plausibility score prediction. And top m conceptualizations are retrieved with m being a hyperparameter to control the number of retrievals. The same conceptualization h_a is removed from the candidate pool to avoid repetition. For triple conceptualization, the task is to determine whether a conceptualized head h_a, with concept c, is plausible given r and t as the context. we perform instantiation in cascade to instantiate c to some concrete instances to assist the learning process. Possible instantiations of c are extracted from annotated and pseudo-labeled event conceptualizations by searching for conceptualized events $h'_a \in H_a$ other than h_a with c as the concept and extracting their corresponding instances $i \subset h'_a$. Similarly, the instances are then scored by the teacher model, and the top n of them are retrieved.

We also instantiate the concept within the abstracted event for abstract triple verification by retrieving instances. New abstracted events, formed by replacing the target concept with the retrieved candidates, are sent to the previously trained PseudoReasoner for evaluation. A plausibility threshold $T_r = 0.5$ is fixed to filter positive retrieval only. The retrieved candidates are then ranked according to their plausibility score. Intuitively, alternative event conceptualizations can serve as hints for discriminating the correctness of the target conceptualization, and instantiations can carry additional contextualized information to help verify the plausibility of a conceptualized triple, which meets the objective of deriving abstract commonsense knowledge

Training Data	BLEU-1		BLEU-2		METEOR		ROUGE-L		CIDEr		**Human**	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test
$D_h^l + D_{0.95}^u$	**73.0**	71.1	**70.2**	63.0	**48.1**	**47.1**	**71.4**	70.7	**63.6**	66.9	**92.8**	**93.3**
$D_h^l + D_{0.9}^u$	71.3	**71.9**	65.2	63.8	45.7	46.7	69.8	**71.3**	63.4	**67.9**	90.5	91.0
$D_h^l + D_{0.8}^u$	68.2	68.4	65.9	**64.0**	44.8	44.0	66.6	66.7	60.0	62.0	86.0	85.7
$D_h^l + D_{0.7}^u$	66.5	67.2	57.2	62.6	43.0	43.4	65.9	65.8	60.4	61.2	79.0	80.3
$D_h^l + D_{0.5}^u$	64.9	62.4	58.3	51.1	41.2	40.9	63.8	63.0	58.2	59.4	74.5	79.0
D_h^l	67.6	65.3	56.8	53.1	43.5	43.1	65.7	66.6	60.2	60.9	70.0	81.5
Zero-Shot	20.2	17.0	6.80	4.11	5.80	4.70	3.80	3.00	1.90	1.60	15.0	11.5

Table 4.2: Performance (%) of GPT2 (XL) on the generative event conceptualization task. D_h^l stands for annotated labeled data, and D^u stands for the data acquired by CAT. The underfoot value indicates the threshold for selecting plausible pseudo labels. The best performances are bold-faced, and the second-best ones are underlined.

that is context-sensitive.

4.3.3 Prompt Aggregation

We then bootstrap the retrieved alternative conceptualizations/instantiations via natural language prompts. Here bootstrap [140] can be understood as binding the alternative retrievals and the target concept/triple together to strengthen the discrimination of the target concept/triple. As shown in Figure 4.2 step (3), the initially given input and retrieved concepts/instances are concatenated via human-defined prompts for both conceptualization tasks. Alternative concepts/instances are sorted in the order of their plausibility score ranking. `[SEP]` tokens are added to separate different components in the prompt. Two student models $\mathcal{S}_h$ and $\mathcal{S}_t$ for both tasks are trained using the modified text with such prompts as inputs. They are expected to learn the bootstrapping connectionism between the target and the additional retrievals we provided.

4.3.4 Pseudo-Label Refinement

All pseudo labels, initially derived by a teacher model trained on the original labeled dataset, are re-labeled according to the plausibility score predicted by our newly enhanced student models $\mathcal{S}_h$ and $\mathcal{S}_t$. Similar to the teacher model, two thresholds, T^+ and T^-, are applied to distinguish positive and negative examples for both tasks. In addition, negative labels are assigned to triples whose conceptualized head events are predicted as wrong conceptualizations by $\mathcal{S}_h$, as wrong conceptualizations will not yield plausible abstract commonsense knowledge.

4.3.5 Application and Evaluation of CAT

The resulting models of CAT include an event conceptualization model and a triple conceptualization model, both fine-tuned on the refined pseudo labels and the labeled data. These two models can be used to conceptualize ATOMIC to a larger commonsense knowledge base on a more abstract level. We further conduct intrinsic evaluations on the acquired event conceptualization model under a generative event conceptualization paradigm and extrinsic evaluations on the resulting conceptualized CSKB with commonsense inference modeling task (COMET [31]) in Section 4.4. Here we select COMET as the representative because it is a general commonsense model that can be applied to various downstream commonsense reasoning tasks such as SocialIQA [80], self-talk [27], and CSKB completion [12]. Meanwhile, generative event conceptualization enables performing automatic conceptualization scalably. Both are important applications and evaluations of CAT.

4.4 Experiments

We conduct conceptualization experiments using CAT in Section 4.4.1 and generative experiments as evaluations in Section 4.4.2. These experiments demonstrate that CAT has a strong capability in conceptualizing CSKBs, and better conceptualization modeling can help populate more novel and diverse commonsense knowledge and thus help commonsense modeling (COMET).

4.4.1 CSKB Conceptualization

Baselines. We collectively introduce the baselines for both event and triple conceptualization tasks, as they are inherently classification tasks. AUC is used as the evaluation metric. Under a supervised learning setting, we apply KG-BERT [158] model with BERT [74], BART [92], RoBERTa [75], DeBERTa [76, 77], and ELECTRA [159] as the backbone language models. We also attempt to leverage supervised generative language models as baselines. GPT2 [91] models are trained with a text generation objective only on positive examples, and we use perplexity as the prediction scores to calculate AUC. For the semi-supervised learning baselines, we leverage UDA [160], NoisyStudent [161], and PseudoReasoner [155] with RoBERTa-large being the backbone model.

Training Data	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE-L		CIDEr	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test
Zero-Shot	5.42	4.89	1.84	1.51	0.65	0.52	0.26	0.21	6.50	5.70	6.40	5.90	1.60	1.20
ATOMIC (subset)	38.1	38.1	25.4	25.7	18.7	18.8	15.5	15.7	14.9	14.9	33.0	33.2	27.6	27.8
$+D^l_t$	38.1	38.5	24.8	25.5	17.8	18.4	14.7	15.2	15.3	15.6	33.1	33.7	26.8	27.3
+Finetune	38.6	39.0	25.8	26.6	18.9	19.7	15.7	16.4	15.1	15.4	33.6	34.4	28.8	30.0
$+D^u_{\text{Abs.ATM.}}$	40.0	40.3	27.1	27.8	20.0	20.8	16.5	17.5	16.1	16.3	35.3	35.7	31.6	31.7
+Finetune	40.1	40.5	27.1	27.8	20.1	20.8	16.7	17.4	16.2	16.4	35.4	35.9	31.8	31.7
$+D^l_t + D^u_{\text{Abs.ATM.}}$	40.2	40.6	26.2	27.4	19.0	20.4	15.1	16.8	16.3	16.5	35.0	35.4	31.0	31.3
+Finetune	40.0	40.4	26.0	26.9	18.7	19.7	15.0	16.1	16.3	16.4	35.0	35.4	30.3	30.7
$+D^u_{\text{CAT}}$	**41.2**	41.9	**28.1**	**29.0**	**20.7**	**21.5**	**16.5**	**17.8**	**16.6**	16.9	35.9	36.5	**33.4**	33.7
+Finetune	41.1	**42.0**	28.0	29.0	20.4	21.5	16.4	17.6	16.6	**17.0**	**36.0**	**36.8**	33.2	**33.8**
$+D^l_t + D^u_{\text{CAT}}$	39.9	40.5	26.2	27.4	19.3	20.6	16.0	17.4	16.0	16.2	35.0	35.4	30.8	31.3
+Finetune	40.4	41.0	26.6	27.6	19.5	20.7	16.1	17.1	16.2	16.5	35.4	35.8	31.3	31.5

Table 4.3: Performances (%) of GPT2 (XL) on commonsense inference modeling task (COMET). D^l_t stands for annotated abstract triples, and D^u_{CAT} stands for abstract triples acquired by CAT. $D^u_{\text{Abs.ATM.}}$ contains triples that are pseudo-labeled by a supervised RoBERTa discriminator [78]. The best performances are bold-faced. Finetune refers to fine-tuning back on the ATOMIC subset.

Discriminative Results. The results for both tasks are presented in Table 4.6. Under a supervised learning setting, KG-BERT family mostly performs better on both tasks than GPT2 due to the fact that GPT2 is only fine-tuned on positive examples and thus cannot learn from negative examples that contain wrong conceptualizations and implausible abstract commonsense knowledge. As for the semi-supervised learning setting, previous SSL baselines are rather limited in improving the performance against supervised learning. The best PseudoReasoner only improves by 0.5% and 0.3% on the test set for both tasks compared with supervised RoBERTa-large models. Instead, models trained with CAT can outperform all other training methodologies. Comparing the test set performance with PseudoReasoner, small backbone models (BERT-base) can improve by 3.4% and 2.2%, and large models (RoBERTa-large) can be improved by 2.1% and 2.2%. This shows pipelining two-step conceptualizations as a loop and leveraging our proposed bootstrapping-based method can yield a larger performance gain compared with simply applying a semi-supervised learning strategy. For example, the results indicate that bootstrapping alternative conceptualization and instantiation plays the most important role in assisting learning conceptualization among all components of CAT.

4.4.2 Application and Evaluation of CAT

As CAT is a framework for acquiring conceptualized commonsense knowledge, including both conceptualized head events (from h_o to h_a) and abstract commonsense triples (h_a, r, t), we as-

sess these pseudo-labeled outcomes via two generative tasks with various threshold tuning as evaluations.

Generative Event Conceptualization. To intrinsically evaluate the effectiveness of CAT's event conceptualization, we use the acquired conceptualized head events as training data to learn a generative event conceptualizer. Specifically, the models are trained with instance-conceptualizations pairs in the format of "<*instance*> *is an instance of* <*concept*>". At the evaluation phase, the model is prompted with "<*instance*> *is an instance of* `[GEN]`" where <*instance*> is the instance to be conceptualized and `[GEN]` is the generation token. We then retrieve the top-1 generation and compare it against the target set from the evaluation dataset to compute four NLG metrics: BLEU [162], METEOR [163], ROUGE-L [164], and CIDEr [165] scores. These scores can be regarded as an approximation of the top-1 generations' recall. Additionally, we uniformly sample 500 generations from each evaluation split and conduct expert annotations on the plausibility of each conceptualization to ensure that out-of-domain concepts can be properly evaluated. The experts are asked to determine whether each top-1 generation is indeed a plausible conceptualization or not, such that the top-1 generations' precision is reflected. Thus, current evaluation measures jointly evaluate the top-1 generations' precision and recall, which makes it robust and non-easy to be impacted by repetition problems [166]. Zero-shot GPT2 and GPT2 fine-tuned on the originally labeled event conceptualizations in D_h^l are used as baselines. We also study the effect of the threshold T^+ that selects plausible conceptualized heads, where higher thresholds indicate higher plausibility regarded by CAT. The results are presented in Table 4.2. With a relatively high threshold, generators trained on a mixture of pseudo-labeled data by CAT and annotated concepts significantly outperform the baselines in every automated metric. A plausible rate of 93.3% is maximally achieved on the test set, which is 11.8% higher than the baseline. Gradually reducing the threshold also decreases the performance, indicating abstract heads with lower plausibility scores can be of poorer quality. Such results indicate that CAT can produce high-quality event conceptualizations for generative models to learn better conceptualizers without the need to annotate a large number of data.

Commonsense Inference Modeling (COMET). The second component of CAT produces triple-level abstract commonsense knowledge. We evaluate these abstract commonsense triples with a commonsense inference task that generates commonsense tails given heads and relations as inputs, as in COMET [31]. Following prior work [78], we apply the same training and evaluation process to the models. The base training data we use are a subset of ATOMIC triples

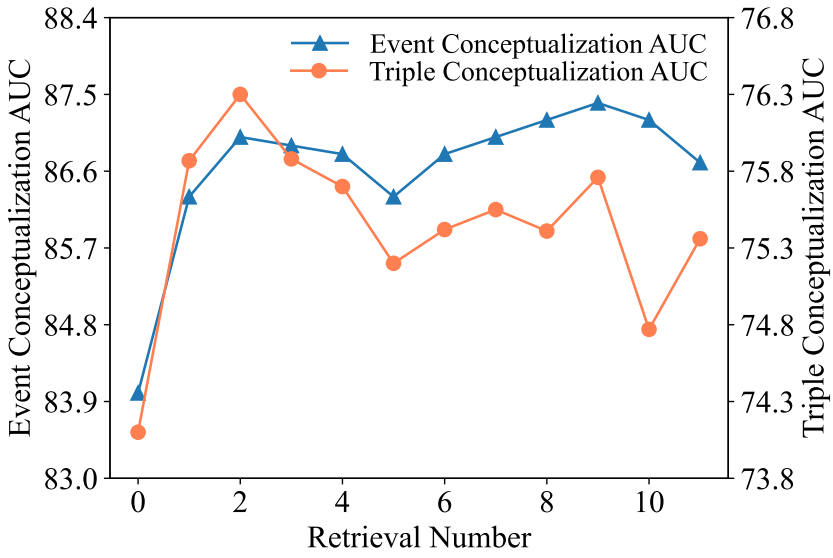


Figure 4.3: Ablation study on the number of retrieved conceptualizations/instantiations for CAT framework.

corresponding to those annotated abstract triples in D_t^l, which contains 17K (3.7%) among the original ATOMIC. We derive abstract commonsense knowledge using CAT from a subset of D_t^u where the heads correspond to those in the ATOMIC subset to ensure no data leakage, denoted as D_{CAT}^u. GPT2 is fine-tuned on the ATOMIC subset, the annotated abstract triples D_t^l, the abstract knowledge verified by CAT, or their combinations. The commonsense generation results are presented in Table 4.3. Similar to COMET [31], all models are evaluated on the original ATOMIC's full validation and testing sets. The best result is achieved using a mixture of the ATOMIC subset and abstract triples pseudo-labeled by our framework, with 0.95 as the threshold for selecting plausible triples. This indicates high-quality abstract commonsense triples can indeed provide a more general view of the original commonsense knowledge, thus helping commonsense inference. Additionally, training with our pseudo-labeled examples outperforms training with those annotated triples in AbstractATOMIC, which also validates the effectiveness of our model that leverages a large amount of unlabeled data. To further investigate how conceptual knowledge improves commonsense inference modeling, we conduct more empirical analysis in Section 4.4.4.

4.4.3 Number of Retrieved Alternative Conceptualizations and Instantiations.

We then study the ablation of bootstrapping different numbers of alternative conceptualizations/instantiations (denoted as #retrieval) in our CAT framework. For simplicity, when tuning the

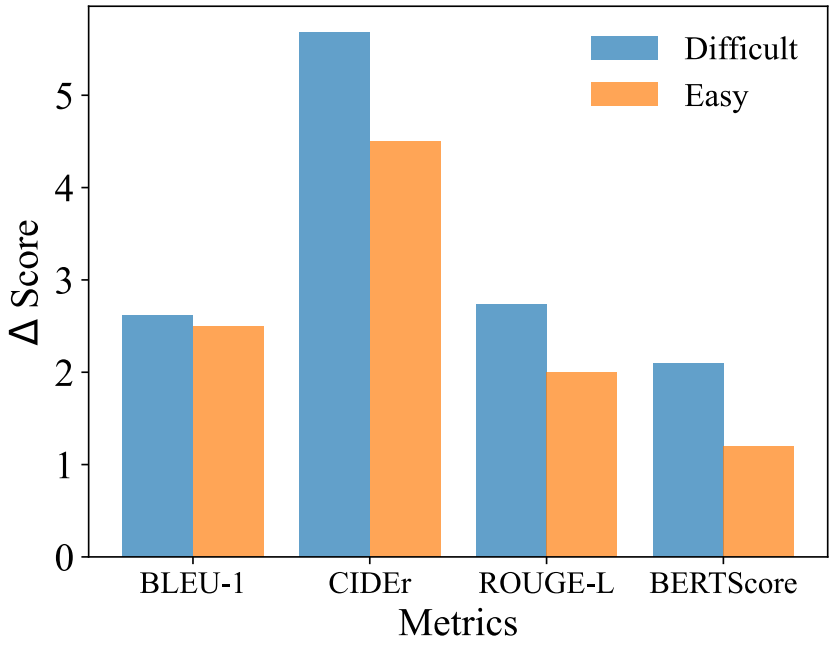


Figure 4.4: Comparison of performance improvement by GPT2 generator trained on the conceptualization-aided ATOMIC subset for two groups of testing head events.

#retrieval for one task, the #retrieval of the other task is fixed at the best value we acquired. We plot the test AUC score with #retrieval from 0 to 11 using BERT-base as the backbone model in Figure 4.3. #retrieval=0 refers to training with a simple student-teacher framework without bootstrapping alternative conceptualizations and instantiations. For event conceptualization, the performance generally positively correlates with the number of retrievals, while it starts dropping after 9. A reversed trend is observed for triple conceptualization, where using only two instances achieves the best performance. One possible reason is that in triple conceptualization, the retrieved instances are events and much longer than the retrieved concepts in event conceptualization, and aggregating various alternative events for a triple will cause language models to be less sensitive to the semantics of the original triple [167].

4.4.4 The Effect of Abstract Knowledge

We finally study the effect of abstract commonsense knowledge acquired by CAT by studying the semantic overlaps between training and testing data. We sort the test set by the BERTScore [168] between each individual testing entry against the whole training set in the original ATOMIC and split them in half to acquire two test groups. The testing entries with lower BERTScore on the training set indicate a larger semantic shift from the training set [169], which is also harder for models to discriminate [170]. We denote the testing group with a lower BERTScore as "Difficult" and the other half as "Easy". The performance gain on the two test set splits between the

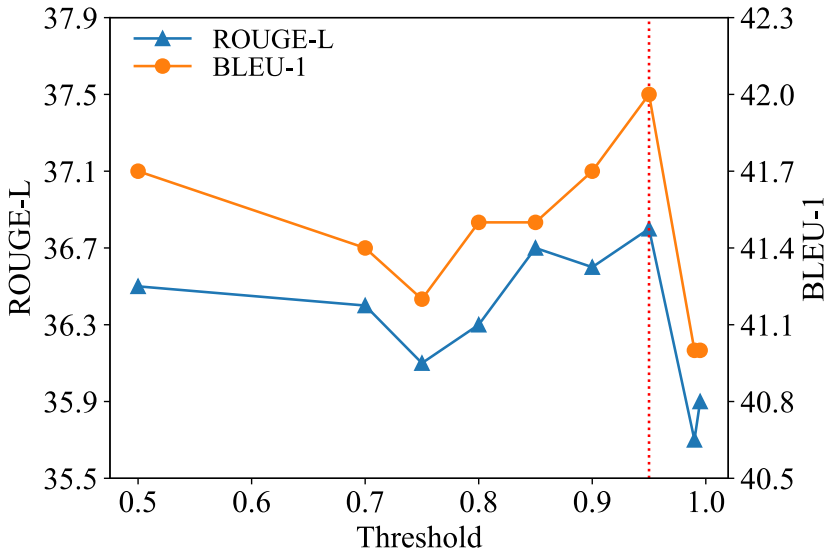


Figure 4.5: Performance (%) curve by COMET (GPT2-XL) on commonsense inference generation task with different thresholds for determining positive pseudo labels. Performance with the best threshold of 0.95 is marked as the red dotted line.

best conceptualization-aided COMET and the COMET trained on the ATOMIC subset only is reported in Figure 4.4. We can observe that training COMET with abstract commonsense knowledge leads to a larger improvement for harder test examples dissimilar from the original training set, indicating that introducing extra abstract commonsense knowledge can help COMET become more generalizable to harder test sets.

4.4.5 Ablation of Threshold

We conduct a more comprehensive study on the commonsense inference generation task by experimenting with the effect of threshold tuning when filtering abstract commonsense knowledge. Multiple thresholds ranging from 0.5 to 0.995 are experimented with to derive abstract commonsense knowledge of different qualities. COMET (GPT2-XL) generators are fine-tuned on the ATOMIC subset, augmented by a mixture of annotated and pseudo-labeled abstract triples. The performance curve according to the threshold is plotted in Figure 4.5.

It can be observed that gradually increasing the threshold from 0.75 will lead to better performance, which may be due to the improvement in data quality. However, increasing the threshold over 0.95 will cause a performance drop. One possible reason is the amount of pseudo-labeled triples significantly drops with a relatively high threshold, and COMET fails to learn well from annotated triples only. Using the CAT framework to pseudo-label unlabeled abstract triples leads to better performance than leveraging a RoBERTa-large supervised discriminator to assign pseudo-labels, which also validates the reliability of the triple conceptualization discriminator in

CAT. Also, it is noticeable that training COMET with triples based on our constructed ATOMIC subset is much worse than training with the full ATOMIC dataset. This indicates that exposing the model with substantial factual commonsense knowledge is still important, and only equipping the model with abstract commonsense knowledge is not enough for commonsense inference modeling.

4.4.6 Ablation of Framework Components

In this section, we study the effects of different components in CAT and the training strategy of CAT. These studies indicate that our framework design and the proposed bootstrapping method play an important role in CSKB conceptualization and are more effective than leveraging unlabeled data with pseudo labels.

Our CAT framework consists of three critical components that make CAT different from traditional semi-supervised baselines. They are denoted as:

• Bootstrapping: Assist the training of student models by retrieving alternative conceptualizations and instantiations and bootstrapping them via natural language prompts. Dropping this component will train student models with the original textual prompts that are also used by the teacher models.

• CAT Cycle: Unite event and triple conceptualization tasks by assigning negative pseudo labels to abstract triples whose conceptualized head is predicted as wrong conceptualization. Dropping this component will separate the framework into two lines of training, which are training event conceptualization and triple conceptualization models separately.

• Pseudo-label refinement: Refine the pseudo labels with the latest student models and retrain the student models. Dropping this component will not update any pseudo label and will not re-train the student model.

We then conduct ablation studies regarding these three components with semi-supervised CAT to prove the effectiveness of our framework design and proposed bootstrapping method. Each component is removed separately, and the test set performances by student models are reported. The results are shown in Table 4.4. From the results, bootstrapping alternative conceptualization and instantiation leads to the largest performance gain. Bridging event conceptualization discrimination with triple conceptualization also causes slight improvements. However, refining the pseudo labels and re-train the student models have barely any effect. Thus, our bootstrapping method is the most important component within the entire CAT framework and can effectively assist in learning conceptual knowledge.

Models	Event.	Triple.
CAT (BERT-base)	**87.4**	**76.3**
⋄ w/o Bootstrapping	83.1	73.0
⋄ w/o CAT Cycle	86.5	75.1
⋄ w/o Pseudo-label Refinement	87.4	76.2
CAT (DeBERTa-v3-large)	**89.2**	**80.0**
⋄ w/o Bootstrapping	84.0	77.7
⋄ w/o CAT Cycle	88.1	79.0
⋄ w/o Pseudo-label Refinement	89.1	79.7

Table 4.4: Ablation study on three components of CAT. Three components refer to the explanations above. The column **Event.** indicates test set AUC on the event conceptualization task, and the column **Triple.** indicates test set AUC on the triple conceptualization task.

Method	Event.	Triple.	Total
Supervised Baselines	107,384	65,386	172,770
UDA	412,367	4,916,658	5,329,025
Noisy-Student	412,367	4,916,658	5,329,025
PseudoReasoner	316,601	1,727,865	2,044,466
CAT	317,507	1,595,411	1,912,918

Table 4.5: Comparison between the number of training data for discriminative event conceptualization (Event.) and triple conceptualization (Triple.) tasks.

4.4.7 Ablation of Supervised CAT

We further study training CAT in a supervised learning setting to examine the role of unlabeled data. In supervised CAT, no teacher models are trained to provide pseudo labels. The alternative conceptualizations and instantiations are retrieved directly from the annotated event conceptualization data and bootstrapped later. Two student models are trained on the bootstrapped data only and evaluated on the same testing set, and the results are reported in Table 4.6. Compared with supervised learning baselines, supervised CAT can achieve a comparable result on the event conceptualization task. This may be due to the fact that the diversity of concepts drops without considering unlabeled conceptualizations. Improvements in the triple conceptualization task are more significant, and the results are comparable with semi-supervised CAT. This indicates that our framework design and bootstrapping method are successful in discriminating high-quality abstract commonsense knowledge, and leveraging a semi-supervised learning paradigm benefits more in event conceptualization discrimination.

4.4.8 Computational Cost Analysis

In this section, we compare the number of training data used for both CSKB conceptualization tasks to compare the computational cost across different frameworks and methodologies empirically. Both annotated and pseudo-labeled data are counted. The comparison result is presented in Table 4.5. All semi-supervised learning methods leverage a significant amount of unlabeled data due to the great scarcity of annotations. With threshold filterings, PseudoReasoner [155] and our CAT framework can abandon more than half of pseudo examples with poor quality. Even though our CAT framework can still outperform PseudoReasoner and achieve the best performance among all methods. Additionally, there is no notable increase in the number of model parameters as CAT also applies a teacher-student paradigm that is similar to Noisy-Student and PseudoReasoner. Even compared with the supervised baselines, CAT only doubles the parameters used. In conclusion, with comparable training data and parameters against other baselines, CAT can achieve much better results and state-of-the-art performances.

4.5 Conclusions

In conclusion, this chapter proposes CAT, a semi-supervised learning framework for commonsense reasoning, by leveraging the power of abstract commonsense knowledge. By achieving state-of-the-art performances in CSKB conceptualization tasks, we remarkably improve modeling commonsense inference, as an important cornerstone of many commonsense reasoning tasks. Our analysis also demonstrates that high-quality abstract commonsense knowledge can benefit commonsense inference modeling by providing more generalizability on hard commonsense knowledge. We hope this method can draw insights toward commonsense reasoning from a conceptualization perspective.

Framework	Backbone PTLM / Method	Event Conceptualization		Triple Conceptualization	
		Validation	Testing	Validation	Testing
Supervised Learning	BERT-base *110M*	$82.4_{\pm 0.05}$	$82.5_{\pm 0.31}$	$71.2_{\pm 0.58}$	$72.6_{\pm 0.71}$
	BERT-large *340M*	$82.8_{\pm 0.48}$	$83.1_{\pm 0.80}$	$72.4_{\pm 0.01}$	$73.7_{\pm 0.00}$
	BART-base *139M*	$83.8_{\pm 0.28}$	$84.4_{\pm 0.32}$	$72.0_{\pm 0.09}$	$72.6_{\pm 0.15}$
	BART-large *406M*	$85.0_{\pm 0.13}$	$85.2_{\pm 0.22}$	$74.5_{\pm 0.13}$	$76.2_{\pm 0.19}$
	RoBERTa-base *110M*	$84.1_{\pm 0.04}$	$84.5_{\pm 0.19}$	$72.2_{\pm 0.00}$	$74.1_{\pm 0.00}$
	RoBERTa-large *340M*	$85.2_{\pm 0.24}$	$85.5_{\pm 0.02}$	$75.3_{\pm 0.00}$	$76.9_{\pm 0.01}$
	DeBERTa-v3-base *214M*	$85.1_{\pm 0.08}$	$85.8_{\pm 0.07}$	$73.9_{\pm 0.10}$	$75.9_{\pm 0.04}$
	DeBERTa-v3-large *435M*	$85.8_{\pm 0.05}$	$86.2_{\pm 0.15}$	$76.9_{\pm 0.03}$	$78.0_{\pm 0.02}$
	ELECTRA-base *110M*	$85.4_{\pm 0.05}$	$85.8_{\pm 0.02}$	$74.3_{\pm 0.27}$	$76.2_{\pm 0.12}$
	ELECTRA-large *340M*	$84.7_{\pm 0.47}$	$85.3_{\pm 0.38}$	$75.6_{\pm 0.01}$	$77.9_{\pm 0.06}$
	GPT2-base *117M*	$60.0_{\pm 0.06}$	$59.1_{\pm 0.14}$	$52.8_{\pm 0.14}$	$55.9_{\pm 0.11}$
	GPT2-medium *345M*	$61.2_{\pm 0.11}$	$60.3_{\pm 0.08}$	$54.6_{\pm 0.17}$	$57.4_{\pm 0.09}$
	GPT2-large *774M*	$64.1_{\pm 0.05}$	$62.7_{\pm 0.08}$	$60.5_{\pm 0.11}$	$59.8_{\pm 0.06}$
	GPT2-XL *1558M*	$64.2_{\pm 0.19}$	$63.6_{\pm 0.22}$	$62.2_{\pm 0.08}$	$61.5_{\pm 0.10}$
Semi-Supervised Learning	UDA (TF-IDF)	$83.6_{\pm 0.29}$	$83.6_{\pm 0.24}$	$75.8_{\pm 1.26}$	$76.8_{\pm 1.34}$
	UDA (back-trans.)	$83.4_{\pm 0.27}$	$83.6_{\pm 0.24}$	$75.8_{\pm 1.25}$	$76.8_{\pm 1.34}$
	Noisy-Student	$86.4_{\pm 0.05}$	$86.5_{\pm 0.09}$	$75.4_{\pm 0.64}$	$76.7_{\pm 0.59}$
	PseudoReasoner (BERT-base)	$83.3_{\pm 0.11}$	$84.0_{\pm 0.24}$	$73.0_{\pm 0.14}$	$74.1_{\pm 0.33}$
	PseudoReasoner (RoBERTa-large)	$86.6_{\pm 0.25}$	$86.7_{\pm 0.33}$	$76.3_{\pm 0.12}$	$77.2_{\pm 0.21}$
CAT *(Supervised)*	BERT-base *110M*	$83.9_{\pm 0.42}$	$84.5_{\pm 0.43}$	$73.4_{\pm 0.32}$	$73.3_{\pm 0.23}$
	BERT-large *340M*	$82.8_{\pm 0.48}$	$83.1_{\pm 0.80}$	$72.4_{\pm 0.01}$	$73.7_{\pm 0.00}$
	BART-base *139M*	$84.9_{\pm 0.05}$	$85.4_{\pm 0.08}$	$75.2_{\pm 0.06}$	$76.9_{\pm 0.21}$
	BART-large *406M*	$86.2_{\pm 0.05}$	$86.0_{\pm 0.06}$	$76.8_{\pm 0.21}$	$78.7_{\pm 0.31}$
	RoBERTa-base *110M*	$85.5_{\pm 0.06}$	$86.0_{\pm 0.06}$	$76.6_{\pm 0.12}$	$77.2_{\pm 0.18}$
	RoBERTa-large *340M*	$86.2_{\pm 0.31}$	$86.2_{\pm 0.31}$	$77.7_{\pm 0.19}$	$78.5_{\pm 0.28}$
	DeBERTa-v3-base *214M*	$85.8_{\pm 0.15}$	$86.2_{\pm 0.07}$	$76.8_{\pm 0.28}$	$79.0_{\pm 0.20}$
	DeBERTa-v3-large *435M*	$\mathbf{86.3_{\pm 0.11}}$	$\mathbf{86.7_{\pm 0.08}}$	$\mathbf{78.4_{\pm 0.20}}$	$\mathbf{79.5_{\pm 0.18}}$
	ELECTRA-base *110M*	$85.5_{\pm 0.12}$	$85.7_{\pm 0.08}$	$76.7_{\pm 0.05}$	$77.3_{\pm 0.16}$
	ELECTRA-large *340M*	$86.2_{\pm 0.66}$	$86.0_{\pm 0.62}$	$77.8_{\pm 0.11}$	$78.5_{\pm 0.09}$
CAT *(Semi-Supervised)*	BERT-base *110M*	$87.1_{\pm 0.06}$	$87.4_{\pm 0.11}$	$74.3_{\pm 0.26}$	$76.3_{\pm 0.38}$
	BERT-large *340M*	$87.7_{\pm 0.16}$	$88.0_{\pm 0.19}$	$75.8_{\pm 0.23}$	$77.8_{\pm 0.36}$
	BART-base *139M*	$88.2_{\pm 0.09}$	$88.2_{\pm 0.09}$	$75.7_{\pm 0.09}$	$78.0_{\pm 0.14}$
	BART-large *406M*	$88.6_{\pm 0.07}$	$88.7_{\pm 0.10}$	$77.2_{\pm 0.12}$	$79.0_{\pm 0.14}$
	RoBERTa-base *110M*	$88.4_{\pm 0.12}$	$88.3_{\pm 0.08}$	$76.9_{\pm 0.16}$	$78.0_{\pm 0.19}$
	RoBERTa-large *340M*	$89.0_{\pm 0.15}$	$88.8_{\pm 0.20}$	$78.2_{\pm 0.08}$	$79.4_{\pm 0.14}$
	DeBERTa-v3-base *214M*	$88.8_{\pm 0.12}$	$88.9_{\pm 0.08}$	$77.5_{\pm 0.10}$	$79.9_{\pm 0.07}$
	DeBERTa-v3-large *435M*	$\mathbf{89.1_{\pm 0.05}}$	$\mathbf{89.2_{\pm 0.14}}$	$\mathbf{78.7_{\pm 0.16}}$	$\mathbf{80.0_{\pm 0.33}}$
	ELECTRA-base *110M*	$88.7_{\pm 0.10}$	$88.9_{\pm 0.10}$	$74.9_{\pm 0.15}$	$75.5_{\pm 0.40}$
	ELECTRA-large *340M*	$88.6_{\pm 0.77}$	$88.5_{\pm 0.70}$	$74.9_{\pm 0.15}$	$75.5_{\pm 0.40}$

Table 4.6: Performance (%) by our CAT framework on the discriminative event conceptualization and triple conceptualization tasks. We report the average AUC score and standard deviation across experiments with three random seeds. The best performances within each framework are underlined, and the best among all models are bold-faced.

CHAPTER 5

CONCEPTUALIZATION-AUGMENTED COMMONSENSE QUESTION ANSWERING

Chapters 3–4 established conceptualization as a concrete mechanism for generalization and studied how to *learn* the conceptualization–instantiation mapping itself—together with a plausibility model—so that abstract commonsense triples can be acquired from largely unlabeled CSKB data. In particular, Chapter 4 (CAT) treated conceptualization and instantiation as a coupled "lift-and-ground" loop and showed that, when supervision is scarce, the two directions can regularize and bootstrap each other to scale the acquisition of *contextualized* abstract commonsense knowledge in CSKBs. This gives us a practical answer to a foundational question in this thesis: where can concept-level knowledge come from, and how can we obtain it reliably beyond manual curation.

This chapter moves from *acquiring* conceptualized knowledge to *using* it as supervision for downstream reasoning. Commonsense question answering is a natural next step: it operationalizes generalization as the ability to select the most plausible option under new distributions, often without access to labeled training data from the target benchmark. At the same time, it exposes a key limitation of prior CSKB-to-QA pipelines: even if a CSKB is large, converting it into training instances with heuristic negative sampling can inject systematic noise (false negatives) and can leave important abstractions implicit rather than learnable.

Conceptualization provides a unifying way to address both issues. By lifting instance-level triples to concept-level variants, we can expand the coverage of CSKB-derived supervision with abstract knowledge that transfers across surface forms; by using concept-level overlap as an additional constraint, we can generate harder yet cleaner distractors that better reflect semantic plausibility. Building on the conceptualization models and plausibility filtering developed in CAT, we propose CAR (Conceptualization-Augmented Reasoner), which injects conceptualized knowledge into CSKB-based QA synthesis and uses concept-aware distractor sampling to reduce false negatives—thereby translating the lift-and-ground principle into improved robustness in zero-shot commonsense QA.

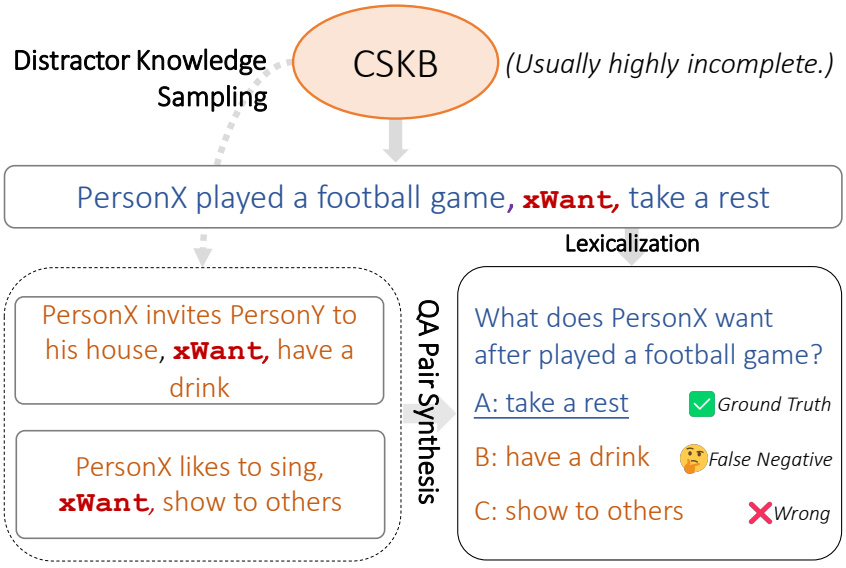


Figure 5.1: An example of constructing synthetic QA pairs from CSKB [35]. The simple heuristic used in this process can result in false negative options.

5.1 Introduction

Enabling machines to reason with commonsense knowledge is a long-standing goal of artificial intelligence [126, 171]. Pre-trained Language Models (PLMs; [74, 159]) fine-tuned on task-specific training sets achieve remarkable near-human performance on held-out test sets, yet struggle to generalize to examples that are distributionally different from their training sets [172–175]. This discrepancy arises because fine-tuned PLMs often rely on spurious, dataset-specific correlations to learn a task rather than learning to fully leverage implicit commonsense knowledge required for reasoning [176]. For reasoning systems to be effective, though, they must be robust across domains and generalize beyond the specificities of individual datasets.

To confront the generalization issue in commonsense reasoning tasks, the task of zero-shot commonsense Question-Answering (QA) requires models to answer questions for evaluation benchmarks without access to their corresponding training data [27, 177]. Among several methods that tackle this task, the most performant ones inject commonsense knowledge from CSKBs [129, 178] into PLMs by fine-tuning them on synthetic QA pairs transformed from commonsense knowledge triples, where the head and relation are transformed to a question, and the tail serves as a ground answer. Negative examples are randomly sampled with keyword-overlap constraints [35]. Such knowledge injection benefits not only QA tasks that are derived from CSKBs, such as SocialIQA [80], which is derived from ATOMIC [128], but also QA datasets in other domains [114].

Despite recent advancements in this area, two major challenges remain. First, manually cu-

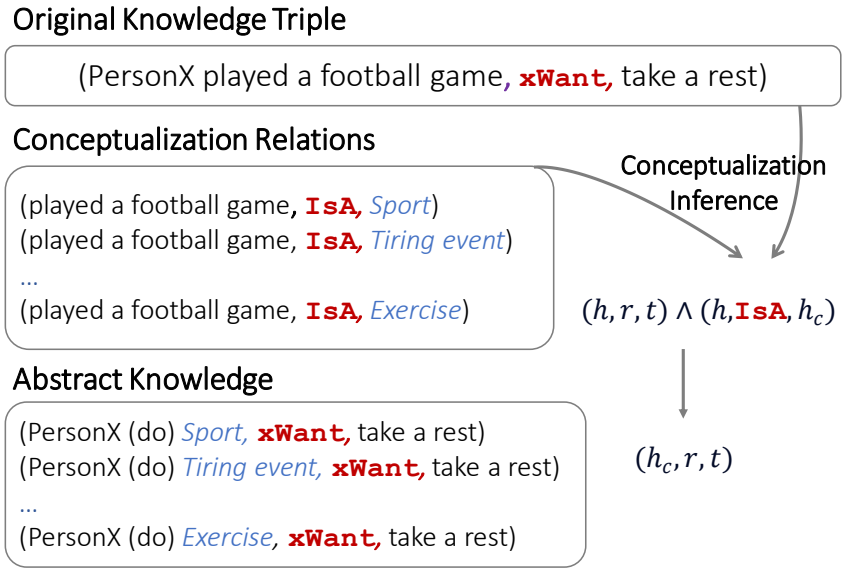


Figure 5.2: An example of conceptualization inference. More abstracted knowledge, such as (Do sport, `xWant`, take a rest), can be obtained through conceptualization.

rated CSKBs, such as ATOMIC, are incomplete [179]. While consolidating multiple CSKBs can improve coverage, it remains infeasible to cover all conceivable knowledge for the vast range of entities and situations in the real world [78]. Automatic methods for expanding CSKBs exist, such as knowledge base completion [8, 12], and knowledge distillation from large language models [42, 180], but they either fail to provide knowledge about novel entities or only provide highly accurate yet less informative knowledge (e.g., vague adjectives, such as *happy*, as situation descriptors). Second, in zero-shot commonsense QA, negative examples are required for models to learn to distinguish the validity of commonsense scenarios [181]. However, existing negative QA examples are synthesized using simple heuristic-based negative sampling without considering deeper semantics, resulting in too many false negative options. For instance, in Figure 5.1, "have a drink" is also plausible in the context of "after playing a football game." This indicates that the lack of common keywords does not ensure the implausibility of distractors, as the sampled knowledge may still be semantically related. These questions that label plausible options as negative instances confuse the model during training, impeding its ability to discern correct commonsense knowledge. To identify dissimilar knowledge for distractors, signals representing higher-level semantics of both the question and commonsense knowledge are required.

We tackle both of these challenges by utilizing *conceptualization*. As prior work [103] posits, humans rely on conceptual induction to draw inferences about unseen situations without the need for memorizing specific knowledge. Conceptualization [78] offers a similar capability by abstracting a set of instances into concepts, which allows for the derivation of abstract commonsense knowledge associated with each concept that can be instantiated to assist reasoning

on specific downstream situations. For example, in Figure 5.2, "play a football game" can be conceptualized as a *tiring event*, which further generalizes as abstract knowledge. The benefits of conceptualization are twofold. First, conceptualized commonsense knowledge introduces abstract knowledge through a one-step concept inference based on the original CSKB, enhancing knowledge coverage. Second, as the abstract knowledge is conditioned on the original knowledge, the *recall* of knowledge regarding the same head is increased, leading to more fine-grained constraints for negative option sampling.

Inspired by these advantages, we propose **CAR** (**C**onceptualization-**A**ugmented **R**easoner), a simple yet effective zero-shot commonsense QA framework that leverages *conceptualization* to expand existing CSKBs and reduce false-negative distractors. We first augment the original CSKB with conceptualization to infuse abstract commonsense knowledge to improve knowledge coverage. Then, we propose a conceptualization-constraint sampling strategy that generates distractors with concept-level constraints to prevent false negative options (Section 5.4). Under our evaluation protocol and prompts (Section 5.5), CAR achieves higher average accuracy than the two prompted LLM baselines we tested (GPT-3.5 `text-davinci-003` and ChatGPT `gpt-3.5-turbo`). In Section 5.6, we analyze why CAR works by providing human evaluations that show a significant reduction of false negative options compared to other methods. Finally, our analysis reveals that conceptualization-augmented training examples tend to be more *ambiguous* [182] than those produced by prior heuristics, leading to better out-of-domain generalization.

The contributions of this chapter are three-fold: (1) We propose a novel zero-shot commonsense QA framework that incorporates conceptualization-based abstract commonsense knowledge to enhance the generalizability of QA models. (2) We introduce a new conceptualization-constrained distractor generation strategy for synthesizing QA pairs from CSKB, which greatly improves the fairness of distractors. (3) We validate the effectiveness of CAR by achieving state-of-the-art performances on five QA benchmarks, surpassing all existing methods and large language models, such as ChatGPT. We also conduct comprehensive analyses to evaluate the advantages brought by *conceptualizations*.

5.2 Related Works

Zero-shot Commonsense QA. Zero-shot commonsense QA evaluates a model's reasoning generalizability on unseen QA entries without any supervision signals from the corresponding annotated training data. To tackle this task, two primary pipelines have emerged in existing

works. The first paradigm employs off-the-shelf language models without changing the parameters, either using vanilla language modeling with prompts [25, 26], or with some inference-time mechanisms specifically designed for reasoning, such as self-talk [27], cloze translation [28], and dynamic generation of reasoning sub-graphs and graph reasoning [29]. The second pipeline leverages external CSKBs as knowledge sources to provide PLMs with additional supervision signals for further fine-tuning [34–36]. A common strategy involves converting knowledge triples in CSKBs to synthetic QA pairs by transforming the head and relation to a question, the tail to a gold answer, and (randomly) sample tails from other heads as distractors. Such fine-tuning paradigm benefits from incorporating CSKBs within different domains [37, 38] and exploiting multi-hop graph structures with graph neural networks [39], and heightens the model's commonsense sensitivity in a QA context, which leads to state-of-the-art performances. By comparing plausible commonsense knowledge with negative examples, these methods teach the model to better distinguish between them.

Conceptualization. Conceptualization refers to the process of abstracting a group of instances or events into a general concept [68, 69]. In commonsense reasoning, it simulates conceptual induction [103] and enables the derivation of abstract commonsense knowledge under the specific contextualization of the original commonsense knowledge [108], which is often lacking in existing CSKBs. Around many existing works studying conceptualization [88, 136, 137, 142], AbstractATOMIC [78] investigates it at event-level semantics and construct AbstractATOMIC, an event conceptualization benchmark and knowledge base based on ATOMIC [128]. Recently, CAT [85] proposes to conceptualize CSKBs at scale with semi-supervised learning and demonstrate abstract knowledge can enhance commonsense inference modeling [31, 183]. With current works mostly investigating the problem of *conceptualization* itself, none of them have extrinsically evaluated the impact of conceptualization on downstream tasks, such as commonsense QA [21] or machine reading comprehension [184].

Data Augmentation. Data augmentation aims at generating new examples from existing data to expand the size and diversity of a training set without requiring costly data annotations [185]. Various methods have been proposed to augment textual data, including those using random perturbation [185], text embeddings [186], lexical semantics [187], back translation [188], and large language models [42, 180, 189] for CSKB construction. Nevertheless, text-perturbation-based augmentations do not provide new knowledge to CSKBs, and knowledge mining from large language models suffers from high typicality (e.g., favoring simple commonsense over

informative yet rare commonsense) and low density, still making negative sampling subject to false negatives [12].

5.3 Problem Definition

5.3.1 Definitions

Conceptualization. Formally, denote a CSKB as D with knowledge triples in the format of $D = \{(h, r, t)|h \in H, r \in R, t \in T\}$, where H, R, and T are the sets of heads, relations, and tails in the original CSKB. Following prior work [78], the conceptualized CSKB, conditioned on D, can be denoted as $D^C = \{(h_c, r, t)|h_c \in H_c, r \in R, t \in T\}$, where H_c is the set of conceptualized head events. Specifically, each conceptualized head h_c is obtained by replacing a component $i \in h$ with its abstract concept c while ensuring that the formed (h_c, r, t) triple is still plausible in the original context (r, t). Such (h_c, r, t) triples are commonly referred to as abstract commonsense knowledge.

Zero-shot Commonsense QA. In this chapter, we employ the zero-shot commonsense QA task proposed [35] to study our framework. First, the CSKB D is transformed into multiple (Q_i, A_i) pairs where Q_i is a natural langauge question and $A_i = \{A_{i,1}, A_{i,2}, ..., A_{i,m}\}$ is a set of options with m candidates. Specifically, for a given knowledge triple $(h, r, t) \in D$, we convert h, r into Q_i via natural language templates and use t as the ground answer. Additionally, we retrieve $m - 1$ distractors from other triples sampled from D using a manually defined strategy, such as keyword overlap filtering. The objective of our task is to train a QA model from the synthetic QA sets $D^Q = \{(Q_i, A_i)|(h_i, r_i, t_i) \in D\}$. Once trained, the model is tested on held-out test entries (Q^{test}, A^{test}) from QA benchmarks. This requires the model to perform zero-shot commonsense reasoning since the training data from the target benchmarks are unavailable to the model.

5.3.2 Dataset

We use ATOMIC [80] as the source CSKB D and AbstractATOMIC [78] as the conceptualized CSKB D^C conditioned on ATOMIC. ATOMIC contains inferential commonsense knowledge, in the format of (h, r, t) triple, that is associated with commonly seen events. Specifically, the heads of ATOMIC triples are events, whereas the tail nodes are either events or attributes. For conceptualization, we use the human-annotated abstract knowledge from AbstractATOMIC [78]

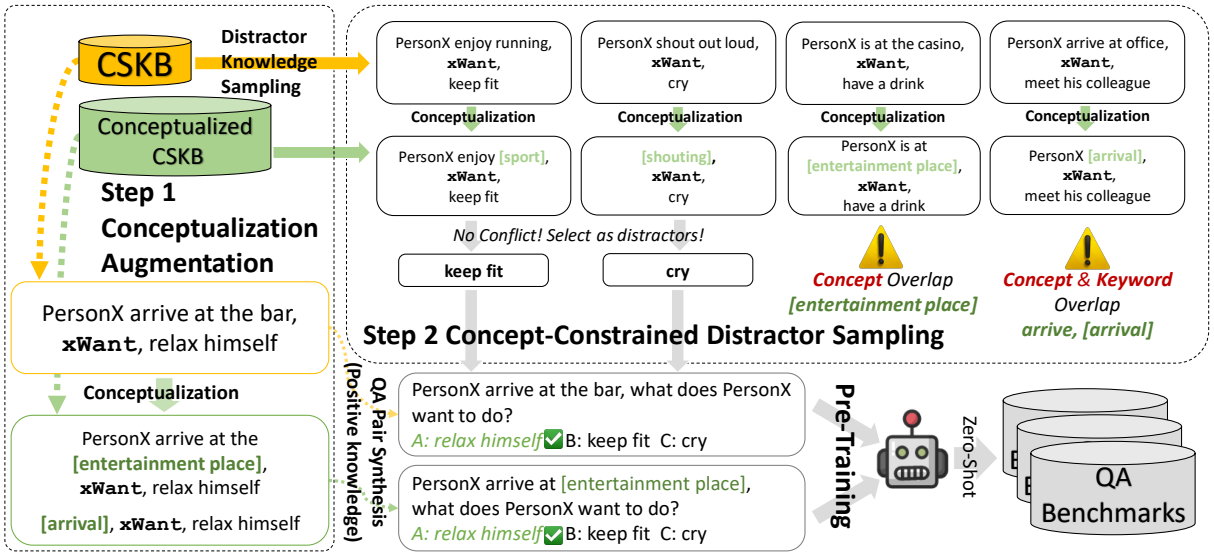


Figure 5.3: An overview of the CAR framework, which shows the process of synthesizing (PersonX arrive at the bar, `xWant`, relax himself) into QA pairs. The triple is conceptualized first, and potential distractor triples are sampled and filtered by keyword and concept overlap. Only those triples that have no overlap are used as distractors.

to train a generative conceptualizer for acquiring D^C.

AbstractATOMIC conceptualizes ATOMIC by identifying instances $i \in h$ via syntactic parsing and replacing it with concept c that is heuristically matched from Probase [62] and WordNet [57] to form a conceptualized head event h_c. Abstract commonsense knowledge, in the form of (h_c, r, t) triple, is collected by connecting the conceptualized head event h_c with its non-abstract counterparts (r, t) from ATOMIC. Human annotations are then conducted to verify the plausibility of the matched conceptualizations and collected abstract knowledge triples. In this chapter, we only use the plausible ones and abandon the implausible counterparts.

5.3.3 Evaluation Benchmarks

We evaluate our framework on the validation split of five commonsense QA benchmarks: Abductive NLI (aNLI; [190]), CommonsenseQA (CSQA; [21]), PhysicalIQA (PIQA; [114]), SocialIQA (SIQA; [80]), and WinoGrande (WG; [191]). These manually constructed benchmarks evaluate various knowledge types essential for robust commonsense reasoning [37].

5.4 The CAR Framework

This section introduces our proposed CAR framework. A general sketch is presented in Figure 5.3. Our framework can be summarized into three steps: (1) Conduct one-step conceptu-

alization inference on existing triples in the CSKB to obtain abstract commonsense knowledge triples. (2) Transfer the triples into QA pairs and generate distractors using keywords and conceptualizations as constraints. (3) Train the QA model using marginal ranking loss.

5.4.1 Conceptualization Augmentation

To incorporate abstract knowledge into the CSKB, we begin by augmenting the $(h, r, t) \in D$ triples by conducting a one-step conceptualization inference. Initially, given a head event h, we retrieve all plausible conceptualizations $C_h = \{c_{i_1,1}, c_{i_1,2}, ...\}$ for all identified instances $i \in \{i_1, i_2, ...|i \in h\}$ using entity-linking heuristics to retrieve concepts from Probase [62] and WordNet [57]. The conceptualized head event h_c is then obtained by replacing an $i \in h$ with one of its retrieved conceptualization $c \in \{c_{i,1}, c_{i,2}, ...\}$. This is done for all identified instances and their retrieved conceptualizations, thereby constructing the set of conceptualized head events of h. Subsequently, we link the non-abstract counterpart (r, t) after h_c to generate candidate abstract knowledge triples (h_c, r, t), where we adopt a discriminator trained with a semi-supervised conceptualization-instantiation framework to determine their plausibility [85]. Only plausible triples are kept to form D^C.

Why abstract knowledge augmentation is preferable to direct instance-level augmentation. An important design choice in CAR is to augment the training resource with conceptualized knowledge rather than only adding more instance-level world knowledge. The main reason is that direct augmentation at the instance level often increases data volume without substantially improving the model's access to reusable structure. Additional concrete triples may enrich lexical coverage, but they also tend to preserve the same sparsity problem: semantically related situations remain distributed across many surface forms, and the model must relearn a similar regularity separately for each phrasing or entity combination. As a result, direct augmentation can improve memorization while offering only limited gains in systematic transfer.

By contrast, abstract knowledge augmentation changes the granularity of supervision. When multiple concrete events are lifted into a shared conceptual representation, the common inferential pattern becomes explicit and can be reused across a broader family of situations. This helps the reasoning model focus on what should remain invariant under substitution, paraphrase, or modest distribution shift. For example, different leisure-related events may map to a concept such as *relaxing event*, making the associated consequences easier to learn as a stable pattern rather than as a collection of disconnected facts. In this sense, conceptualization acts as a form of structural compression: it reduces redundancy among related examples while making higher-

level regularities more visible to the learner.

This does not mean that abstract knowledge should replace all instance-level knowledge. Instance-level supervision remains necessary for grounding, contextual specificity, and lexical realization. The advantage of abstract augmentation is instead that it complements concrete knowledge with a more transferable layer of supervision. In the framework of this thesis, the strongest setting is not "abstract only" but rather a lift-and-ground combination, where concept-level knowledge provides generalizable structure and instance-level knowledge preserves contextual faithfulness. This is precisely why conceptualization-based augmentation is especially effective for robustness-oriented tasks such as commonsense question answering: the model is encouraged to learn not only which answer is correct in one particular case, but also which underlying conceptual regularity makes that answer correct across related cases.

Furthermore, analysis experiments in Section 5.6.1 also support that conceptualization is a better augmentation paradigm compared to several prior knowledge-level augmentation baselines.

5.4.2 Concept-Constrained QA Synthesis

To synthesize a commonsense triple (h, r, t) into a (Q_i, A_i) pair, we first transfer h, r into Q_i by using natural language templates and set t as the ground-truth answer A_1. For example, the triple in Figure 5.3 becomes "PersonX arrives at the bar, what does PersonX want to do?" with the answer being "relax himself." Additional distractors are generated by transforming sampled distractor triples from the original CSKB, where only triples with the same commonsense relation r are sampled to ensure informativeness. To prevent sampling false negative options, we constrain sampling distractor knowledge by filtering keywords and conceptualizations. Formally, denote the keywords of a head event h as $T_h = \{t_1, t_2, \cdots\}$ and the full set of plausible conceptualizations for all identified instances in h as $C_h = \{c_{i_1,1}, c_{i_1,2}, \cdots, c_{i_2,1}, \cdots\}$, we associate a triple (h, r, t) with $T_h + C_h$ to form its constraint. Only knowledge triple (h', r, t') which satisfies $(T_{h'} + C_{h'}) \cap (T_h + C_h) = \varnothing$ can be sampled as a distractor candidate. This constraint requires that the two triples have no common keywords, and their instances cannot be abstracted into the same conceptualization. For example, in Figure 5.3, "(PersonX is at the casino, `xWant`, have a drink)" cannot be used as a distractor triple because "casino" can be conceptualized as "entertainment place," which is the same as "bar" in the original triple. Finally, we sample two distractor triples for the triple (h, r, t) and use the tails of these two triples as the distractors. To guarantee that the abstract commonsense knowledge from our previous augmentation is

learnable by the QA model, we synthesize both the original triple (h, r, t) and its conceptualized versions (h_c, r, t) into QA pairs.

5.4.3 Model Training

We train our QA model by fine-tuning a pre-trained Masked Language Model (MLM) using the Marginal Ranking (MR) loss. Let C represent the original context (if any), Q represent the question, and $(A_1, A_2, ...)$ be the list of options. We first concatenate C, Q, and an answer option A_i together via natural language prompts to generate input sequences $(T_1, T_2, ...)$. For example, the synthesized question with its correct answer in Figure 5.3 will be transformed as: "PersonX arrives at the bar, as a result, PersonX want to, relax himself." We then repeatedly mask out a token at one time and calculate the masked loss. The final MLM score for an input sequence $T \in \{T_1, T_2, ...\}$ with n tokens is:

$$\mathcal{S}(T) = -\frac{1}{n}\sum_{i=1}^{n} \log P(t_i|..., t_{i-1}, t_{i+1}, ...) \tag{5.1}$$

After calculating the scores $S_1, S_2, ...$ for all answer candidates $A_1, A_2, ...$, we compute the marginal ranking loss based on Equation 5.2, where η represents the margin and y is the index of the correct answer.

$$\mathcal{L} = \frac{1}{m}\sum_{i=1, i\neq y}^{m} \max(0, \eta - S_y + S_i) \tag{5.2}$$

During the evaluation phase, we use the same scoring procedure to assign a score to each option and select the one whose concatenated sentence achieves the lowest score as the model's prediction.

5.5 Experiments

5.5.1 Setup

Baselines First, we use random voting (*Random*) and most-frequent labeling (*Majority*) to demonstrate the characteristics of each benchmark. Vanilla RoBERTa-Large [75], and DeBERTa-v3-Large [77] PLMs are used to demonstrate the power of fine-tuning. The performances of these two models under a supervised training regime are also included to show the upper bound of our results. We also include the results of several existing approaches that tackle the same task, including Self-talk [27], COMET-DynaGen [29], SMLM [34], MICO [36], and the previous

Augmentation	Div↑	Exp.Div↑	Plau.↑	%F.Neg.↓	aNLI	CSQA	PIQA	SIQA	WG
N/A *(Baseline)*	N/A	N/A	88.0	45.7	76.0	67.0	78.0	62.1	76.0
EDA [185]	8.10	4.67	9.33	33.0	76.5	65.6	76.6	61.4	74.9
Word2Vec [186]	11.8	4.00	9.00	55.0	74.3	65.8	75.1	62.9	74.7
GLOVE [186]	8.21	6.67	4.67	44.3	74.7	64.2	74.6	61.1	74.4
BERT-base [198]	0.81	8.33	14.3	41.7	70.4	63.9	72.4	63.5	61.0
Synonym [187]	6.92	11.0	5.67	45.0	75.5	64.9	74.5	62.5	75.7
GPT3-distil [42]	35.6	24.3	**95.7**	42.7	75.4	**71.8**	75.6	63.4	76.0
Conceptualization (Ours)	**48.5**	**37.0**	90.0	**22.7**	**79.6**	69.3	**78.6**	**64.0**	**78.2**

Table 5.1: Comparison results (%) of different augmentation methods against conceptualization. N/A stands for not using any augmentation. Plau. is the expert-evaluated ratio of plausible augmented knowledge, %F.Neg. represents the expert-annotated proportion of false negative options. Div. and Exp.Div. are diversities measured by embedding similarity and expert annotated knowledge coverage. Performances on the right refer to accuracies achieved by the QA model trained on data augmented by each method. The best performances are **bold-faced**.

state-of-the-art STL-Adapter [37]. Most importantly, we compare our framework with previous baseline [35] to validate the efficacy of conceptualization since both methods share similar model architecture and training procedures. Both RoBERTa-Large and DeBERTa-v3-Large are used as the backbones for fair comparisons. There are, in total, 534,833 synthetic QA pairs [35].

With the recent advances in Large Langauge Models (LLMs) [192–194], we also benchmark the performances of GPT3.5 [32] and ChatGPT [40] as baselines. We prompt the LLM directly in a zero-shot setting, where no in-context learning [195] or chain-of-thought reasoning [196] are applied. For every QA entry, the LLM is presented with a question, several choices, and a natural language command that asks it to choose the index of the correct answer directly [197]. We then parse the generated outputs to obtain the "predictions" of LLM by using meticulously designed rules and compare them with the ground-truth labels. With this approach, we ensure that the input given to LLMs is mostly identical to that provided to models trained within our framework.

Implementation Details We use accuracy as the evaluation metric and compare our framework with the following baseline methods. For conceptualization, we leverage an off-the-shelf conceptualizer [85], which is a semi-supervised conceptualization discriminator fine-tuned on labeled conceptualization data from AbstractATOMIC and unlabeled data from ATOMIC. We use a plausibility score $T = 0.9$ to filter out plausible conceptualizations, which results in 440K conceptualization-aided synthetic QA pairs for training. We employ an AdamW optimizer [199] with a learning rate of 7e-6 and a max sequence length of 128 to accommodate QA pairs with different lengths. We select the best checkpoint according to the highest accuracy achieved on

the synthetic validation QA set. Each experiment is repeated using three different random seeds, and the average performance is reported. The model is warmed up with 5% of total iterations and evaluated every 1000 global steps, while the margin η for the marginal ranking loss is set to 1.

5.5.2 Results

The main results are reported in Table 5.5. For the baselines, DeBERTa-v3-Large (MR) trained on ATOMIC achieves the best performance, followed by ChatGPT. Both achieve an accuracy of more than 70% on average. Our best system, based on DeBERTa-v3-Large and trained on our conceptualization-augmented ATOMIC, achieves state-of-the-art results and significantly outperforms all PLM-based baselines on every benchmark, and can advance the average accuracy by 2.1% compared with the same baseline model. It also significantly surpasses the performance of the same model that is trained on ATOMIC-10X with only 10% amount of data. Notably, compared with LLMs, our system champions three benchmarks and performs better on average with a 3.7% leap. This indicates that supervision signals from CSKBs are important for downstream applications, and CSKBs aided by conceptualization can significantly enhance this process. Moreover, as an ablation, we study the role of concept-level distractor sampling by discarding conceptualization augmentation and only training the models on ATOMIC, synthesized to QA format with our proposed constraint technique. Comparing the results in Table 5.5, it can be observed that the concept-level distractor sampling improves the average performance by approximately 1.5%. This demonstrates that our proposed technique is effective, and generating distractors with a stronger positive knowledge recall is helpful in synthesizing QA pairs that are both fair and informative.

5.6 Analysis and Discussion

In this section, we study the effects of conceptualization and the reasons contributing to CAR's success. First, we conduct expert evaluations on the synthetic QA pairs to study the quality and diversity of different CSKB augmentation methods in comparison with conceptualization. Second, we conduct training dynamics [182] analysis to show that conceptualization-aided QA pairs can provide more *ambiguous* examples helpful for training. Finally, we study the impact of filtering ATOMIC-10X with different critic thresholds, the ablations of CAR, and the effect of conceptualization from an out-of-domain generalization perspective.

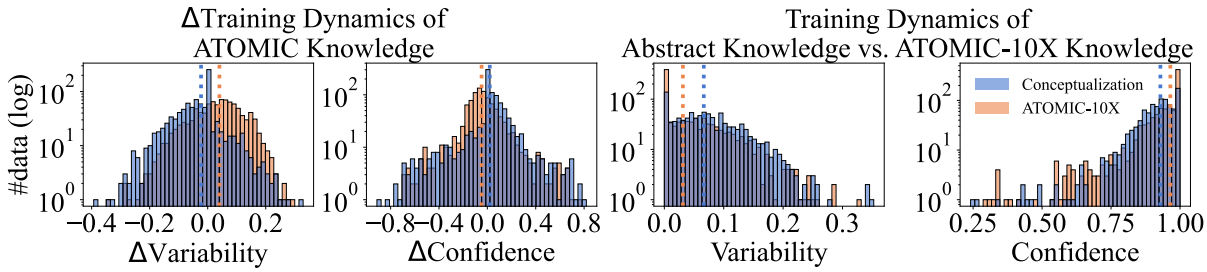


Figure 5.4: Analyses on training dynamics of different knowledge. The dotted lines refer to the median values.

5.6.1 Comparisons with Data Augmentations

To demonstrate the effectiveness of our conceptualization-based augmentation, we conduct comprehensive comparisons with alternative data augmentation methods that aim to expand the semantic coverage of commonsense knowledge bases (CSKBs) in a manner analogous to conceptualization. We consider the following baselines: EDA, word-embedding-based substitutions (Word2Vec [200] and GloVe [201]), contextual-embedding-based substitutions (BERT [74]), and synonym-based substitutions (WordNet [57]). For fair comparison, for each ATOMIC triple with head event h, we only augment the identified instance token $i \in h$, and we generate the same number of augmented siblings as the number of valid conceptualizations $|C_h|$. In addition, we include LLM-distilled knowledge from ATOMIC-10X [42] as another baseline augmentation source by randomly sampling a matched number of triples from ATOMIC-10X and merging them into ATOMIC (details in Section 5.6.2).

We analyze these augmentation methods along three dimensions: (i) diversity of augmented knowledge, (ii) quality of synthesized QA pairs, and (iii) zero-shot commonsense QA performance. We recruit three expert annotators (undergraduate or graduate students actively involved in commonsense research). They exhibit strong agreement, with an inter-annotator agreement of 83% in pairwise agreement and a Fleiss' Kappa [202] of 0.64, comparable to the 0.62 reported in prior work [35].

Diversity. We first study whether each augmentation method introduces knowledge that is meaningfully new relative to the original training set. For each ATOMIC triple, we compute the average cosine similarity between the triple and its augmented siblings using SentenceBERT embeddings [203]. For ATOMIC-10X, we treat the sampled triples as augmentations. We define the complement of the average similarity (aggregated over all triples) as an automatic diversity score (Div.). In parallel, we retrieve the top-10 most similar ATOMIC triples for each augmented triple (by SentenceBERT similarity) and ask experts to judge whether the augmented triple is

semantically covered by those retrievals. We define expert-evaluated diversity as the ratio of uncovered triples among 300 samples. As shown in Table 5.1, conceptualization achieves the best performance on both diversity metrics, indicating that the induced abstract knowledge is diverse and largely absent from existing CSKBs, thereby expanding coverage.

Quality of Synthetic QA Pairs. Next, we transform augmented triples into synthetic QA pairs, using the head-event keywords as constraints and the augmented knowledge as the target commonsense signal. We sample 300 QA pairs per method and ask the same experts to annotate (i) whether the ground-truth answer is correct and (ii) whether distractors are also plausible given the augmented head event (i.e., whether they form false-negative distractors). These annotations measure the plausibility ratio of augmented knowledge and the proportion of QA pairs containing false-negative distractors. Table 5.1 shows that many baseline augmentations produce implausible knowledge and do not effectively improve distractor quality. In contrast, conceptualization remains highly plausible and substantially reduces false-negative distractors. Expert annotators reach 86% accuracy on 300 randomly sampled QA pairs, exceeding the 80% reported by the baseline QA synthesis pipeline [35].

Zero-shot Commonsense QA Performance. Finally, we train DeBERTa-v3-Large models on QA pairs synthesized from the union of original ATOMIC and augmented triples produced by each method. We use only the head-event keywords as constraints. Models are trained with the marginal ranking loss (Section 5.4.3) and evaluated in a zero-shot manner on five commonsense QA benchmarks. Results in Table 5.1 show that conceptualization outperforms all other augmentation methods on average and consistently improves zero-shot commonsense reasoning.

Comparison with ATOMIC-10X (augmentation view). ATOMIC-10X contains a large amount of machine-distilled commonsense knowledge, and therefore appears promising as an augmentation source. However, despite its scale and apparent diversity, Table 5.1 shows that models trained with ATOMIC-10X augmentation do not reliably improve. A plausible explanation is the prevalence of false-negative distractors induced by overly general and versatile tail events distilled from GPT-3, which can apply to many heads and thus confound the ranking-based objective.

5.6.2 ATOMIC-10X Usage and Additional Experiments

ATOMIC-10X is a machine-generated corpus developed by [42] via symbolic knowledge distillation from large language models such as GPT-3 [32]. In brief, GPT-3 is prompted with ATOMIC-style head events and relations to generate tail events; a student model is trained to produce symbolic graphs, and a critic model is used to score and filter generated knowledge. ATOMIC-10X substantially increases scale relative to human-curated ATOMIC2020 [129].

We use ATOMIC-10X in two scenarios and additionally study the effect of critic-threshold filtering.

Scenario I: training solely on ATOMIC-10X. We train QA models using only ATOMIC-10X (without integrating ATOMIC or AbstractATOMIC), following the original ATOMIC-10X train/dev/test splits. We synthesize QA pairs using the same pipeline as [35]. For each triple, we extract head-event keywords by lemmatizing tokens and removing common subjects, prepositions, and stopwords. To control knowledge quality, we filter ATOMIC-10X with multiple critic thresholds (0.9, 0.8, 0.7, 0.5), and we also include training on the unfiltered set. We then evaluate the resulting QA models on five zero-shot commonsense QA benchmarks (Table 5.4). Overall, even with high critic thresholds, training on ATOMIC-10X does not yield consistent improvements beyond marginal gains, and it typically fails to surpass training on ATOMIC. This suggests that performance is not determined by raw scale alone, but rather by the diversity and reliability of knowledge; in this regard, human-annotated ATOMIC remains competitive.

Scenario II: augmenting ATOMIC with ATOMIC-10X. We also use ATOMIC-10X as an augmentation source for ATOMIC, matching the number of sampled ATOMIC-10X triples to the total number of plausible abstract commonsense triples in AbstractATOMIC. We filter ATOMIC-10X with critic thresholds (0.9, 0.8, 0.7, 0.5) before sampling, merge the sampled triples into ATOMIC, and synthesize QA pairs from the combined pool. In this setting, distractors may originate from both ATOMIC and ATOMIC-10X. We train and evaluate QA models as above; the best ATOMIC-10X augmentation configuration (DeBERTa-v3-Large with a critic threshold of 0.8) is the variant reported in Table 5.1, and full results are reported in Table 5.4. We find that ATOMIC-10X can occasionally improve a particular benchmark, but it does not reliably improve the average performance across benchmarks, which we treat as a closer proxy for generalizable commonsense reasoning. A likely reason is that ATOMIC-10X may contain noise that is not beneficial for zero-shot commonsense QA, consistent with observations in prior work [204]. In contrast, conceptualization resolves these issues and yields consistent

gains across benchmarks.

5.6.3 Training Dynamics Analysis

Training dynamics provide a lens into a model's confidence and variability for individual training instances when learning from large-scale data. In the context of QA, we interpret *confidence* as the model's certainty in ranking the ground-truth option above distractors, and *variability* as the fluctuation of this confidence over training time. Such signals help explain how different knowledge sources affect optimization and generalization.

In this section, we examine the impact of abstract commonsense knowledge (conceptualization) versus GPT-3-distilled knowledge (ATOMIC-10X) through training dynamics on two sets of data. We train three QA models on synthetic QA pairs from (i) conceptualization-augmented ATOMIC, (ii) ATOMIC-10X-augmented ATOMIC, and (iii) the original ATOMIC (baseline).

Dynamics on the same ATOMIC QA pairs (optimization effect). We randomly select 1,000 QA pairs synthesized from the original ATOMIC and compute their training dynamics under all three models. As shown on the left of Figure 5.4, conceptualization augmentation reduces the average variability while increasing confidence on these ATOMIC instances, suggesting that abstract knowledge can make the original ATOMIC knowledge easier to learn. In contrast, ATOMIC-10X augmentation exhibits the opposite trend, indicating potential interference during learning.

Dynamics on abstract vs. ATOMIC-10X knowledge (generalization-oriented effect). We also compute training dynamics on 1,000 QA pairs synthesized from abstract knowledge (conceptualizations) and another 1,000 from ATOMIC-10X. The rightmost plots in Figure 5.4 show that, compared to ATOMIC-10X, abstract knowledge tends to be more ambiguous: it induces higher variability and lower confidence. Prior work on data maps [182] suggests that such ambiguous instances contribute disproportionately to out-of-distribution (OOD) generalization. Consistent with this hypothesis, we observe that conceptualization is superior to ATOMIC-10X: it both aids optimization on original ATOMIC knowledge and injects ambiguity that is beneficial for downstream OOD benchmarks.

Training-dynamic definitions (adapted to marginal ranking loss). Training dynamics were introduced by [182] to analyze a model's per-instance behavior via confidence and variability over epochs. Because our QA models are trained with marginal ranking loss (Section 5.4.3),

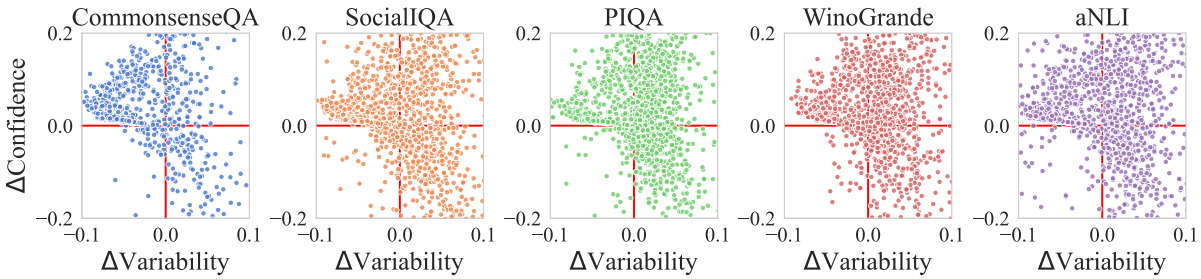


Figure 5.5: The change of training dynamics on various commonsense QA benchmarks by a DeBERTa-v3-Large model trained on abstract commonsense knowledge injected ATOMIC (ours) compared with the one trained only on ATOMIC [35].

they do not directly output class probabilities. Instead, we define confidence using the model's score gap between the ground-truth option and distractors.

Let n be the number of saved checkpoints within an epoch, and let a QA instance be (Q_i, A_i) with m options $A_i = \{A_{i,1}, \ldots, A_{i,m}\}$ where the ground-truth option is $A_{i,j}$. Let $S^c_{i,d}$ be the model score for option $A_{i,d}$ at checkpoint c, and let $\sigma(\cdot)$ be the sigmoid function. We define confidence as:

$$\mathcal{C}(Q_i, A_i) = \frac{1}{n}\sum_{c=1}^{n}\sigma\left(\frac{1}{m-1}\sum_{d\neq j}\left(S^c_{i,j} - S^c_{i,d}\right)\right). \tag{5.3}$$

Intuitively, this quantity averages (over checkpoints) the mean score gap between the ground-truth option and distractors; a larger gap implies higher confidence in ranking the correct answer above distractors.

Variability follows [182] and is computed as the standard deviation of the (sigmoid-transformed) score-gap signal across checkpoints:

$$\mathcal{V}(Q_i, A_i) = \sqrt{\frac{1}{n}\sum_{c=1}^{n}\left(\sigma\left(\frac{1}{m-1}\sum_{d\neq j}\left(S^c_{i,j} - S^c_{i,d}\right)\right) - \mathcal{C}(Q_i, A_i)\right)^2}. \tag{5.4}$$

Revisiting Figure 5.4 with these definitions, we find that injecting abstract commonsense knowledge increases confidence and reduces variability on ATOMIC instances, whereas ATOMIC-10X induces a reversed trend. Moreover, abstract knowledge is more ambiguous (higher variability, lower confidence) than ATOMIC-10X, aligning with the hypothesis that ambiguity contributes to better OOD generalization [182]. We additionally plot training-dynamics changes on downstream QA benchmarks (Figure 5.5) and observe that abstract knowledge increases model confidence on these benchmarks, providing further evidence for improved generalization.

Models	aNLI	CSQA	PIQA	SIQA	WG
CAR (RoBERTa)	72.7	66.3	73.2	64.0	62.0
⋄ w/o CA	72.3	64.8	73.2	64.8	61.3
⋄ w/o CCQS	71.5	67.3	72.1	61.8	62.7
CAR (DeBERTa)	79.6	69.3	78.6	64.0	78.2
⋄ w/o CA	78.9	67.2	78.6	63.8	78.1
⋄ w/o CCQS	78.2	68.1	78.1	63.5	78.3

Table 5.2: Ablation study on two components of CAR. CA stands for Conceptualization Augmentation, and CCQS stands for Concept-Constrained QA Synthesis. The following five columns denote the accuracy (%) on each benchmark.

5.6.4 Ablation Study

We ablate two critical components of the CAR framework relative to traditional zero-shot QA systems [35]:

• **Conceptualization Augmentation (CA):** We augment the original CSKB with conceptualizations to derive abstract commonsense knowledge, and synthesize QA pairs from this augmented CSKB. Without CA, abstract knowledge is not incorporated; conceptualizations are only used as constraints during QA synthesis, resembling applying our QA synthesis protocol directly to ATOMIC.

• **Concept-Constrained QA Synthesis (CCQS):** We constrain distractor generation by requiring that distractors' head events share neither keywords nor conceptualizations with the question. Without CCQS, the constraint is weakened to keyword-only exclusion, which can introduce more false-negative distractors.

We train RoBERTa-Large and DeBERTa-v3-Large models while dropping one component at a time. Results in Table 5.2 show that both components matter, with CCQS contributing more on average, highlighting the importance of eliminating false-negative distractors and demonstrating that conceptualization is effective for achieving this goal.

5.6.5 The Effect of Conceptualization on Generalization

Finally, we examine whether conceptualization particularly helps on benchmark questions that are semantically distant from the ATOMIC training set. For each benchmark, we compute the average BERTScore [168] between each question and the ATOMIC training split, and we split questions into "Difficult" (low semantic overlap) and "Easy" (high overlap). We then compare two QA models trained following [35]: one trained on conceptualization-augmented ATOMIC

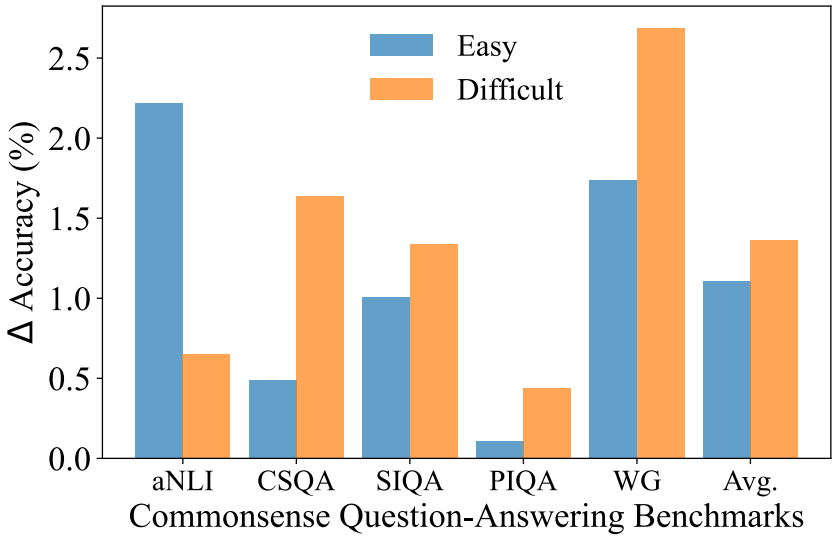


Figure 5.6: Comparison of accuracy improvement (%) with/without conceptualization-augmentation for two groups of QA entries across five benchmarks. Avg. stands for averaging across all benchmarks.

and the other trained on ATOMIC only. Figure 5.6 shows that conceptualization yields larger gains on semantically distant questions across multiple benchmarks, indicating improved generalizability to out-of-distribution queries.

5.6.6 Generalization to Other CSKBs

While our main experiments use AbstractATOMIC as the conceptualization source for ATOMIC, we also evaluate whether CAR transfers to other CSKBs. Following [35], we experiment on CWWV, a dataset that combines multiple CSKBs including ConceptNet [58], WordNet [57], and Wikidata [205]. We use two flexible generative conceptualizers: (i) a GPT-2-based conceptualization generator [91] and (ii) ChatGPT [40]. Generated conceptualizations are converted into abstract knowledge and integrated into CWWV, producing an augmented CSKB used to train a zero-shot commonsense QA reasoner under CAR.

Table 5.3 shows a modest but consistent improvement (about 1% on average) over baselines that leverage CWWV, demonstrating that CAR can incorporate conceptualizations from other CSKBs and transfer beyond ATOMIC. In future work, it would be valuable to explore automatic construction of conceptualization resources for additional CSKBs and further investigate their benefits for general commonsense reasoning.

Model	CSKB	a-NLI	CSQA	PIQA	SIQA	WG	Avg.
RoBERTa-L (MR) [35]	CWWV	70.0	67.9	72.0	54.8	59.4	64.8
MTL [37]	CWWV	69.6	67.3	72.5	52.0	57.2	63.7
ZS-Fusion [37]	CWWV	69.6	67.6	73.1	53.7	59.5	64.7
CAR-RoBERTa-L (Ours)	CWWV^C	71.6	68.4	73.0	55.4	60.6	65.8
GPT-3.5 (`text-davinci-003`)	N/A	61.8	68.9	67.8	68.0	60.7	65.4
ChatGPT (`gpt-3.5-turbo`)	N/A	69.3	**74.5**	75.1	69.5	62.8	70.2

Table 5.3: Zero-shot evaluation results (%) on five commonsense question answering benchmarks by models trained on the CWWV dataset. CWWV^C refers to the augmented CWWV dataset using generated conceptualizations from a trained GPT2 generator and ChatGPT.

5.7 Conclusions

In this chapter, we present CAR, a pioneering framework for zero-shot commonsense QA empowered by conceptualization. Our approach surpasses even large language models on five QA benchmarks, achieving state-of-the-art performance on average. Our analyses reveal that conceptualization can improve the sampling of negative examples, and abstract knowledge is more helpful compared with those distilled from GPT3 as it provides more ambiguous knowledge to support OOD generalization. These findings demonstrate the substantial benefits of introducing conceptualization and abstract knowledge into zero-shot commonsense reasoning. This chapter underscores the critical role of conceptualization in achieving generalizable commonsense reasoning, setting the stage for further advancements in this field.

Model	CSKB	Critic	a-NLI	CSQA	PIQA	SIQA	WG	Avg.
Backbone: RoBERTa-Large *340M*								
RoBERTa-L (MR)	ATOMIC	N/A	70.8	64.2	72.1	63.1	59.2	65.9
RoBERTa-L (MR)	ATM-10X	0.9	69.6	58.1	72.3	58.3	57.2	63.1
RoBERTa-L (MR)	ATM-10X	0.8	70.1	58.9	71.5	58.2	57.7	63.3
RoBERTa-L (MR)	ATM-10X	0.7	70.8	59.4	72.1	58.5	58.3	63.8
RoBERTa-L (MR)	ATM-10X	0.5	68.7	56.8	71.7	58.4	60.1	63.1
RoBERTa-L (MR)	ATM-10X	0.0	70.7	58.3	71.7	58.2	57.5	63.3
RoBERTa-L (MR)	$\text{ATM}^{\text{ATM-10X}}$	0.9	71.7	66.3	73.2	62.8	60.7	66.9
RoBERTa-L (MR)	$\text{ATM}^{\text{ATM-10X}}$	0.8	71.8	66.0	73.2	61.7	59.5	66.4
RoBERTa-L (MR)	$\text{ATM}^{\text{ATM-10X}}$	0.7	71.6	65.6	72.9	62.2	59.8	66.4
RoBERTa-L (MR)	$\text{ATM}^{\text{ATM-10X}}$	0.5	72.0	65.4	72.9	62.0	60.5	66.6
RoBERTa-L (MR)	$\text{ATM}^{\text{ATM-10X}}$	0.0	71.6	66.3	73.3	62.9	61.0	67.0
CAR-RoBERTa-L (Ours)	ATOMIC	N/A	72.3	64.8	73.2	64.8	61.3	67.3
CAR-RoBERTa-L (Ours)	ATM^{C}	N/A	72.7	66.3	73.2	64.0	62.0	67.6
Backbone: DeBERTa-v3-Large *435M*								
DeBERTa-v3-L (MR)	ATOMIC	N/A	76.0	67.0	78.0	62.1	76.0	71.8
DeBERTa-v3-L (MR)	ATM-10X	0.9	74.5	70.8	78.9	59.7	72.2	71.2
DeBERTa-v3-L (MR)	ATM-10X	0.8	74.2	70.6	**79.5**	59.2	70.7	70.8
DeBERTa-v3-L (MR)	ATM-10X	0.7	74.6	69.9	79.3	60.0	70.2	70.8
DeBERTa-v3-L (MR)	ATM-10X	0.5	74.1	70.4	78.8	58.9	70.1	70.5
DeBERTa-v3-L (MR)	ATM-10X	0.0	75.1	71.6	79.0	59.7	71.7	71.4
DeBERTa-v3-L (MR)	$\text{ATM}^{\text{ATM-10X}}$	0.9	75.4	71.3	73.4	61.7	75.3	71.4
DeBERTa-v3-L (MR)	$\text{ATM}^{\text{ATM-10X}}$	0.8	75.4	71.8	75.6	63.4	76.0	72.4
DeBERTa-v3-L (MR)	$\text{ATM}^{\text{ATM-10X}}$	0.7	74.9	71.2	77.4	61.8	76.2	72.3
DeBERTa-v3-L (MR)	$\text{ATM}^{\text{ATM-10X}}$	0.5	74.8	71.2	77.1	61.7	75.7	72.1
DeBERTa-v3-L (MR)	$\text{ATM}^{\text{ATM-10X}}$	0.0	76.2	71.0	75.8	62.8	75.8	72.3
CAR-DeBERTa-v3-L (Ours)	ATOMIC	N/A	78.9	67.2	78.6	63.8	78.1	73.3
CAR-DeBERTa-v3-L (Ours)	ATM^{C}	N/A	**79.6**	69.3	78.6	64.0	**78.2**	**73.9**
Large Language Models								
GPT-3.5 (`text-davinci-003`)	N/A	N/A	61.8	68.9	67.8	68.0	60.7	65.4
ChatGPT (`gpt-3.5-turbo`)	N/A	N/A	69.3	**74.5**	75.1	**69.5**	62.8	70.2

Table 5.4: Zero-shot evaluation results (%) on five commonsense question answering benchmarks using different critic thresholds for filtering ATOMIC-10X. The best results are **boldfaced**, and the second-best ones are underlined. ATM^{C} stands for the ATOMIC with abstract commonsense knowledge injected. ATM-10X stands for using ATOMIC-10X [42] as the source CSKB D. $\text{ATM}^{\text{ATM-10X}}$ indicates the ATOMIC with sampled knowledge from ATOMIC-10X injected. Critic indicates the lower bound for filtering knowledge from ATOMIC-10X, which means that only knowledge with a critic score above the threshold will be selected.

Model	CSKB	a-NLI	CSQA	PIQA	SIQA	WG	Avg.
Random	-	50.0	20.0	50.0	33.3	50.0	40.7
Majority	-	50.8	20.9	50.5	33.6	50.4	41.2
GPT2-L [91]	-	56.5	41.4	68.9	44.6	53.2	52.9
RoBERTa-L [75]	-	65.5	45.0	67.6	47.3	57.5	56.6
DeBERTa-v3-L [77]	-	59.9	25.4	44.8	47.8	50.3	45.6
Self-talk [27]	-	-	32.4	70.2	46.2	54.7	-
COMET-DynGen [29]	ATOMIC	-	-	-	50.1	-	-
SMLM [34]	*	65.3	38.8	-	48.5	-	-
Backbone: RoBERTa-Large *340M*							
RoBERTa-L (Vanilla) [75]	-	65.5	45.0	67.6	47.3	57.5	56.6
MICO [36]	ATOMIC	-	44.2	-	56.0	-	-
RoBERTa-L (MR) [35]	ATM_{10X}	70.8	64.2	71.7	61.0	60.7	65.7
RoBERTa-L (MR) [35]	ATOMIC	70.8	64.2	72.1	63.1	59.2	65.9
RoBERTa-L (MR) [35]	CWWV	70.0	67.9	72.0	54.8	59.4	64.8
RoBERTa-L (MR) [35]	CSKG	70.5	67.4	72.4	63.2	60.9	66.8
STL-PLM [37]	ATOMIC	71.6	64.0	72.2	63.2	60.5	66.3
MTL [37]	CWWV	69.6	67.3	72.5	52.0	57.2	63.7
MTL [37]	CSKG	69.8	67.1	72.0	61.9	59.3	66.0
STL-Adapter [37]	ATOMIC	71.3	66.5	71.1	64.4	60.3	66.7
STL-Adapter [37]	CSKG	71.5	66.7	72.1	64.7	59.0	66.8
ZS-Fusion [37]	CWWV	69.6	67.6	73.1	53.7	59.5	64.7
ZS-Fusion [37]	CSKG	72.4	68.3	73.0	66.7	60.9	68.3
MKIF [39]	CSKG	72.5	<u>71.0</u>	73.1	-	61.0	-
CAR-RoBERTa-L (Ours)	ATOMIC	72.3	64.8	73.2	64.8	61.3	67.3
CAR-RoBERTa-L (Ours)	ATM^C	72.7	66.3	73.2	64.0	62.0	67.6
Backbone: DeBERTa-v3-Large *435M*							
DeBERTa-v3-L (MR) [35]	ATM_{10X}	74.0	65.4	73.8	59.5	73.9	69.3
DeBERTa-v3-L (MR) [35]	ATOMIC	76.0	67.0	<u>78.0</u>	62.1	76.0	71.8
CAR-DeBERTa-v3-L (Ours)	ATOMIC	<u>78.9</u>	67.2	**78.6**	63.8	<u>78.1</u>	<u>73.3</u>
CAR-DeBERTa-v3-L (Ours)	ATM^C	**79.6**	69.3	**78.6**	64.0	**78.2**	**73.9**
Large Language Models							
GPT-3.5 (`text-davinci-003`)	-	61.8	68.9	67.8	<u>68.0</u>	60.7	65.4
ChatGPT (`gpt-3.5-turbo`)	-	69.3	**74.5**	75.1	**69.5**	62.8	70.2
Supervised Learning & Human Performance							
RoBERTa-L (Supervised)	-	85.6	78.5	79.2	76.6	79.3	79.8
DeBERTa-v3-L (Supervised)	-	89.0	82.1	84.5	80.1	84.1	84.0
Human Performance	-	91.4	88.9	94.9	86.9	94.1	91.2

Table 5.5: Zero-shot evaluation results (%) on five commonsense question answering benchmarks with baselines trained on multiple CSKBs. The best results are **bold-faced**, and the second-best ones are <u>underlined</u>. ATM^C stands for the ATOMIC with abstract commonsense knowledge injected and ATM_{10X} stands for ATOMIC-10X [42]. All baseline results are consistent with their original papers. CWWV refers to the combination of ConceptNet [58], VisualGenome [206], WikiData [205], and WordNet [57]. CSKG [207] consists of ATOMIC [128] and CWWV.

CHAPTER 6

SCALABLE CONCEPTUALIZATION DISTILLATION FOR INFINITE COMMONSENSE KNOWLEDGE ACQUISITION

The previous chapters positioned *conceptualization* as a unifying lens for generalizable commonsense reasoning and progressively made that lens operational. At a high level, we cast commonsense reasoning as a *lift-and-ground* loop: lift concrete events to concepts to expose reusable regularities, then ground those concepts back into concrete contexts where they can support inference. Methodologically, we asked two questions: *(i) can this loop be learned at scale under scarce supervision while remaining contextualized?* and *(ii) does conceptualization materially improve downstream reasoning within realistic pipelines?* CAT addressed (i) by jointly modeling conceptualization and instantiation in a semi-supervised pipeline, learning plausibility-aware operators from abundant unlabeled CSKB data. CAR addressed (ii) by showing that even a single conceptualization step can improve robustness in zero-shot commonsense QA via coverage expansion and reduced false-negative distractors.

This chapter asks the next question implied by those results: *can we turn the conceptualization–instantiation loop into a scalable knowledge acquisition primitive, rather than a one-off augmentation?* In principle, iterating the loop can extrapolate a relatively small CSKB into a richer space of event variations and inferential consequences: conceptualization induces abstractions from a concrete triple; instantiation proposes diverse grounded realizations; and the accepted realizations can be fed back as new inputs for further abstraction. In practice, two constraints dominate. First, the chain must remain *contextualized*, since the validity of an abstraction or instantiation depends on the original relational context. Second, it must be *scalable*, since naïvely verifying multi-step expansions with human annotation is quickly prohibitive.

We therefore introduce **CANDLE**, a **C**onceptu**A**lization and I**N**stantiation **D**istillation framework from **L**arge Language Mod**E**ls (LLMs). CANDLE uses strong LLM teachers to sequentially generate contextualized conceptualizations and contextualized instantiations conditioned on an original CSKB triple, applies critic models to filter low-quality generations, and re-injects accepted instantiations back into the CSKB to enable iterative expansion. In this way, CANDLE

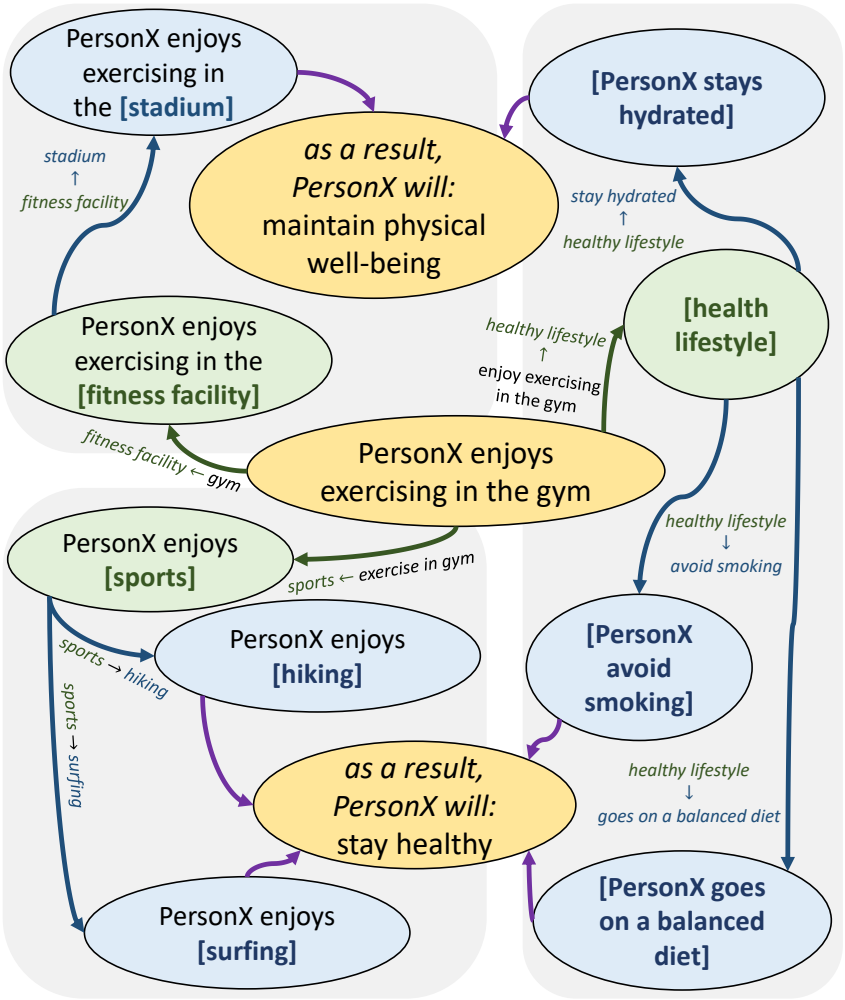


Figure 6.1: Examples showing several chains of **conceptualization** and **instantiation** over the event *PersonX enjoys exercising in the gym*. **New inferential commonsense knowledge** can be induced when placing the **instantiation** back into the **original context**.

directly builds on CAT's emphasis on contextualized plausibility and bidirectional modeling, while extending CAR's downstream motivation from "conceptualization helps" to "the full lift-and-ground loop can be distilled and scaled" to produce large pools of high-quality abstract and grounded commonsense knowledge for training and evaluation.

6.1 Introduction

Commonsense reasoning refers to the cognitive ability to make logical inferences and draw conclusions based on general knowledge and understanding of the world that is typically shared among individuals [5, 7]. However, a longstanding challenge is generalizability, as commonsense reasoning often necessitates applying knowledge to novel situations beyond simple pattern recognition or memorizing all special cases [208, 209]. One promising approach to address this

is the chain of conceptualization [103] and instantiation [134], which, akin to the process of conceptual induction and deduction in human reasoning [108], involves conceptualizing instances derived from known commonsense knowledge and subsequently instantiating these concepts in new situations to obtain the knowledge required for downstream reasoning. For example, in Figure 6.1, one can first conceptualize *enjoys exercising in the gym* as a *healthy lifestyle*, and then further instantiate it to *go on a balanced diet*. This process allows for the derivation of a novel event, *PersonX goes on a balanced diet*, which may entail new commonsense knowledge when connected with the original event's commonsense inferential tail. By possessing substantial knowledge to initiate the process of conceptualization and instantiation, one can extrapolate limited commonsense knowledge to a wide array of diverse scenarios.

Yet, replicating this fundamental ability on machines remains challenging due to the absence of both types of knowledge in widely used CommonSense Knowledge Bases (CSKBs [24, 58, 128, 210]). Language models that are fine-tuned on these CSKBs are, therefore, unable to effectively utilize such a chain for reasoning during inference. To compensate, various methods compensating the lack of conceptualization ability of language models have been proposed for entity-level [68, 69, 88, 136–138] and event-level [78, 85, 139] conceptualizations by matching against concept taxonomies like Probase [62] and WordNet [57]. Meanwhile, [135] address instantiation through controllable generation. However, several limitations still persist.

Firstly, despite the importance of both conceptualization and instantiation, most existing works underestimate the importance of the second step while focusing solely on conceptualization and using the resulting abstract knowledge directly. Other studies that concentrate on instantiations either overlook the conceptualization step entirely or only retrieve instances from the original CSKB, failing to introduce novel entities and events. Secondly, most conceptualization methods heavily depend on matching instances with concepts in concept taxonomies, such as Probase and WordNet, which have a limited scope and lack contextual information. Consequently, the derived conceptualizations are constrained in scale by these taxonomies and are formulated without considering proper contextualization, necessitating further verification in the original context. Lastly, the chain of conceptualization and instantiation can easily bring more than two orders of magnitude of data on top of the original CSKB. However, current acquisition and verification methods for both steps heavily rely on human annotation, which can be extremely costly as the scale of the CSKB increases.

To address these gaps, we introduce CANDLE, a **C**onceptu**A**lization and I**N**stantiation **D**istillation framework from **L**arge Language Mod**E**ls (LLMs) to aid commonsense reasoning. Specifically, CANDLE marks the first to complete the chain of conceptualization and instantiation by in-

structing powerful LLMs to sequentially generate both types of knowledge based on concrete commonsense triples while carefully considering the original context throughout the process. We further alleviate the human annotation cost by employing two critic filtering models to eliminate low-quality generations. The instantiated knowledge, representing concrete commonsense knowledge again, can be fed back into CANDLE as input, iteratively augmenting the original CSKB significantly.

By applying CANDLE to ATOMIC [128], we construct a large-scale knowledge base comprising 6.18 million conceptualizations and instantiations from two powerful LLMs, ChatGPT [40] and LLAMA2 [97]. We demonstrate the intrinsic efficacy of CANDLE through automatic and human evaluations, highlighting the ability to generate high-quality and diverse knowledge (Section 6.5.1). We further show the extrinsic benefits of CANDLE by leveraging the generated knowledge as complementary training data to distill student models that yield improvements across three downstream tasks, including CSKB conceptualization, generative commonsense inference, and zero-shot commonsense question answering (Section 6.5.2).

6.2 Related Works

6.2.1 Conceptualization and Instantiation

Conceptualization aims to abstract a set of entities or events into a general concept, thereby forming abstract commonsense knowledge within its original context [103]. Subsequently, instantiation grounds the derived concept into other instances and events to introduce new commonsense knowledge. Existing works primarily focused on entity-level conceptualization [68, 69, 88, 136, 142], with [78] pioneering the construction of an event conceptualization benchmark by extracting concepts for social events from WordNet [57] synsets and Probase [62]. [85, 111] further proposed a semi-supervised framework for conceptualizing CSKBs and demonstrated that abstract knowledge can enhance commonsense inference modeling and question answering. [81] constructed an abstraction benchmark based on eventualities from ASER [83]. Regarding instantiation, [135] introduced a controllable generative framework to identify valid instantiations for abstract knowledge automatically. However, none of the existing studies have fully completed the chain of conceptualization and instantiation, with each focusing on only one aspect. Human annotation is also frequently applied for data collection and verification, which is both expensive and limited in scalability. Additionally, the downstream benefits of instantiated commonsense knowledge have not been thoroughly explored, leaving a significant gap in improving

commonsense reasoning models.

6.2.2 Commonsense Knowledge Distillation

Recent breakthroughs in LLMs [40, 41] have led to numerous efforts in distilling commonsense knowledge into datasets for training performant student models. [42–45] followed the pipeline of symbolic knowledge distillation, which uses human-crafted prompts to extract specific types of knowledge from LLMs for training downstream models. [46] proposed to transfer distilled knowledge from a ranker to a retriever, resulting in a more robust commonsense generator. [47] and [48] focused on distilling conversational responses from LLMs to enhance dialogue agents with commonsense knowledge and high-quality rationales. In this chapter, we share similar aspirations and propose a chain of distillation framework that sequentially obtains abstract and instantiated knowledge from powerful LLMs. Empirical results show that our framework offers more substantial downstream benefits than traditional symbolic knowledge distillation methods.

6.3 Definitions and Datasets

We follow the definitions proposed by [78] and [85] to formulate conceptualization and instantiation. Denote the triples in the original CSKB as $D_o = \{(h_o, r, t) | h_o \in H_o, r \in R, t \in T\}$, where H_o, R, and T are the set of heads, relations, and tails in the original CSKB. The objective of conceptualization is to form a conceptualized head event, denoted as h_a, from the original head h_o. This is achieved by linking a component $i \subseteq h_o$ to a concept c, forming h_a by replacing i with c. Consequently, abstract knowledge is formed by combining the conceptualized head event with the original relation and tail, represented by (h_a, r, t). In the next step, the goal of instantiation is to associate the concept $c \subseteq h_a$ with a new instance i'. This process enables the formation of new commonsense knowledge in the format of $(h_{i'}, r, t)$, where $h_{i'}$ is obtained by replacing $c \subseteq h_a$ with i'. In this chapter, we use ATOMIC [128] as the original CSKB D_o, which contains 310K (h_o, r, t) triples after dropping those with wildcards and 18,839 unique h_o head events. AbstractATOMIC [78] is used as the source of instances i for every head event h_o.

6.4 The CANDLE Framework

This section introduces our CANDLE framework, illustrated in Figure 6.2. Our framework can be outlined in three steps: (1) Instruct ChatGPT to generate contextualized conceptualizations

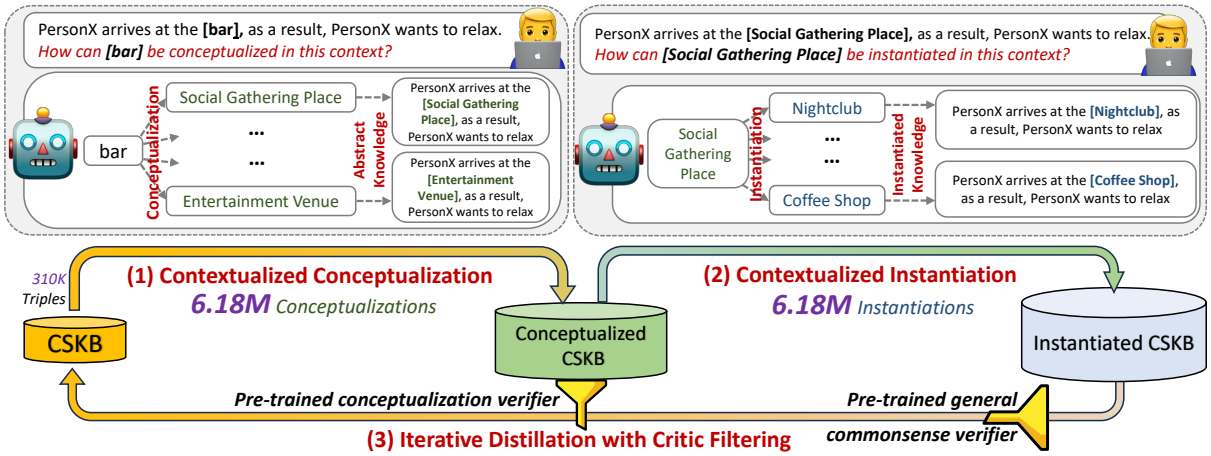


Figure 6.2: Overview of our CANDLE framework. A running example with *PersonX arrives at the bar, as a result, PersonX wants to relax* is shown in the figure, where *bar* is first conceptualized and then instantiated by LLMs. The instantiations can be integrated back into the original CSKB and become input for the framework again.

based on the triples in the original CSKB. (2) Instruct LLAMA2 to instantiate the conceptualizations obtained in Step 1. (3) Apply critic-filtering to the generations in both steps and close the loop by reintroducing the instantiations back to the CSKB.

6.4.1 Contextualized Conceptualization

Previous methods for collecting conceptualizations rely on heuristically matching instances against concepts from WordNet and Probase. However, they suffer from limited concept coverage, resulting in a lack of knowledge diversity after instantiation, and require additional verification to ensure that concept c fits into the original context (h_o, r, t). To address both issues, we propose to utilize ChatGPT as a loose teacher to collect conceptualizations in a one-step inference manner. To verify the feasibility of such choice instead of other open-source LLMs, we carry out a pilot study, in which we randomly select 1000 events and asked ChatGPT and LLAMA2-7B to generate their conceptualizations. The results of our expert evaluation show that ChatGPT has 98% plausible generations, while LLAMA2 only achieves 81%. Therefore, we choose ChatGPT as our core conceptualizer due to its exceptional performances. Following [32] and [42], we use a few-shot prompt to instruct ChatGPT:

<TASK-PROMPT>

<EX$_1$-INP><EX$_1$-OUT>

...

<EX$_{N-1}$-INP><EX$_{N-1}$-OUT>

<EX$_N$-INP>

where **<TASK-PROMPT>** is a task instruction that explains how to conceptualize an event and **<EX$_1$-INP><EX$_1$-OUT>** are human authored examples of conceptualizations for events sampled from ATOMIC. For each example, (h_o, r, t, i) are included in the input, and c is the output. Finally, we provide the N_{th} input as **<EX$_N$-INP>** and ask ChatGPT to generate the corresponding conceptualization as **<EX$_N$-OUT>**. This ensures that ChatGPT not only learns the relationship between instances i and their conceptualizations c but also performs such abstraction in a contextualized manner, ensuring the plausibility of the generated conceptualization c within the original context (h_o, r, t). In this chapter, we set $N = 6$ and obtain $N_c = 20$ conceptualizations for every event h_o.

6.4.2 Contextualized Instantiation

After conceptualizing all events, we proceed to instantiate them by instructing an open-source LLM to reduce the cost as the scale of instantiation is $N_c = 20$ times larger than that of conceptualization. Similarly, we carry out a round of pilot study to demonstrate the feasibility of employing LLAMA2. We ask both ChatGPT and LLAMA2 to instantiate 1,000 ChatGPT-generated conceptualizations, and find that both models are able to produce approximately 95% plausible instantiations with critic filtering. Considering the significant cost of using ChatGPT to generate 6.18 million conceptualizations, we decide to use LLAMA2-13B as our core instantiater. We employ a similar prompt as described in Section 6.4.1, with the modification of replacing **<TASK-PROMPT>** with the explanation of instantiating a conceptualized event and changing **<EX$_1$-INP><EX$_1$-OUT>** to human-authored examples of instantiations for abstract commonsense knowledge triples. (h_a, r, t, c) are included in the input and i' is the expected output. By learning from these examples, LLAMA2 is expected to generate the corresponding instantiation i' (**<EX$_N$-OUT>**) based on the given abstract knowledge triple (h_a, r, t, c) (**<EX$_N$-INP>**). We set $N = 11$ and produce only one instantiation for each conceptualized event h_a due to the significant amount of conceptualizations obtained in the previous step.

6.4.3 Iterating with Critic Filtering

Following [42], we use critic filtering models to eliminate low-quality generations from LLMs. Specifically, we utilize a DeBERTa-v3-large conceptualization discriminator, provided by [85], and VERA-T5-xxl, provided by [102], to evaluate the quality of the generated conceptualizations and instantiations, respectively. We set an empirical threshold value t to serve as the cutoff point for discarding generations with scores below t. In Section 6.5.1, we present evaluations

conducted to determine the optimal value for t. For all downstream applications, we set $t = 0.9$. Post-filtering, the instantiated triples $(h_{i'}, r, t)$ can be reintroduced as the input for conceptualizations again as they continue to represent concrete commonsense knowledge. This iterative process of conceptualization and instantiation forms a loop, which enables continuously augmenting a CSKB. In this chapter, we execute the loop only once, but multiple iterations hold the promise of significantly enhancing the CSKB's knowledge coverage.

6.4.4 Distillation Details

This section provides additional details about the CANDLE distillation process not covered in previous sections. First, we present the prompts used to instruct ChatGPT to perform contextualized conceptualizations and LLAMA2 to perform contextualized instantiation. For prompting ChatGPT to distill conceptualizations, we use a few-shot prompt as shown below:

> **Following the given examples, you are required to conceptualize the instance (enclosed by []) in the last given event into abstract concepts. The concept should still fit into the instance's original sentence. Make sure that the generated abstract concepts are general and not simply hypernyms of the instance.**
>
> ...
>
> **Event <i>: PersonX enjoys drinking in the [bar], as a result, PersonX feels relaxed. [bar] can be conceptualized as Social Gathering Place**
>
> ...
>
> **Event <N>: PersonX likes [painting on the beach], as a result, PersonX will go to the beach. [painting on the beach] can be conceptualized as**

Similarly, for prompting LLAMA2-13B to distill instantiations based on previously generated conceptualizations, we use a few-shot prompt as shown below:

> **Following the given examples, you are required to instantiate the concept (enclosed by []) in the last given event into entities or events. If the event only contains the concept, then instantiate it to an event starting with a subject**

PersonX or PersonY. If the event contains other words, then instantiate it to an entity. The instance should still fit into the original sentence. Make sure that the generated instance is specific.

...

Event <i>: PersonX enjoys drinking in the [Social Gathering Place], as a result, PersonX feels relaxed. [Social Gathering Place] can be instantiated as beer festival

...

Event <N>: PersonX likes [exercise], as a result, PersonX will go to the stadium. [exercise] can be conceptualized as

These prompts are consistent with our descriptions in Section 6.4.1 and Section 6.4.2, where the task description is first presented, followed by human-authored examples, and finally, the event we want to conceptualize or instantiate. We also leverage several tricks in the prompt, such as numbering the examples, generating concepts instead of hypernyms, and keeping the generated responses concise. Finally, we parse the generations via manually defined rules and compile them into a dataset.

Additionally, we introduce some generation settings when prompting LLMs. For ChatGPT, we access it through the official OpenAI APIs[1]. The code of the accessed version is `gpt-3.5-turbo-0613`. We set the temperature to 1.0 and the maximum length for generated tokens to 200. To conceptualize all events in ATOMIC into 20 conceptualizations each, the time required for the distillation process is approximately ten days and the financial budget is around 1500 USD.

For LLAMA2, we access it via the Huggingface Library [211]. The code of the accessed model is `meta-llama/Llama-2-13b-chat-hf`[2]. When prompting, we use the Top-k sampling decoding strategy and set $k = 10$. We set the maximum length of generated tokens to 200. The models are hosted on sixteen NVIDIA-V100 GPUs, and the time required to distill the entire dataset is approximately one month.

After collecting 20 conceptualizations for every head event in ATOMIC and further instantiating them to new entities and events, we construct an expanded knowledge base of ATOMIC. We also include more statistics, as shown in Table 6.1 and Table 6.2. For instantiations, they share the same relational distribution as abstract commonsense triples since we only instanti-

[1]https://chat.openai.com/

[2]https://huggingface.co/meta-llama/Llama-2-13b-chat-hf

	Abs.ATM	CANDLE
#Unq. event	15,388	15,359
#Unq. instance	21,493	21,442
#Unq. conceptualization	31,227	**853,499**
#Tot. conceptualization	503,588	**6,181,391**
#Unq. instantiation	-	**676,737**
#Tot. instantiation	-	**6,181,391**
Avg. #concept/event	32.73	**173.33**
Avg. #Unq. concept/event	28.33	**167.76**
Avg. #concept/instance	23.43	**124.16**
Avg. #Unq. concept/instance	17.27	**100.88**

Table 6.1: Statistics of conceptualizations and instantiations in AbstractATOMIC (Abs.ATM [78]) and CANDLE. Tot. stands for total, Unq. stands for unique, and Avg. stands for average.

ate them once. These statistics indicate that, compared to AbstractATOMIC, which is the only available conceptualization benchmark based on ATOMIC, CANDLE contains more abstract commonsense triples and many more unique conceptualizations. According to our results, it can also be expected that the abstract knowledge distilled from CANDLE is of better quality than AbstractATOMIC, which human annotations or any filtering have not verified.

Relation	ATOMIC	Abs.ATM	CANDLE
xEffect	78,832	938,330	964,765
oEffect	28,351	333,845	346,363
xWant	101,249	1,170,835	1,322,810
oWant	43,079	484,570	551,391
xReact	62,969	510,476	480,259
oReact	26,570	224,706	208,538
xNeed	74,272	900,429	894,338
xAttr	110,791	838,191	810,958
xIntent	45,490	519,813	601,969
Total	572,053	5,921,195	6,181,391

Table 6.2: Statistics of abstract commonsense knowledge triples by relations in ATOMIC, AbstractATOMIC (Abs.ATM [78]), and CANDLE.

For critic filtering, we use the state-of-the-art conceptualization discriminator [85]. This discriminator is utilized to assess the plausibility of CANDLE distilled conceptualizations. It considers the original event, the instance being conceptualized, and the target concept as its inputs and generates a score ranging from 0 to 1 to represent plausibility. For instantiation, we use the pre-trained VERA model [102]. We convert the instantiated commonsense knowledge triple into a declarative statement and request an estimation of its plausibility from VERA. This estimation is provided as a score ranging from 0 to 1. The output scores from both models

Corpus	Conceptualization		Instantiation	
	Size (Unq.)/K	Accept	Size (Unq.)/K	Accept
AbsATM	503.5 (31.22)	-	None	-
EXEM	0.650 (0.650)	-	25.12 (25.12)	-
CANDLE	6,181 (853.5)	82.6%	6,181 (676.7)	77.9%
(critic$_{0.5}$)	4,002 (498.4)	88.1%	4,176 (512.7)	84.4%
(critic$_{0.7}$)	3,272 (382.2)	93.5%	3,098 (455.9)	89.1%
(critic$_{0.9}$)	2,137 (219.4)	97.2%	2,208 (382.1)	94.5%

Table 6.3: Statistics and expert acceptance rates of CANDLE in comparison to AbstractATOMIC (AbsATM [78]) and Exemplar (EXEM [135]). Unq stands for unique.

serve as the critical values assigned to each CANDLE distillation. These critical values are then subjected to further filtering based on various thresholds.

Additionally, we calculate the percentage of unique abstract concepts using BLEU soft uniqueness [42, 212]. We define a concept, denoted as x, as unique if $BLEU_1(C, x) < 0.5$, where C represents all concepts that share the same head event and identified instance with x in AbstractATOMIC. Here, 0.5 serves as an empirical threshold. Our distillation process yields 92.3% unique conceptualizations, indicating a significantly higher diversity than previous datasets.

Similarly, we evaluate the uniqueness of the newly introduced head events resulting from our chain of conceptualization and instantiation. To determine uniqueness, we define an instantiated head event, referred to as $h_{i'}$, as unique if $BLEU_1(h_o, h_{i'}) < 0.5$, where h_o represents the original head event in ATOMIC. The threshold of 0.5 is an empirical threshold. Our empirical results demonstrate that 78.6% of the instantiated events are unique compared to ATOMIC, highlighting the effectiveness of CANDLE in enhancing the semantic coverage of the CSKB.

6.5 Main Evaluations

In this section, we evaluate CANDLE from both intrinsic and extrinsic perspectives. Intrinsically, we demonstrate the high quality and diversity of conceptualizations and instantiations generated by CANDLE (Section 6.5.1). Extrinsically, we explore the benefits by applying the distilled knowledge to downstream tasks (Section 6.5.2).

6.5.1 Distillation Evaluations

Statistics and Quality. We present CANDLE distillation statistics based on ATOMIC in Table 6.3, showing its superiority in scale and concept coverage compared to other benchmarks.

Even with a strict critic filtering threshold ($t = 0.9$), CANDLE maintains its leading position, having the highest count of total and unique knowledge for both types. To assess the quality of the distilled knowledge, we recruit four expert annotators to conduct human evaluations on the plausibility of the generated conceptualizations and instantiations. They are asked to annotate the plausibility of 3,000 randomly sampled abstract commonsense triples (h_a, r, t) and 3,000 instantiated triples $(h_{i'}, r, t)$ from the distilled knowledge set. Accepted triples are those deemed plausible by all annotators. We then analyze accepted triple ratios for different levels of critic filtering, as shown in Table 6.3. Our findings show that LLMs have impressive conceptualization and instantiation abilities, with initial plausibility rates of 82.6% and 77.9% for both types of knowledge, respectively. Critic filtering improves plausibility by up to 14.6% and 16.6%, demonstrating the effectiveness of our measures in maintaining high-quality distilled knowledge.

Conceptualization Diversity. The process of abstracting an event into highly diverse conceptualizations plays a crucial role in CANDLE. It is of significant importance because the greater the diversity of conceptualizations, the broader the knowledge coverage becomes upon instantiation. This, in turn, enhances the overall knowledge coverage within the distillation process. To examine the diversity of the top 10,000 popular distilled conceptualizations, we obtain their hypernyms by matching against Probase [62] and present a visualization in Figure 6.3. It reveals that our distilled conceptualizations possess a high level of diversity across various categories, forming a comprehensive and intricate knowledge base.

6.5.2 Downstream Applications

In this section, we explore the downstream applications of CANDLE. By applying CANDLE to ATOMIC, the distilled conceptualizations and instantiations form a large-scale expansion of the original CSKB, which contains high-quality abstract and concrete commonsense knowledge. Leveraging both types of knowledge as supplementary training data, we enhance various downstream commonsense reasoning models. Specifically, we utilize distilled conceptualizations in the CSKB conceptualization task [85], while instantiations are used in generative commonsense inference (COMET [31]) and zero-shot commonsense QA tasks [35].

CSKB Conceptualization

Task Setup. The CSKB conceptualization task evaluates a model's ability to conceptualize a CSKB through two binary classification subtasks, which are crucial for performing CSKB

Figure 6.3: Hypernyms distribution of the top 10,000 popular conceptualizations distilled from CANDLE.

conceptualization inference upon concept taxonomies [78]. The first subtask, event conceptualization, aims to determine whether h_o can be correctly conceptualized using h_a, where h_a is derived by replacing an instance $i \subset h_o$ with its linked concept c. The second subtask, triple conceptualization, aims to assess the plausibility of a conceptualized triple (h_a, r, t) that represents abstract commonsense knowledge. Accuracy is used as the evaluation metric. Following [85], we use the AbstractATOMIC dataset provided by [78] as the evaluation benchmark.

To obtain our distilled models for these tasks, we first synthesize negative samples from CANDLE distilled conceptualizations. For event conceptualizations, a random concept from another head event without common words is selected as the negative candidate, while for triple conceptualization, a tail of another head event without common words under the same relation is selected. We then fine-tune language models on a balanced mixture of CANDLE distillation and synthesized negative samples to train two models, each serving as a pre-trained general discriminator in their respective task domain. These two models are subsequently fine-tuned on the training sets of AbstractATOMIC to fit into the benchmark, and their performances on the validation and test sets are reported.

Model Type	Backbone Model / Method	Event Conceptualization		Triple Conceptualization	
		Validation	Testing	Validation	Testing
Pre-trained Langauge Models	RoBERTa-large *340M*	77.28	77.99	81.77	82.69
	DeBERTa-v3-large *435M*	78.02	78.27	82.18	82.96
	GPT2-XL *1.5B*	53.71	56.10	47.65	47.21
	PseudoReasoner (RoBERTa-large)	78.33	78.91	79.69	80.27
	PseudoReasoner (DeBERTa-v3-large)	79.03	79.21	79.89	80.07
	CAT (RoBERTa-large) *340M*	78.51	78.53	82.27	83.02
	CAT (DeBERTa-v3-large) *435M*	79.55	79.39	82.88	83.52
Large Language Models	ChatGPT (`openai/gpt-3.5-turbo`)	69.29	68.65	68.54	68.12
	+ Five-shot Exemplars	69.42	70.40	70.27	72.08
	+ Chain-of-thought	74.82	72.32	71.48	72.85
	LLAMA2 *7B*	46.29	43.90	40.81	41.25
	+ Five-shot Exemplars	47.92	44.89	74.67	76.80
	LLAMA2 *13B*	48.17	48.59	48.31	48.55
	+ Five-shot Exemplars	49.29	49.90	80.67	82.08
	Mistral-v0.1 *7B*	46.29	43.90	58.09	58.07
	+ Five-shot Exemplars	51.00	50.06	65.09	69.80
	LLAMA2 (LoRA Fine-tuned) *7B*	75.80	76.27	79.89	82.15
	Mistral-v0.1 (LoRA Fine-tuned) *7B*	75.71	76.76	79.59	80.35
	VERA-T5 *5B*	70.76	70.29	72.60	76.85
	VERA-T5 (Fine-tuned) *5B*	75.69	76.21	80.13	81.25
CANDLE Distilled *(Ours)*	RoBERTa-large *340M*	$80.69_{\uparrow 2.18}$	$80.99_{\uparrow 2.46}$	$83.11_{\uparrow 0.84}$	$84.50_{\uparrow 1.48}$
	DeBERTa-v3-large *435M*	$\mathbf{80.97}_{\uparrow 1.42}$	$\mathbf{81.14}_{\uparrow 1.75}$	$\mathbf{83.64}_{\uparrow 0.76}$	$\mathbf{84.64}_{\uparrow 1.12}$
	LLAMA2 (LoRA Fine-tuned) *7B*	$77.48_{\uparrow 1.68}$	$78.27_{\uparrow 2.00}$	$81.68_{\uparrow 1.79}$	$83.40_{\uparrow 1.25}$
	Mistral-v0.1 (LoRA Fine-tuned) *7B*	$77.77_{\uparrow 2.06}$	$78.29_{\uparrow 1.53}$	$81.95_{\uparrow 2.36}$	$82.54_{\uparrow 2.19}$
	VERA-T5 (Fine-tuned) *5B*	$77.54_{\uparrow 1.85}$	$78.03_{\uparrow 1.82}$	$82.79_{\uparrow 2.66}$	$83.61_{\uparrow 2.36}$

Table 6.4: Performances (Accuracy%) on CSKB conceptualization tasks. The best performances within each model type are underlined, and the best among all models are **bold-faced**. ↑ signifies the improvement compared to the best baseline with the same backbone model or method.

Baselines. We evaluate our distilled models by comparing them against several baselines. These include supervised fine-tuned language models like RoBERTa-Large [75], DeBERTa-V3-Large [77], GPT-2 [91], LLAMA2 [97], Mistral [213], and VERA [102], as well as semi-supervised methods such as PsuedoReasoner [155] and CAT [85]. Due to computational power limitations, we utilize LoRA [214] for fine-tuning LLMs. As additional baselines, we also consider prompting LLMs, including LLAMA2, Mistral, and ChatGPT. We explore both direct zero-shot prompting and alternative methods, such as with five-shot exemplars [215] and chain-of-thought reasoning [196].

Results and Analysis. Table 6.4 shows the results. CAT trained with DeBERTa-v3-large outperforms all other baselines for both tasks. Among LLMs, LLAMA and Mistral perform well after fine-tuning, but they struggle in prompting scenarios. However, pre-training on CANDLE's distilled conceptualizations consistently improves results for both tasks. For example, Mistral shows a significant improvement of 1.54% and 2.19% on two tasks compared to directly fine-

Training Data	Bleu-1	Bleu-2	Bleu-3	Bleu-4	METEOR	ROUGE-L	CIDEr	BERTScore
Backbone: GPT2-XL [91] *1.5B*								
Zero-shot	4.350	1.598	0.732	0.293	5.702	5.030	0.792	37.11
ATOMIC	32.23	19.06	13.27	10.28	17.63	25.50	20.15	58.39
+ Finetune	45.72	29.18	21.12	16.15	29.97	49.69	64.61	76.09
ATOMIC$^{20}_{20}$	42.15	25.77	17.82	13.14	29.82	47.61	63.70	70.39
ATOMIC-10X	33.69	18.82	11.71	7.910	18.78	25.69	19.29	61.47
+ Finetune	45.38	29.20	21.09	16.15	30.09	49.86	65.02	75.89
AbstractATOMIC	29.46	17.16	11.89	9.019	17.42	24.30	19.95	57.83
+ Finetune	45.30	29.08	21.00	16.06	29.98	48.61	63.98	75.56
CANDLE Distilled	26.91	16.44	12.31	10.28	17.66	23.66	21.36	57.15
+ Finetune	**50.71**	**33.85**	**25.55**	**20.43**	**32.45**	**51.91**	**69.68**	**76.86**
Backbone: ChatGPT [40] (`openai/gpt-3.5-turbo`)								
Zero-shot	11.82	4.258	1.891	0.926	13.87	13.73	4.350	49.28
Five-shot	26.32	12.50	7.160	4.415	18.60	24.65	8.313	58.69
Chain-of-thought	9.906	3.568	1.556	0.736	11.85	11.02	2.905	46.17
Backbone: LLAMA2 [97] *7B*								
Zero-shot	18.26	7.453	3.594	1.945	15.90	20.28	8.872	48.23
Five-shot	31.22	16.87	9.767	5.989	19.74	27.67	17.83	58.41
ATOMIC	29.94	16.44	10.03	6.631	19.02	25.75	18.71	59.68
+ Finetune	42.04	23.01	14.10	9.125	27.80	42.90	53.17	71.52
ATOMIC$^{20}_{20}$	41.07	22.46	13.62	8.619	27.74	42.42	53.28	71.77
ATOMIC-10X	33.06	17.65	9.986	6.078	19.22	25.32	17.80	61.25
+ Finetune	42.07	23.08	14.14	9.198	28.14	42.75	53.69	71.93
AbstractATOMIC	26.08	13.27	7.799	5.018	15.08	21.20	14.78	56.83
+ Finetune	42.78	23.64	14.58	9.471	27.74	42.55	53.12	71.51
CANDLE Distilled	28.93	15.56	9.468	6.140	18.60	25.37	17.20	60.27
+ Finetune	43.86	24.40	15.12	10.00	28.36	43.86	54.25	72.94

Table 6.5: Performances (%) of the commonsense inference modeling task (COMET) on the full test set of ATOMIC$^{20}_{20}$. The best ones within each backbone are underlined, and the best among all is **bold-faced**.

tuning on AbstractATOMIC. Additionally, the distilled DeBERTa-v3-large surpasses all baseline models and achieves state-of-the-art performance. This can be attributed to the distilled conceptualizations obtained from CANDLE, which grant the model a more comprehensive understanding of conceptualizations and subsequently enhance its discriminatory capabilities.

Generative Commonsense Inference

Task Setup. The task of generative commonsense inference modeling (COMET [31]) asks the model to generate commonsense tails t based on given head h_o and relation r inputs. Following [129], we use the full test set of ATOMIC$^{20}_{20}$ as our evaluation benchmark. We use several automatic metrics for evaluation, including BLEU [162], ROUGE-L [164], METEOR [163], CIDEr [165], and BERTScore [168]. Meanwhile, four expert annotators are recruited to conduct expert evaluations of the generations. They are asked to annotate the plausibility of 200

randomly selected commonsense triple generations under each setting, and the resulting plausibility rates are reported.

Similar to training distilled models in previous tasks, we first pre-train GPT2 and LLAMA2-7B on critic-filtered CANDLE instantiations, where each $(h_{i'}, r, t)$ triple is concatenated into a sentence via natural language templates. Subsequently, we fine-tune these models on the training split of ATOMIC$^{20}_{20}$ to fit them into the benchmark. Finally, we report their performances on the test set.

Baselines. For baselines, we separately train GPT2 and LLAMA2-7B on the training sets of ATOMIC, ATOMIC$^{20}_{20}$, ATOMIC10X [42], and AbstractATOMIC. These models are then fine-tuned on the training split of ATOMIC$^{20}_{20}$ and evaluated on its test set. We also include their zero-shot prompting performances, with LLAMA2 being evaluated with five-shot exemplars. ChatGPT's performances under zero-shot, five-shot, and chain-of-thought settings are also reported.

Results and Analysis. Table 6.5 shows the results. Among the baselines, models pre-trained on ATOMIC-10X achieve the highest expert acceptance rate, surpassing those trained on AbstractATOMIC. This may be because ATOMIC-10X covers a wider range of commonsense relations consistent with ATOMIC$^{20}_{20}$. However, CANDLE distilled models achieve the highest scores compared to baselines with the same backbone model. For example, the CANDLE distilled LLAMA-7B model improves BERTScore by 1.01% and expert-plausibility by 3.00% compared to the best baseline. It also outperforms ChatGPT in all automatic metrics while maintaining a high plausibility rate of around 80%. This emphasizes the advantages of using CANDLE distilled instantiations for COMET training over traditional symbolic knowledge distillation methods or conceptualization augmentation. Interestingly, we also observe that LLAMA2 has a tendency to generate long and contextually rich commonsense knowledge. On the other hand, GPT2, when fine-tuned on ATOMIC-like data, may generate shorter and more concise knowledge, which aligns with the format and length of knowledge in ATOMIC2020, thus achieving better results in automatic evaluations. However, human annotators tend to consider long and contextually rich commonsense statements, generated by LLAMA2, as more plausible.

Zero-shot Commonsense QA

Task Setup. The task of zero-shot commonsense QA involves selecting the most plausible option for commonsense questions without any supervision signals from benchmark data. We

follow the most effective pipeline by [35], which fine-tune language models on QA pairs synthesized from knowledge in CSKBs. The head h_o and relation r of a (h_o, r, t) triple are transformed into a question using natural language prompts, with the tail t serving as the correct answer option. Distractors or negative examples are generated by randomly sampling tails from triples that do not share common keywords with the head. In addition to directly synthesizing from knowledge triples in ATOMIC, we augment ATOMIC by sampling triples from ATOMIC-10X, AbstractATOMIC, and CANDLE instantiations. The number of sampled triples is the same as in the original ATOMIC dataset. We then synthesize them into QA pairs to train different baseline models and CANDLE distilled models. For our distilled models, we utilize QA pairs sourced from CANDLE-instantiation augmented ATOMIC to train a DeBERTa-v3-large model using the marginal ranking loss and a T5-xxl model [93] following the training regime of VERA. We evaluate the performance of all models on the validation split of Abductive NLI (aNLI [190]), CommonsenseQA (CSQA [21]), PhysicalIQA (PIQA [114]), SocialIQA (SIQA [80]), and WinoGrande (WG [191]). Accuracy is used as the evaluation metric.

Baselines. First, we report performances of vanilla RoBERTa-Large, DeBERTa-v3-Large, Self-talk [27], COMET-DynaGen [29], SMLM [34], MICO [36], MR [35], STL-Adapter [37], and the previous state-of-the-art method, CAR [111]. For MR and CAR, DeBERTa-v3-Large is used as the backbone, and their performances on ATOMIC-10X and AbstractATOMIC are also reported. For LLMs, we report the performances of prompting GPT3.5 [32], ChatGPT, GPT4 [41], LLAMA2, and Mistral in a zero-shot manner. For ChatGPT, its performances with chain-of-thought [196] and self-consistency chain-of-thought [216] prompting are also reported. We also train several VERA-T5-xxl baselines on different sets of QA pairs as LLM baselines.

Results and Analysis. Table 6.8 shows the results, demonstrating that CANDLE distilled models generalize better than the baselines across several commonsense QA benchmarks. For instance, VERA demonstrates an average improvement of 1.4% compared to the best baseline. This can be attributed to the inclusion of new entities and events in CANDLE instantiations that are absent in other CSKBs, where CANDLE instantiations can aid in answering commonsense questions that require knowledge of these new instances. Furthermore, the distilled DeBERTa-v3-large model outperforms all baselines, including methods utilizing LLMs. This also indicates that augmenting with CANDLE distilled instantiations provides a more significant advantage compared to using symbolically distilled or abstract knowledge as training data.

6.6 Analysis

6.6.1 Feasibility of Iterating CANDLE

We first demonstrate the feasibility of iterating the CANDLE framework with more than one round. To do so, we randomly sample 10,000 distilled instantiations from LLAMA2 as the input for the CANDLE framework and execute the framework again, resulting in 200,000 conceptualizations and 200,000 instantiations. Subsequently, we randomly select 300 from each set and annotate them accordingly. The results are shown in Table 6.6. We observe that iterating the framework produces slightly better results than the first loop. This improvement may be attributed to the fact that the knowledge generated in the initial loop is more easily understood by LLMs compared to the human-annotated data in ATOMIC. Moreover, 58% of the conceptualizations and 44% of the instantiations are novel compared to the first loop. Based on these findings, we believe that our iterative framework is effective, and the iteration process enhances the augmentation of a CSKB through multiple iterations.

Critic	Conceptualization	Instantiation
0.0	92.3%	85.5%
0.5	94.6%	91.2%
0.7	95.9%	93.3%
0.9	98.3%	96.7%

Table 6.6: Annotation results of distillations obtained from the second round of executing CANDLE.

6.6.2 Source of Empirical Gains

Since LLAMA2 has been pre-trained on some evaluation benchmarks, it remains questionable whether the empirical gains in downstream tasks are due to knowledge overlap between distillations from LLAMA2 and the evaluation benchmarks. To this extent, we further demonstrate that CANDLE distilled models perform better due to improved generalizability rather than relying on data overlap with the evaluation data. We use SentenceBERT [203] to measure the textual similarity between the distilled knowledge and the evaluation data for each task. We then calculate the ratio of data that exhibits semantic overlap with a similarity score exceeding 0.5 and also report the average similarity. The results are shown in Table 6.7. Based on the results, we observe that the distilled knowledge has minimal overlap with the evaluation set. This indicates that the empirical gain primarily stems from our distilled knowledge, which improves the

generalizability of the models, rather than relying on knowledge overlap with the evaluation sets.

Task	CSKB Concept.	COMET	CSQA
Overlap Ratio	10.1%	8.7%	5.3%
Avg. Similarity	0.39	0.38	0.31

Table 6.7: Knowledge overlap ratio and average similarity between distilled knowledge and evaluation data.

6.6.3 Ablation Study

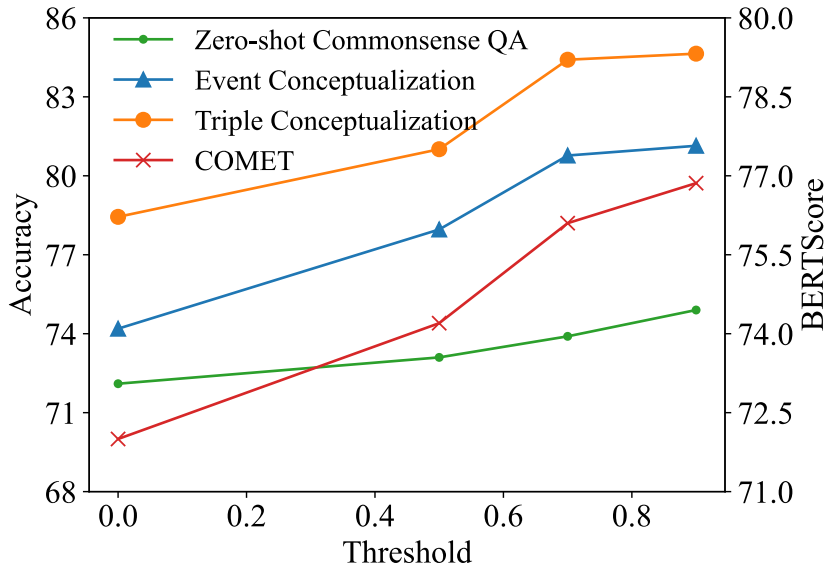


Figure 6.4: Ablation results examining the impact of different threshold values in CANDLE's critic filtering.

In this section, we examine the impact of our critic filters on the ablation of CANDLE. Specifically, we investigate the effect of different levels of critic threshold or completely abandoning critic filtering on downstream tasks. We conduct four experiments with different settings, denoted as $t \in \{0, 0.5, 0.7, 0.9\}$, where $t = 0$ corresponds to abandoning critic filtering and using all distilled knowledge as complementary training data. For detailed statistics, please refer to Table 6.3. For each value of t, we select the distilled knowledge with a critic score higher than t and utilize it as complementary training data to train student models for the three downstream tasks. We employ the same training strategies described in prior sections. In the case of CSKB conceptualization and zero-shot commonsense QA tasks, we utilize DeBERTa-v3-large as the backbone model, with accuracy as the evaluation metric. For COMET, we use GPT2 and evaluate using the BERTScore as the evaluation metric. The results are visualized in Figure 6.4. Our analysis reveals a consistent trend where higher threshold values yield improved performance,

indicating the reliability of our critic filter. However, it is worth noting that setting the threshold above 0.9 may potentially lead to even better performance. Nevertheless, such a trade-off comes with a downside: it reduces the amount of usable knowledge in each distillation round, which can impede the iterative process. The reason for this is that when the number of distilled conceptualizations and instantiations decreases significantly in each round, CANDLE is unable to incorporate new instantiated data for future distillation iterations. As a result, the "convergence" of those high-critic data occurs prematurely in CANDLE.

6.6.4 Case Study

We present some examples in Table 6.9 to show conceptualizations and instantiations generated by CANDLE, along with their corresponding critic values assigned by our critic-filtering discriminators. It can be observed that both ChatGPT and LLAMA2 exhibit the ability to generate high-quality knowledge based on given instructions. Furthermore, they can introduce novel conceptualizations and events during the distillation chain, effectively meeting our expectations of CANDLE. Future works can investigate the feasibility of incorporating conceptualization and abstract knowledge into more downstream tasks, such as metaphysical reasoning [217], complex reasoning [101, 218, 219], theory-of-mind reasoning [220], text to table generation [221], and commonsense knowledge graph denoising [204].

6.7 Conclusions

This chapter introduces CANDLE, a distillation framework that realizes the chain of conceptualization and instantiation over CSKBs. We demonstrate the efficacy of CANDLE through comprehensive evaluations of the distilled knowledge and its positive impact on downstream tasks. Our research sheds light on distilling LLMs to enable more robust and generalizable commonsense reasoning.

Model/Method	CSKB	a-NLI	CSQA	PIQA	SIQA	WG	Avg.
Pre-trained Language Models							
Random Vote	-	50.0	20.0	50.0	33.3	50.0	40.7
Majority Vote	-	50.8	20.9	50.5	33.6	50.4	41.2
GPT2-L [91]	-	56.5	41.4	68.9	44.6	53.2	52.9
RoBERTa-L [75]	-	65.5	45.0	67.6	47.3	57.5	56.6
DeBERTa-v3-L [77]	-	59.9	25.4	44.8	47.8	50.3	45.6
Self-talk [27]	-	-	32.4	70.2	46.2	54.7	-
SMLM [34]	*	65.3	38.8	-	48.5	-	-
COMET-DynGen [29]	ATOMIC	-	-	-	50.1	-	-
MICO [36]	ATOMIC	-	44.2	-	56.0	-	-
STL-PLM [37]	ATOMIC	71.6	64.0	72.2	63.2	60.5	66.3
MTL [37]	CWWV	69.6	67.3	72.5	52.0	57.2	63.7
MTL [37]	CSKG	69.8	67.1	72.0	61.9	59.3	66.0
STL-Adapter [37]	ATOMIC	71.3	66.5	71.1	64.4	60.3	66.7
STL-Adapter [37]	CSKG	71.5	66.7	72.1	64.7	59.0	66.8
RoBERTa-L (MR) [35]	ATM_{10X}	70.8	64.2	71.7	61.0	60.7	65.7
RoBERTa-L (MR) [35]	ATOMIC	70.8	64.2	72.1	63.1	59.2	65.9
RoBERTa-L (MR) [35]	CWWV	70.0	67.9	72.0	54.8	59.4	64.8
RoBERTa-L (MR) [35]	CSKG	70.5	67.4	72.4	63.2	60.9	66.8
DeBERTa-v3-L (MR) [35]	ATM10X	75.1	71.6	79.0	59.7	71.7	71.4
DeBERTa-v3-L (MR) [35]	ATOMIC	76.0	67.0	78.0	62.1	76.0	71.8
ZS-Fusion [37]	CWWV	69.6	67.6	73.1	53.7	59.5	64.7
ZS-Fusion [37]	CSKG	72.4	68.3	73.0	66.7	60.9	68.3
MKIF [39]	CSKG	72.5	71.0	73.1	-	61.0	-
CAR-RoBERTa-L [111]	ATOMIC	72.3	64.8	73.2	64.8	61.3	67.3
CAR-RoBERTa-L [111]	AbsATM	72.7	66.3	73.2	64.0	62.0	67.6
CAR-DeBERTa-v3-L [111]	ATOMIC	78.9	67.2	78.6	63.8	78.1	73.3
CAR-DeBERTa-v3-L [111]	AbsATM	<u>79.6</u>	69.3	78.6	64.0	<u>78.2</u>	<u>73.9</u>
DeBERTa-v3-L (CANDLE Distilled)	CANDLE	**81.2**$_{\uparrow 1.6}$	69.9$_{\uparrow 0.6}$	80.3$_{\uparrow 1.7}$	65.9$_{\uparrow 1.9}$	**78.3**$_{\uparrow 0.1}$	**74.9**$_{\uparrow 1.0}$
Large Language Models							
GPT-3.5 (`text-davinci-003`)	-	61.8	68.9	67.8	68.0	60.7	65.4
ChatGPT (`gpt-3.5-turbo`)	-	69.3	74.5	75.1	69.5	62.8	70.2
+ Chain-of-thought	-	70.5	<u>75.5</u>	79.2	**70.7**	63.6	71.9
+ Self-consistent chain-of-thought	-	73.2	**75.7**	<u>81.7</u>	<u>69.7</u>	64.1	72.9
GPT-4 (`gpt-4`)	-	75.0	43.0	73.0	57.0	77.0	65.0
LLAMA2 (7B [97])	-	57.5	57.8	78.8	48.3	69.2	62.3
LLAMA2 (13B [97])	-	55.9	67.3	80.2	50.3	72.8	65.3
Mistral-v0.1 (7B [213])	-	51.0	59.6	**83.0**	42.9	75.3	62.4
VERA-T5-xxl [102]	ATOMIC	71.2	61.7	76.4	57.7	67.5	66.9
VERA-T5-xxl [102]	ATM10X	70.3	59.5	75.1	58.2	67.2	66.1
VERA-T5-xxl [102]	AbsATM	73.2	63.0	77.2	58.1	68.1	68.0
VERA-T5-xxl (CANDLE Distilled)	CANDLE	73.8$_{\uparrow 0.6}$	64.7$_{\uparrow 1.7}$	77.6$_{\uparrow 0.4}$	59.4$_{\uparrow 1.2}$	71.3$_{\uparrow 3.2}$	69.4$_{\uparrow 1.4}$
Supervised Learning & Human Performance							
RoBERTa-L (Supervised)	-	85.6	78.5	79.2	76.6	79.3	79.8
DeBERTa-v3-L (Supervised)	-	89.0	82.1	84.5	80.1	84.1	84.0
VERA-T5 (Multitask Supervised)	-	83.9	77.8	88.5	80.1	92.4	84.5
Human Performance	-	91.4	88.9	94.9	86.9	94.1	91.2

Table 6.8: Full zero-shot evaluation results (Accuracy%) on five commonsense question answering benchmarks. The best results are **bold-faced**, and the second-best ones are <u>underlined</u>. ↑ signifies the improvement CANDLE-distilled models achieve compared to the best baseline with the same backbone model. ATM10X stands for ATOMIC-10X [42] and AbsATM stands for AbstractATOMIC [78]. All scores are retrieved from their original papers.

Original	Concept./Instant.	Critic
PersonX swims in the lake, *as a result, PersonX feels*, tired.	PersonX swims in **freshwater**, *as a result, PersonX feels*, tired.	0.97
	PersonX swims in **the sea**, *as a result, PersonX feels*, tired.	0.87
	PersonX **swims**, *as a result, PersonX feels*, tired.	0.89
	PersonX **swims every week**, *as a result, PersonX feels*, tired.	0.81
PersonX is sitting in class, *as a result, PersonX will*, learns something.	PersonX is sitting in **instructional period**, *as a result, PersonX will*, learns something.	0.54
	PersonX is sitting in **a math class**, *as a result, PersonX will*, learns something.	0.75
	PersonX **study**, *as a result, PersonX will*, learns something.	0.78
	PersonX **learns how to do the exam**, *as a result, PersonX will*, learns something.	0.81
PersonX buys PersonY a gift, *as a result, PersonY feels*, joyful.	**remembrance**, *as a result, PersonY feels*, joyful.	0.19
	PersonX reminisce, *as a result, PersonY feels*, joyful.	0.27
	PersonX **shopping**, *as a result, PersonY feels*, joyful.	0.61
	PersonX **buys a new toy for PersonY**, *as a result, PersonY feels*, joyful.	0.90
PersonX always fought, *as a result, PersonY feels*, angry.	PersonX always **violent behavior**, *as a result, PersonY feels*, angry.	0.98
	PersonX always **punch others hardly**, *as a result, PersonY feels*, angry.	0.91
	combative personality, *as a result, PersonY feels*, angry.	0.98
	PersonX **PersonX likes to join a fight**, *as a result, PersonY feels*, angry.	0.85
PersonX gets a new bike, *as a result, PersonX wants*, to ride it.	PersonX gets a **transportation tool**, *as a result, PersonX wants*, to ride it.	0.92
	PersonX gets a **motor**, *as a result, PersonX wants*, to ride it.	0.98
	bike possession, *as a result, PersonX wants*, to ride it.	0.93
	PersonX has a nice bicycle, *as a result, PersonX wants*, to ride it.	0.89
PersonX spends time with PersonY, *PersonX is seen as*, social.	PersonX spends **love-building period** with PersonY, *PersonX is seen as*, social.	0.05
	PersonX spends **time in love** with PersonY, *PersonX is seen as*, social.	0.37
	social activity, *PersonX is seen as*, social.	0.64
	PersonX enjoys going to parties, *PersonX is seen as*, social.	0.73
PersonX hears sirens, *as a result, PersonX will*, make way to the siren.	**emergency response**, *as a result, PersonX will*, make way to the siren.	0.37
	PersonX sees an ambulance coming, *as a result, PersonX will*, make way to the siren.	0.74
	PersonX hears **loud noise**, *as a result, PersonX will*, make way to the siren.	0.67
	PersonX hears **a fire truck beeping**, *as a result, PersonX will*, make way to the siren.	0.77

Table 6.9: Case studies of **conceptualizations** and **instantiations** distilled from CANDLE in their original context. Original stands for the original triple sampled from ATOMIC. In the Concept./Instant. column, each box contains an abstract commonsense triple that includes **conceptualization**, followed by an instantiated commonsense triple with **instantiation**. We demonstrate two ways to conceptualize each original triple from ATOMIC.

CHAPTER 7

CONCEPTUALIZATION-GUIDED KNOWLEDGE EDITING

The previous chapters establish conceptualization as a core mechanism for *generalizable* commonsense reasoning: by abstracting surface events into reusable concepts and re-grounding them back into contextualized instances, models can expand coverage, reduce spurious supervision artifacts, and improve robustness on downstream tasks. At the same time, they also raise a natural next question: once commonsense knowledge is encoded inside a model, can we *maintain* and *repair* it in a way that preserves this concept-level generalization, rather than treating each error as an isolated exception?

This question becomes more pressing as the field transitions from PTLM-era reasoning systems to modern LLMs. Much of the prior work on commonsense acquisition and augmentation was developed and validated primarily on relatively small or medium-sized pre-trained models, where updating behavior via continued training is still feasible. With increasingly large backbones, however, full retraining or repeated fine-tuning becomes prohibitively expensive, while failures in commonsense plausibility more often reflect broader, concept-level gaps that surface across many paraphrases, contexts, and relational realizations. Knowledge Editing (KE) is therefore an attractive tool for efficiently correcting and updating large models, yet existing KE methods are typically formulated around single, surface-form facts—an assumption that is especially fragile for commonsense knowledge, whose expressions are diverse and whose effects can cascade through tightly coupled generations.

Motivated by this gap, we introduce ConKE, which connects the conceptualization-centric perspective of earlier chapters to the post-training setting of knowledge editing. The key idea is to use conceptualization and instantiation to lift an error from a single statement into a concept-structured neighborhood and then ground it back into diverse, context-consistent instances. This semantic enrichment turns editing from a local patch into a more transferable update, improving coverage and stability while reducing unintended side effects—thereby enabling scalable, concept-guided editing of commonsense knowledge in LLMs.

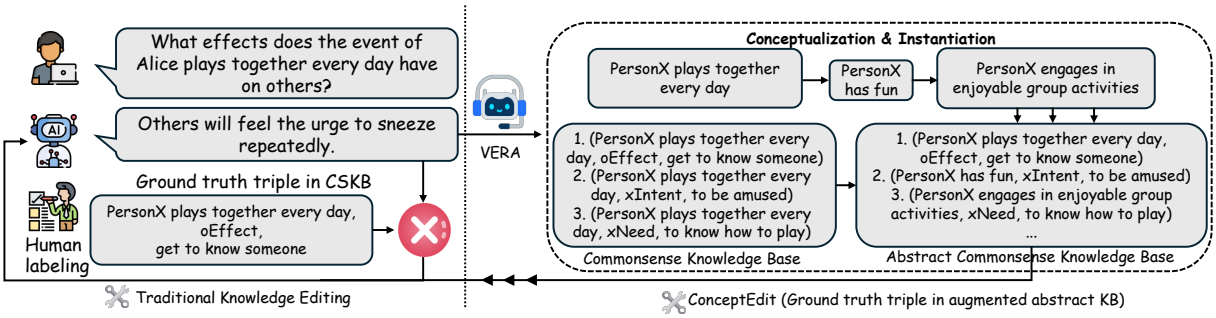


Figure 7.1: An overview of CONKE, which pipelines conceptualization and instantiation, knowledge editing, and LLM verification together for automated and scalable knowledge editing over commonsense knowledge.

7.1 Introduction

Recent advancements in Large Language Models (LLMs [193, 222–224]) have led to Knowledge Editing (KE [225–227]), a computationally efficient strategy to correct inaccurate responses and update LLMs by modifying their internal weights or representations, without re-training the entire model. Such methods have been applied to various domains, including factual reasoning [228, 229], medical knowledge [230], and commonsense reasoning [231], and have proven effective in enhancing domain-specific expertise.

Despite their success, current KE methods face several challenges, including limited knowledge coverage [127] in existing commonsense knowledge bases [11, 24, 45, 210, 232–234] which offer limited coverage and focus on isolated facts, rather than forming hierarchical structures that enable generalization through editing [235, 236]. Furthermore, the unstructured nature of commonsense knowledge complicates scaling, while the flexible representation of commonsense knowledge means that a single fact may manifest in multiple formats. This necessitates editing at the `(relation, tail)` pair level rather than at individual tokens.

To address these issues, we present CONKE, a novel knowledge editing framework tailored for editing commonsense knowledge within LLMs. We use VERA [102], an automated commonsense plausibility verifier, to assess the plausibility of commonsense knowledge in LLMs. For knowledge deemed erroneous and requiring edits, we integrate conceptualization and instantiation [85, 111] to enrich semantic coverage and support more generalizable editing, covering not only the targeted knowledge but also other potentially relevant yet implausible information within the LLM. implausible information within the LLM. This pipeline therefore integrates automated detection, semantic enrichment, and edit application into one closed loop, enabling fully end-to-end scalability.

To ensure flexibility, CONKE adopts an open-ended format for editing, enabling the handling

of arbitrary knowledge structures rather than focusing solely on traditional (h, r, t) triplets. We go beyond traditional Knowledge Editing techniques by combining automated knowledge detection, conceptualization, and instantiation, enhancing the model's ability to generalize and adapt to diverse contexts. Experimental results on AbstractATOMIC [78] demonstrate that LLMs enhanced by CONKE generate commonsense knowledge with improved plausibility. Further evaluations across five commonsense question-answering benchmarks also show performance improvements. These experiments demonstrate the robustness and generalizability of our approach in enhancing commonsense reasoning across diverse architectures and tasks.

7.2 Related Works

7.2.1 Knowledge Editing

Knowledge editing [237] aims to update an LLM's internal knowledge without full retraining or relying solely on prompt engineering, is becoming increasingly crucial. [238] propose ROME, which identifies and updates factual associations within specific MLP layers, achieving precise single-fact edits guided by causal mediation analysis. MEMIT [239] extends ROME's principles to handle large-scale edits simultaneously. By distributing updates across multiple layers and parameters, MEMIT efficiently integrates thousands of facts while maintaining specificity and fluency. GRACE [240], on the other hand, avoids internal parameter changes by integrating external dictionaries and adapters as a modular memory source. This approach allows flexible, inference-time access to new knowledge, though it may sacrifice some internal coherence and interpretability. In this chapter, we build upon these methods to enhance editing commonsense knowledge in LLMs.

7.2.2 Conceptualization in Commonsense

Conceptualization abstracts entities or events into general concepts, forming abstract commonsense knowledge [103], while instantiation grounds these concepts into new instances, introducing additional commonsense knowledge. Previous work largely focused on entity-level conceptualization [68, 69, 88, 136, 142], with [78, 85, 111] pioneering event-level conceptualization from WordNet [57] and Probase [62]. For instantiation, [135] introduced a controllable generative framework that automatically identifies valid instances. In this chapter, we leverage the conceptualization distillation framework proposed by [98] to augment the knowledge being edited, ensuring broader semantic coverage and thereby improving the generalizability of edited

knowledge.

7.3 The ConKE Framework

An overview of ConKE is presented in Figure 7.1. Our framework consists of three main components: (1) automated knowledge verification with VERA [102], (2) abstract knowledge acquisition via conceptualization and instantiation, and (3) LLM knowledge editing. We use the AbstractATOMIC [78] and CANDLE [98] datasets for training and evaluation as two rich sources of abstract knowledge with conceptualization and instantiation. The training set of both datasets are used for editing and the testing sets are used for evaluation.

7.3.1 Automated Knowledge Verification

Since commonsense knowledge is vast, traditional human-in-the-loop methods for detecting and correcting erroneous outputs in LLMs are neither easily scalable nor adaptable. Inspired by recent advances in using LLMs as automated judges [241, 242], we propose a fully automated verification strategy to assess an LLM's internal commonsense knowledge. Our verification process involves VERA [102], a discriminative model trained to score the plausibility of arbitrary commonsense statements, as our evaluation tool. For each triple in the AbstractATOMIC [78] training set, we prompt the LLM with the head event and request it to generate the corresponding relation and tail. VERA then evaluates the plausibility of the generated knowledge by producing a score in the range $[0, 1]$, where values above 0.5 are considered plausible, and those below 0.5 are deemed implausible. By iterating over all triples, this process provides both the LLM's generated responses and VERA's discrimination results, pinpointing which portions of the generated knowledge are incorrect. Consequently, we can identify the exact "areas" within the LLM's internal knowledge that require editing. This automated pipeline eliminates the dependence on costly human annotations for error detection, enabling scalable and efficient improvements of the LLM's commonsense understanding.

7.3.2 Conceptualization and Instantiation

While existing approaches primarily integrate decontextualized commonsense knowledge into LLMs through KE techniques, we hypothesize that capturing the diverse patterns that the same piece of knowledge can exhibit under different contexts is equally important. However, repeated editing may result in knowledge drift, where successive modifications will lead to subtle con-

flicts, causing the model's internal representation to become unstable. To this end, we augment the knowledge to be edited by implementing both conceptualization and instantiation, following [98]. For each triple targeted for editing, we first abstract its instances into more general concepts by prompting GPT-4o, producing abstract knowledge triples (Figure 7.1). We then instantiate these abstract concepts into novel, context-specific instances, again using GPT-4o, thereby forming a rich knowledge base. This process yields approximately 160,000 commonsense knowledge triples, substantially improving the semantic coverage and contextual adaptability of the edited knowledge. Also, this process ensures that the knowledge is rooted in real-world scenarios, enhancing the model's ability to reason about underlying causes and effects, which is a cornerstone of effective commonsense reasoning. When the model's outputs are deemed implausible, we retrieve multiple plausible *(head, relation, tail)* triples from AbstractATOMIC to guide the model's internal revision. Rather than inserting a single missing piece of knowledge, we introduce a set of facts that collectively represent a broader spectrum of commonsense patterns. This approach encourages the model to infer and internalize more generalizable patterns of reasoning, ultimately improving its capacity to handle previously unseen events and scenarios. By automating both the evaluation and the subsequent knowledge integration, our framework scales to a level of complexity that would be prohibitively costly with manual annotation, while still ensuring consistent improvements in the model's commonsense reasoning abilities. Additionally, we are mindful of cascading effects that may arise when modifying a piece of commonsense knowledge. As noted in [226], knowledge is highly interconnected, and modifying one fact can trigger unintended changes in related facts, leading to inconsistencies. To mitigate these cascading effects, we use conceptualization and instantiation to ensure that modifications to abstract concepts are consistently applied to their related instances, hence maintaining coherence and reducing the risk of introducing inconsistencies.

Discussion on global consistency. A further question is whether editing at the concept level can maintain *global* consistency across the model's broader commonsense knowledge. In the current formulation, CONKE improves consistency primarily in a local and semantically structured sense rather than guaranteeing full global coherence over all affected knowledge. More specifically, by enriching an edit target with conceptualized variants and instantiated realizations, CONKE encourages the model to revise a *family* of related expressions together, instead of patching only a single surface form. This substantially reduces brittleness and helps the edit generalize to nearby paraphrases, substitutions, and contexts. However, because the knowledge is stored in distributed model parameters rather than in an explicitly constrained symbolic mem-

ory, interactions with distant or weakly related knowledge cannot be exhaustively controlled.

The practical role of conceptualization in CONKE is therefore to improve *semantic consistency under local transfer*. The edited knowledge is surrounded by abstractions that capture shared structure and by instantiations that anchor that structure in diverse concrete situations. Together with automated plausibility verification, this reduces the likelihood that an edit remains too narrow or contradicts obvious neighboring realizations. Nevertheless, full global consistency would require stronger mechanisms than those implemented here, such as explicit graph-level constraints over edited knowledge, iterative post-edit auditing over broader neighborhoods, or hybrid symbolic-neural representations that expose long-range dependencies among commonsense assertions.

For this reason, the present thesis views global consistency not as a property that is fully solved by CONKE, but as an important next-stage research problem. The contribution of CONKE is to show that conceptualization provides a more principled editing unit than isolated factual strings: it moves the editing process from single-surface correction toward semantically organized revision. This makes consistency easier to *improve*, even if it does not make it possible to guarantee consistency in the strict formal sense. Future work may therefore combine concept-level editing with explicit consistency-checking modules, retrieval-based monitoring of affected neighborhoods, or multi-step editing schedules that jointly optimize local correctness and global compatibility.

7.3.3 LLM Knowledge Editing

Finally, we apply knowledge editing to the LLM using the enriched knowledge base generated through our conceptualization and instantiation processes, correcting errors identified by VERA. To accomplish this, we experiment with three established knowledge editing methods: MEMIT [239], ROME [238], and GRACE [240]. For GRACE, which relies on adapters to determine whether and how to use an external dictionary, we adopt the original deferral mechanism implementation. We evaluate our framework with these editing methods on four representative LLM backbones: `Mistral-7B-Instruct-v0.2`[213], `Meta-Llama-3-8B-Instruct`[224], `Chatglm2-6b`[243], and `GPT-J-6B`[244].

7.4 Experiments and Analyses

In this section, we evaluate LLMs after applying ConKE through expert and automated assessments, illustrating improved performance on downstream tasks and present several ablation

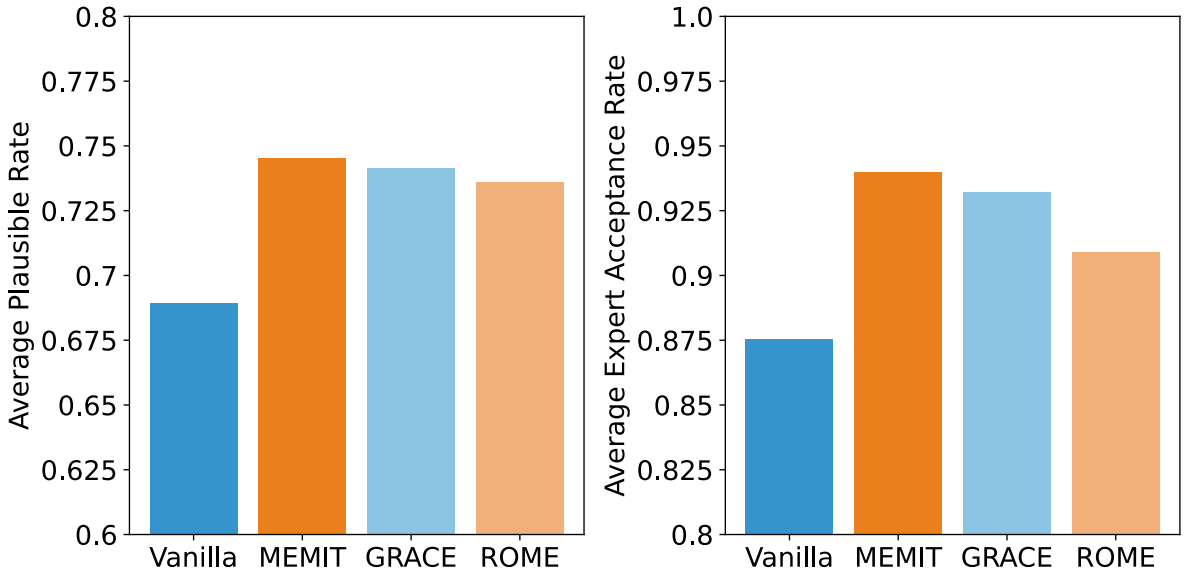


Figure 7.2: Average plausible rate and expert acceptance rate of LLMs' generation after CONKE.

studies.

7.4.1 LLMs-After-Editing Evaluation

We first evaluate LLMs after editing via two measures. First, we prompt these LLMs with head events in the testing set of AbstractATOMIC and ask it to complete the commonsense knowledge. With the generations on the testing set, we ask VERA to score them again and we calculate the plausible ratio whose scores are above 0.5. Then, we sample a subset of 200 generations and recruit two expert annotators to conduct a manual analyses on the acceptance ratio of the plausible assertions that passed VERA's filtering. We compare models after being edited with MEMIT, GRACE, and ROME, and set another vanilla group as baseline comparison. As shown in Figure 7.2, both VERA and human evaluations exhibit consistent trends, with human raters tend to assign higher scores but identifying similar improvements. When applying MEMIT-based editing, both VERA and human evaluations show notable enhancements over the Vanilla baseline. Similarly, GRACE and ROME edits enhance plausibility scores, with MEMIT and GRACE achieving the highest overall performance. The strong results from expert annotations further validate the reliability of VERA's judgments, supporting the use of VERA in our framework as an effective commonsense evaluator to identify implausible knowledge requiring further editing. This approach reduces reliance on manual annotations while preserving robust assessment capabilities.

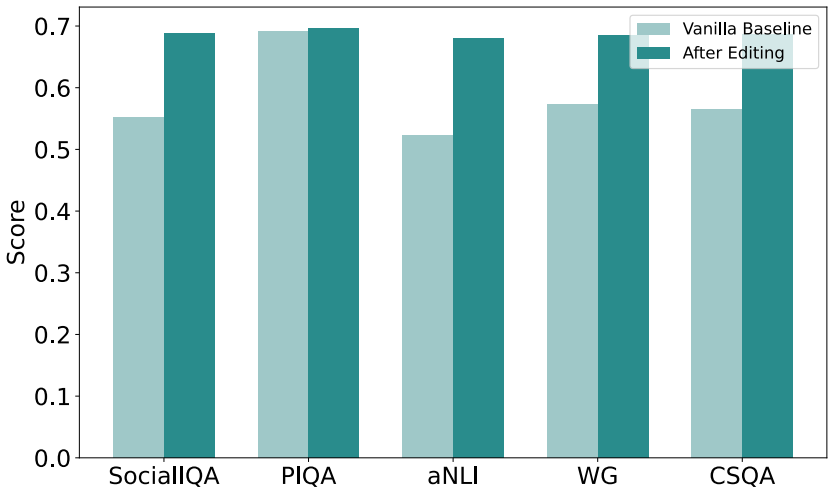


Figure 7.3: Performance of the best LLM after editing on five downstream tasks compared to the vanilla baseline.

7.4.2 Downstream Improvements

To assess whether enhanced internal commonsense reasoning improves downstream task performance, we evaluate the edited models on multiple commonsense reasoning benchmarks. Following [35], we test our framework on the validation splits of five widely-used commonsense QA benchmarks: Abductive NLI (aNLI [190]), CommonsenseQA (CSQA [21]), PhysicalIQA (PIQA [114]), SocialIQA (SocialIQA [80]), and WinoGrande (WG [191]). These benchmarks are designed to evaluate a range of knowledge types crucial for robust commonsense reasoning [38, 217, 245, 246]. We compare the performance of the best LLM edited with ConKE against its corresponding vanilla baseline across all benchmarks, with the results visualized in Figure 7.3. The results show that models edited with ConKE achieve significant performance improvements across all benchmarks, with particularly notable gains in aNLI and SocialIQA. These findings demonstrate the effectiveness of ConKE in enhancing commonsense reasoning capabilities and suggest its potential for broader applications in improving LLM performance on real-world reasoning tasks.

7.4.3 Ablation Study

Finally, to validate the effect of conceptualization, we conducted an ablation study on MEMIT by removing the conceptualization step and comparing performance. In this setup, we edit LLMs both with and without the integration of conceptualization and instantiation, and evaluate their performance by examining the average VERA scores of the generated outputs on the testing set. The conceptualized variant leveraged enriched commonsense triples generated via abstrac-

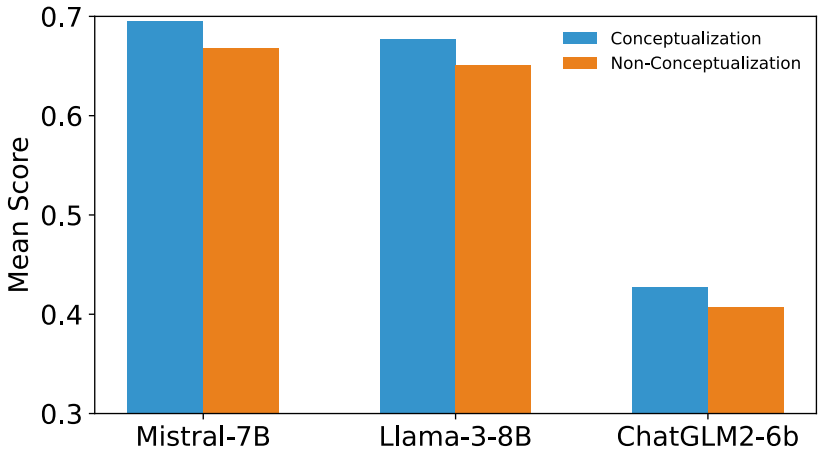


Figure 7.4: VERA evaluation scores of edited LLMs with and without integrating conceptualization.

tion and instantiation prior to the editing process, while the non-conceptualized variant directly applied MEMIT without these pre-processing steps.

Figure 7.4 demonstrates that the conceptualized variants consistently outperform their non-conceptualized counterparts, achieving higher plausibility and improved downstream task accuracy. These results suggest that the enriched conceptual patterns introduced before editing not only enhance plausibility but also enable the model to generalize commonsense knowledge to more complex reasoning tasks, ultimately boosting overall performance [247].

7.5 Conclusions

In this chapter, we presented ConKE, a knowledge editing framework for improving commonsense reasoning in modern LLMs. ConKE addresses two core obstacles in commonsense editing—limited knowledge coverage and poor scalability—by coupling automated verification with VERA and semantic enrichment via conceptualization and instantiation, enabling edits that generalize beyond isolated surface-form facts. Across intrinsic plausibility evaluation and multiple downstream commonsense QA benchmarks, the edited models consistently improve, supporting the effectiveness of our concept-guided editing pipeline.

Despite these gains, several challenges remain. First, commonsense knowledge is highly interconnected, so editing a single statement can trigger cascading changes in related concepts, leading to non-linear interactions that become harder to diagnose as the edited set grows. Second, iterative updates introduce the risk of knowledge drift, where successive edits may subtly conflict with or overwrite earlier edits, highlighting the need for stronger mechanisms to preserve global consistency over time. Third, commonsense often lacks a stable ground truth: it is

context-dependent, culturally variable, and frequently underspecified, which complicates standardization and evaluation. Addressing these issues will likely require more globally coordinated editing objectives, deeper theoretical understanding of edit propagation, and systematic human-in-the-loop validation to ensure that model updates remain coherent and aligned with broader consensus.

CHAPTER 8

FROM CONCEPTUALIZATION TO METAPHYSICAL REASONING

Conceptualization has served as the backbone of this thesis: it turns concrete events into reusable abstractions, exposes latent regularities, and enables models to generalize beyond memorized surface forms. Across the earlier chapters, we used this lens to answer a recurring question: when the world shifts—even slightly—can a model keep its understanding coherent, transferable, and controllable? We showed that conceptual structure helps LLMs acquire commonsense at scale, ground abstractions back into diverse contexts, and even update internal knowledge through targeted edits. Yet these advances also reveal a boundary: conceptualization, by itself, primarily organizes *what* is known; it does not fully specify *how* knowledge should behave when the underlying situation changes, the distribution shifts, or the model must reason about consequences under new conditions.

This chapter breaks out of the "kingdom of conceptualization" because the LLM era raises a more ambitious demand than static generalization: *adaptive reasoning under change*. Modern LLMs increasingly act as decision-making components—agents that plan, interact, and revise beliefs while operating in non-stationary environments. In these settings, robust intelligence is not only about answering questions in-distribution, but about anticipating how actions, contexts, and environmental factors transform what follows. A model must judge whether a changed action is plausible in reality, infer what that change would cause, and determine what further change is required to restore plausibility when the chain breaks. This capability sits at the core of planning and controllable behavior, and it is a crucial ingredient of System II reasoning: deliberate, counter-habitual generalization that remains reliable when assumptions drift.

However, reasoning with distributional change is difficult to study and even harder to evaluate. The space of possible changes is combinatorially large, many changes produce implausible events that terminate real-world trajectories, and existing benchmarks typically cover only a narrow set of perturbations or focus on action–state discrimination without modeling the *transition* induced by change. To confront this gap, we formalize a new target ability—*reasoning with changes in distribution*—as a three-step discriminative process: assessing whether a change

yields a plausible event, judging whether the resulting inference is plausible, and reasoning about what additional change would make an implausible inference plausible again. We call this process **metaphysical reasoning**: not in the traditional philosophical sense, but as a practical name for reasoning that navigates rare, highly abstracted, or reality-violating branches induced by conceptual or numerical shifts.

Crucially, conceptualization is not replaced in this transition—it becomes *the central mechanism that makes metaphysical reasoning tractable*. Distributional change cannot be enumerated; it must be represented. By expressing changes through hierarchical abstractions (and structured numerical/spatial variations), we can cover broad families of perturbations with a small set of conceptual operations. Conceptualization thus provides the "coordinate system" of change: it lets us define where a distribution shifts (which event component changes), how severe the shift is (how abstract the modification becomes), and which consequences should be expected downstream. In other words, metaphysical reasoning is the next stage of the thesis argument: once conceptualization gives models reusable structure, the natural next question is whether models can *reason over that structure when the world moves*. Thus, this final chapter answers that question by introducing a principled formulation and the first large-scale benchmark, MARS, designed to measure and stress-test LLMs' ability to remain coherent across feasibility, consequence, and transition under distributional change.

8.1 Introduction

Recent advances in LLMs have demonstrated superior performance in a variety of reasoning tasks [193, 194, 248–250]. However, to truly achieve conscious processing [251], the integration of System II reasoning ability [104, 252] is essential as it enables LLMs to perform out-of-distribution generalization when encountered with unfamiliar scenarios [107]. Among several components that make up System II reasoning, a critical element of it is the ability to *reason with situational changes in distribution*, triggered by *environmental factors* and *actions by themselves or other agents*, when dealing with non-stationarities [253]. It serves as the core ability in planning tasks [254], which can be achieved by dynamically recombining existing concepts in the given environment or action and learning from the resultant situational changes [255–257]. For instance, in the event that "PersonX is driving a car in a sunny day," a change in the weather from sunny to rainy could cause a different outcome, such as "PersonX becomes more cautious and drives slower." This illustrates that a change in weather conditions can lead to a change in the driver's behavior, which represents an environmental change that triggers situational changes

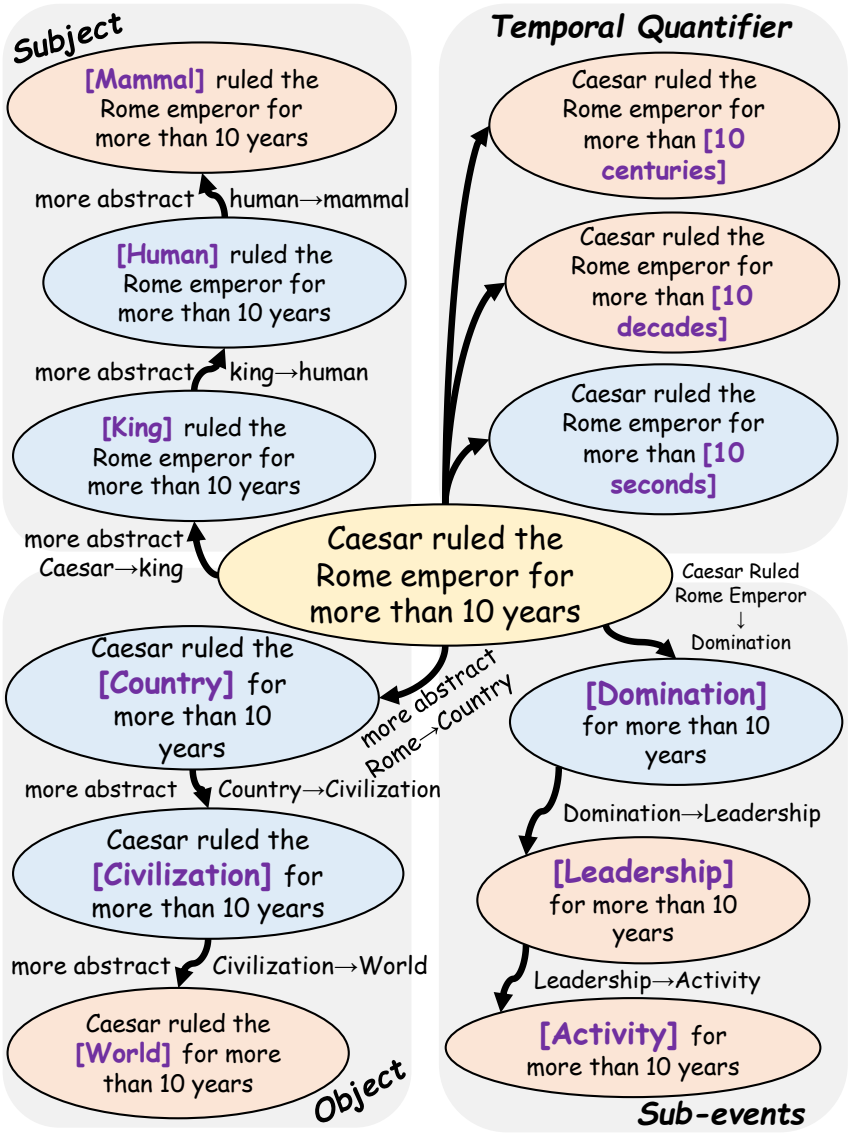


Figure 8.1: Examples of changes in event in our formulation. After changes occur, events may become **metaphysical** as components are abstracted into high-level concepts, while some remain plausible in reality.

within the distribution of different weathers.

Though fundamental, the exploration of this ability has been limited due to several factors. First, the scope for change within an event is vast, with numerous components capable of altering in a wide variety of ways. This results in an overwhelmingly large number of potential changes that are impossible to fully cover with existing knowledge bases. High-level representative ontology, such as abstraction [81], can be applied to represent a large number of changes simultaneously in different components, thereby forming hierarchical distributions. Second, *reasoning with changes in distribution* lacks a clear formulation due to its complexity. Unlike one-step inference reasoning tasks [128], changes in action may lead to implausible events that cannot occur in reality, thus terminating the reasoning process. Such type of changes require

extra care when designing evaluation protocols. A multi-step task formulation is necessary to emulate the entire process. Lastly, there is a lack of a reliable evaluation benchmark. Existing benchmarks [258, 259] typically focus on a limited number of changes within a few scenarios, thus limiting the coverage of formed distributions. The changes in actions and states are also formulated under planning or logical tasks, which neglect transitions (consequences) caused by changes.

To address these gaps, we take a step forward by formally defining *reasoning with changes in distribution* as a *three-step discriminative process*. We start by defining seven categories of changes, each corresponding to different components within an event. To semantically cover more changes in a unified manner, we propose implementing changes by altering each component within the event using their abstractions or numerical variations. This approach creates a hierarchical distribution of various changes, with the abstracted ones offering a more generalized coverage. Inspired by previous work [107], we formulate *reasoning with changes in distribution* as sequentially tasking the model to: (1) assess the plausibility of a potential change in a given event that describes an action, (2) evaluate the plausibility of an inferential state resulting from the modified action, and (3) determine the necessary change in an action to convert an implausible inferential state into a plausible one. This process effectively simulates *reasoning with changes in distribution* by sequentially reasoning through the necessary intermediate steps to comprehend changes. We refer to this process as ***metaphysical reasoning***–a term we adopt to describe a mode of reasoning that deals with highly improbable or abstract scenarios distinct from its traditional philosophical meaning or counterfactual reasoning–as it also requires models to distinguish implausible actions, states, and transitions that exist only in this abstract "metaphysical" realm, indicating their rare occurrence in reality [260]. We refer to this process as ***metaphysical reasoning***, as it also requires models to distinguish implausible actions, states, and transitions that only exist in the metaphysical realm, indicating their rare occurrence in reality [260].

We then construct the first evaluation benchmark, MARS, featuring 355K annotated data across three tasks corresponding to each step. It is constructed by sequentially instructing an LLM to extract events from Wikitext [261] and BookCorpus [262], identify mutable components within each event, generate abstractions and numerical variations for those components, create a metaphysical inference state based on the changes, and generate the necessary modifications to make the metaphysical inference plausible in reality. Large-scale human annotations are then conducted to provide labels of evaluation data entries and verify the quality of our benchmark. Extensive experiments with over 20 (L)LMs demonstrate that all three tasks in this

process present significant challenges, even for LMs after fine-tuning. Further analyses reveal potential reasons for such underperformance and identify possible solutions for enhancing the metaphysical reasoning abilities of language models.

On feasibility in metaphysical event generation. Another important issue is how to assess the feasibility of samples generated in the metaphysical reasoning framework. In this thesis, feasibility is not treated as an unconstrained imaginative property, but as an operational judgment grounded in the compatibility between an event, its modified components, and the resulting inferential state. Concretely, the construction of MARS already imposes several layers of control before an example is admitted into the benchmark: events are first extracted from naturally occurring text, mutable components are then localized within those events, and modifications are generated with respect to those components rather than by unconstrained free-form rewriting. This design narrows the space of possible changes and ensures that the modified events remain structurally comparable to the original ones.

More importantly, the benchmark does not rely solely on raw LLM generation to determine whether a modified event or inference is feasible. The generated entries are subsequently subjected to human annotation, which serves as the final quality-control mechanism for deciding whether the event change is plausible, whether the resulting inferential state is plausible, and whether the proposed transition can restore plausibility. In this sense, feasibility in MARS is operationalized through curated discriminative judgments rather than through a claim that the generation model itself provides a formal proof of realizability. The benchmark is designed to evaluate whether models can distinguish feasible from infeasible changes, not to assert that every generated hypothesis is inherently trustworthy prior to annotation.

This distinction is important because metaphysical reasoning lies precisely at the boundary between abstract variation and real-world constraint. Some modifications are unusual but still feasible; others are linguistically well-formed but incompatible with physical, social, or commonsense regularities. By combining constrained generation with large-scale human verification, MARS keeps the evaluation focused on this boundary. The resulting benchmark therefore reflects feasibility as judged under human commonsense and event-level coherence, which is appropriate for evaluating language models' ability to reason about distributional change, even though it does not amount to a complete symbolic or causal verification framework. Extending MARS with stronger formal feasibility checkers, such as domain-specific simulators or external constraint solvers, would be a valuable direction for future work.

In summary, the main contributions of this chapter are three-fold:

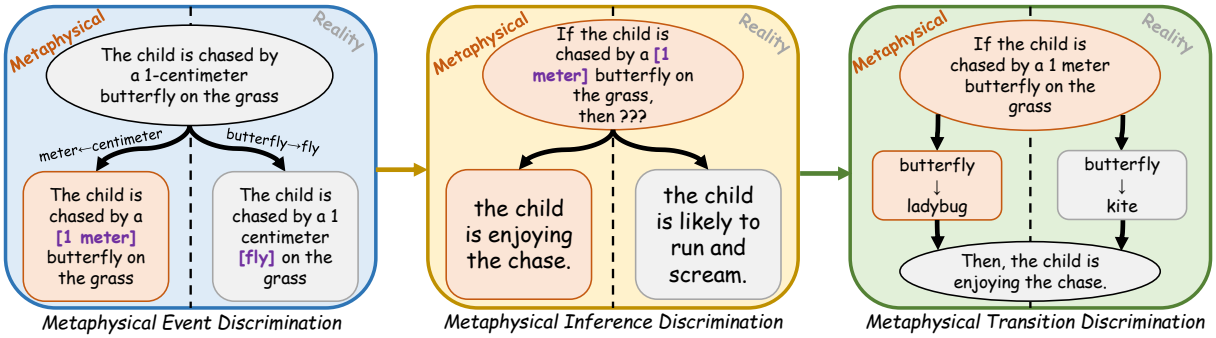


Figure 8.2: The three steps in metaphysical reasoning. Our motivation behind this is that, by conquering all steps sequentially, a conscious agent could answer: (1) Will the change occur in reality? (2) What will the change cause? (3) What change can make a **metaphysical** (desired) inference plausible?

- **Task**: We introduce the task of ***metaphysical reasoning***, which includes three distinct subtasks, designed to assess the ability of (L)LMs to *reason with changes in distribution*.
- **Resource**: We carefully curate a large-scale evaluation benchmark, MARS, to facilitate evaluations of LMs' metaphysical reasoning abilities. The dataset is released.
- **Evaluation**: We experiment with several (L)LMs to demonstrate the difficulty of our proposed tasks and conduct analysis to identify the reasons behind their underperformance.

8.2 Backgrounds and Related Works

Reasoning about Changes in Distribution. Enabling LMs to understand distributional changes due to localized causal interventions, particularly in semantic spaces, has long been a crucial objective in the pursuit of conscious machine intelligence [107, 263]. Previous works have mainly explored this within the context of discriminating changes between actions and states with methods such as commonsense knowledge injection [264], event calculus [265], and fuzzy reasoning [266]. Other studies aim to benchmark this reasoning process through logical reasoning tasks [259] and planning tasks [258, 267]. However, these studies only cover changes in limited formats and scenarios and also overlook the significance of representing changes as a distribution in relation to different variables in actions. Such loss restricts the out-of-distribution generalizability of the resulting LMs when facing unfamiliar scenarios. Moreover, previous evaluations do not cover transitions caused by changes, making subsequent evaluations around reasoning with changes incomplete. We address these issues by proposing to use the abstraction or numerical variations of components as changes to form generalizable distributions. We also design a task in MARS to evaluate LMs' proficiency in understanding how changes motivate situational transitions.

Benchmarking LLMs. The advent of LLMs [40, 41, 95–97] has sparked various studies in investigating LLM's potential in a variety of tasks, including text generation [268–270], temporal reasoning [271, 272], causal reasoning [193, 273, 274], commonsense reasoning [250, 275], and more [194]. The advent of LLMs [40, 41, 95–97] has sparked various studies in investigating LLM's potential in a variety of tasks [193, 194, 250, 269, 270, 272]. These studies have significantly contributed to our understanding of LLMs by evaluating their performance across diverse tasks, using different scales of parameters and prompting methods [276]. However, there is an absence of a comprehensive benchmark for assessing the ability of (L)LMs to *reason with changes in distribution*. This inspires us to formally define it and introduce the first benchmark that evaluates such reasoning capabilities of (L)LMs.

8.3 Definitions of Changes in Event and Metaphysical Reasoning

Modeling changes within an event is inherently complex due to the infinite number of changes that can occur. For simplicity, we only consider events that represent an action and study changes between their inferential states. Given an event e, we first define seven types of changes that could transpire within e. These changes are represented as components of the event, including its subject s, verb v, object o, temporal quantifier t, spatial quantifier l, numerical properties n, and sub-events se. The original event is denoted as a function of these seven components, $e = f(s, v, o, t, l, n, se)$. A change in the event can be represented by altering one of its components, for instance, $e' = f(s', v, o, t, l, n, se)$ if the change impacts the subject s'.

To effectively model the distribution of changes across different types of components, we leverage two types of hierarchical formulations. Specifically, for s, v, o, se, we define changes in these components as conceptualizing their original instance into three concepts with progressively increased abstractedness [108, 277]. For t, l, n, we define their changes as modifications from their original values to three distinct numerical or spatial values with progressively increased units. This brings a hierarchical structure to changes of a certain component, forming a distribution that gradually covers more possible changes. Abstracted components, as high-level concepts, can semantically represent a broader range of combinations for altering an event. Some running examples of how changes impact an action are shown in Figure 8.1. We then propose a *three-step discriminative process*, which we term as ***Metaphysical Reasoning***, to formulate *reason with changes in distribution*. The three steps, as shown in Figure 8.2, are:

(1) Metaphysical Event Discrimination: The first step answers the question, "Will the change happen in reality?" It aims to determine the plausibility of a change based on a given event, as alterations in components may lead to implausible events that defy reality. We refer to such an event, which rarely occurs in reality due to these changes, as a ***metaphysical event***. The goal of the first task is to discriminate whether the modified event e', conditioned on the original event e with a single altered component $c \in (s, v, o, t, l, n, se)$, is metaphysical or not by making a binary prediction.

(2) Metaphysical Inference Discrimination: Considering that distributional changes occur in non-stationary environments, a conscious agent should be able to predict the potential outcomes of the modified event for future reasoning scenarios. Therefore, the second step aims to answer the question, "What will the altered event result in?" Similarly, we term the inferences of an event that rarely occurs in reality as ***metaphysical inference***. The objective of the second task is to determine whether an inferential state i, triggered by the altered event e', is metaphysical or not by predicting a binary answer. Note that e' could be either metaphysical or not, as inferences in both cases can be evaluated.

(3) Metaphysical Transition Reasoning: Finally, with some inferences remain metaphysical, a conscious agent should be able to plan what change is necessary to make such inference plausible in reality. This completes the reasoning chain by covering the feasibility, consequence, and motivation of distributional changes. Thus, the last task answers the question, "What change is needed to make a metaphysical inference plausible?" We refer to this as ***metaphysical transition reasoning*** and set the objective as to determine whether another change, denoted as c', can make a metaphysical inference i plausible in relation to a changed event e' by making a binary prediction regarding c'.

8.3.1 Differentiation from Philosophical Metaphysics and Counterfactual Reasoning

In this thesis, we use the term "metaphysical" to describe a specific mode of reasoning that deals with highly improbable or abstract scenarios, distinct from both its traditional philosophical meaning and the concept of counterfactual reasoning. Philosophically, "metaphysics" refers to the study of the fundamental nature of reality, encompassing questions about existence, causality, and the nature of being [278, 279]. While this classical usage involves conceptual analysis and abstract thought, our focus diverges significantly. We adopt "metaphysical" to signify reasoning that examines transitions between plausible and highly improbable states, emphasizing

the logical structure and abstracted nature of these transitions rather than ontological or existential inquiries.

This distinction is important because our framework does not engage with the philosophical debates about the nature of reality or existence. Instead, it concentrates on how LLMs process and adapt to scenarios that are rare or abstract yet logically consistent. For example, while metaphysical reasoning in our context might involve reasoning about a scenario where "a civilization survives for 100,000 years," it does not explore the metaphysical nature of time, existence, or causality in a philosophical sense.

Furthermore, our concept of metaphysical reasoning is distinct from counterfactual reasoning. Counterfactual reasoning involves evaluating "what if" scenarios that diverge from known realities but remain bounded by plausible causal relationships [280, 281]. For example, a counterfactual might consider, "What if Caesar had lost the battle of Pharsalus?"–a scenario grounded in historical plausibility. In contrast, metaphysical reasoning in our framework extends beyond plausibility to explore scenarios that are structurally coherent but unlikely or abstract, such as "What if Caesar ruled for a millennium?" Here, the focus is not on causal plausibility but on the ability to evaluate transitions to rare, abstract, or highly improbable states.

This differentiation between "metaphysical" in our framework, metaphysics in philosophy, and counterfactual reasoning underscores the novel challenges our benchmarks aim to address. By pushing LLMs to reason about transitions into abstract or improbable scenarios, we aim to probe and enhance their capabilities for adaptive, out-of-distribution reasoning – a necessary step toward achieving generalizable System II reasoning.

8.4 MARS Benchmark Curation Pipeline

We then introduce our sequential pipeline for curating the MARS benchmark. An overview of our curation pipeline is shown in Figure 8.3. To guarantee a comprehensive coverage of events across various domains and topics, we source original text from two publicly available large corpora: Wikitext [261] and BookCorpus [262]. We filter out noisy text that includes hashtags and hyperlinks and segment long text into sentences with no more than 200 tokens to facilitate future processing.

8.4.1 Text Decomposition and Extraction

We first perform text decomposition [282, 283] to break down lengthy text into semantically complete short events, which are then used for fine-grained component extraction. To enable

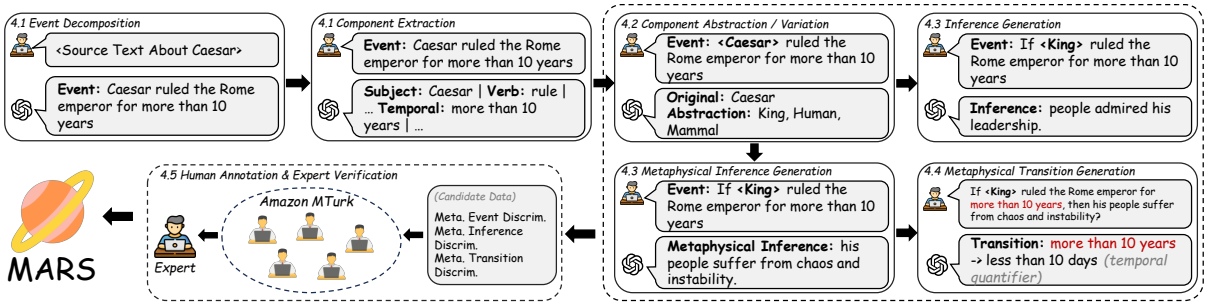


Figure 8.3: An overview of our benchmark curation pipeline with running examples.

large-scale processing, we use ChatGPT [40], a powerful LLM with strong text understanding abilities, as the core processor for all stages. For each stage, we guide it with a few-shot prompt [32, 42] by creating task-specific explanations and exemplars:

<TASK-PROMPT>

<INPUT$_1$**><OUTPUT**$_{(1,1)}$**>** ... **<OUTPUT**$_{(1,N_1)}$**>**

<INPUT$_2$**><OUTPUT**$_{(2,1)}$**>** ... **<OUTPUT**$_{(2,N_2)}$**>**

...

<INPUT$_{10}$**><OUTPUT**$_{(10,1)}$**>** ... **<OUTPUT**$_{(10,N_{10})}$**>**

<INPUT$_{11}$**>**

To perform text decomposition, **<TASK-PROMPT>** clarifies the goal to ChatGPT, which involves extracting semantically complete actions from the given text. **<INPUT**$_{1-10}$**>** and **<OUTPUT**$_{1-10}$**>** are filled with 10 pairs of human-crafted examples, each containing several action events extracted from text sampled from Wikitext and BookCorpus. ChatGPT is expected to learn from these examples and use them as a guide to extract action events (**<OUTPUT**$_{(11,1-N)}$**>**) from the final input text (**<INPUT**$_{11}$**>**). For component extraction, we adjust **<TASK-PROMPT>** to define the task of extracting the seven components from a given event. We populate **<INPUT**$_{1-10}$**>** and **<OUTPUT**$_{1-10}$**>** with 10 pairs of events and seven comma-separated lists of components extracted from the event, each corresponding to one type of components defined in §8.3. ChatGPT then extracts seven lists of components for the final given event (**<INPUT**$_{11}$**>**). If any type of component is absent, "None" will be generated instead.

8.4.2 Component Abstraction and Variation

The next step is designed to implement changes within the event by altering its components, extracted from the previous step, by generating their abstractions or numerical variations. Following prior work [98], we guide ChatGPT by modifying **<TASK-PROMPT>** with the objective of generating abstract concepts for s, v, o, se and numerical variations for t, l, n within a specified

Dataset / Task	#Text	#Event	#Avg.Token	#Train	#Dev	#Test	#Total.	#Unlabel.	Expert.
AbsATM [78]	N/A	7,196	1.060	107,384	12,117	11,503	131,004	372,584	N/A
AbsPyramid [81]	N/A	16,944	1.690	176,691	22,050	22,056	220,797	0	N/A
Meta. Event.	9,998	55,190	1.040	96,004	12,013	11,982	119,999	329,540	94.0%
AbsATM [78]	N/A	7,196	6.413	65,386	8,403	7,408	81,197	5,921,195	N/A
Meta. Inference.	9,837	35,528	10.40	96,009	12,010	11,981	120,000	497,590	96.5%
Propara [284]	9,051	9,051	N/A	7,043	913	1,095	9,051	0	N/A
TRAC [259]	15,000	15,000	N/A	10,000	2,000	3,000	15,000	0	N/A
PlanBench [258]	26,250	26,250	N/A	0	0	26,250	26,250	0	N/A
Meta. Transition.	9,677	31,447	1.810	92,495	11,563	11,560	115,618	273,474	93.5%

Table 8.1: Statistics of the MARS benchmark in comparison against other benchmarks. Meta. refers to three tasks in MARS. Expert. refers to expert verification results.

event. For each **<INPUT$_{1-10}$>** and **<OUTPUT$_{1-10}$>** pair, we populate the input with a specific event and one of its components. The output consists of three human-authored component abstractions or numerical variations that align with the event's context. Subsequently, ChatGPT is tasked with generating three abstractions or numerical variations for the final pair of the given event and a component within the event (**<INPUT$_{11}$>**). Replacing the original components in the event with their generated changes forms changed event candidates for the metaphysical event discrimination task.

8.4.3 Inference Generation

We then collect inferential states of the modified events by similarly instructing ChatGPT to autonomously generate them. For each altered event, we prompt ChatGPT to separately generate one plausible inference and one metaphysical inference. We first modify **<TASK-PROMPT>** to generate a state that could potentially be caused by the altered event, and populate **<INPUT$_{1-10}$>** with 10 modified events and **<OUTPUT$_{1-10}$>** with 10 corresponding plausible inferences authored by human experts. ChatGPT is then requested to generate an additional plausible state inference for the given changed event (**<INPUT$_{11}$>**). Next, we adjust **<TASK-PROMPT>** to generate a metaphysical state that is infrequently caused by the changed event in reality, yet remains contextually relevant. We replace **<OUTPUT$_{1-10}$>** with 10 metaphysical inferences and then collect a metaphysical inference from ChatGPT. This, along with the generated plausible inference, forms two candidate data entries for each changed event in the metaphysical inference discrimination task.

Figure 8.4: Hypernym distribution of the top 5,000 popular component variations.

8.4.4 Metaphysical Transition Generation

Given that half of the inferential states generated in the previous step remain metaphysical, we then collect the additional changes necessary to transform these states into plausible real-world inferences. We adjust the **<TASK-PROMPT>** to describe such required changes and populate **<INPUT$_{1-10}$>** with 10 pairs of modified events and their corresponding metaphysical inferences. **<OUTPUT$_{1-10}$>** are filled with 10 corresponding human-authored changes in events that can render the inferences plausible. Subsequently, ChatGPT generates the required change for the final pair of the modified event and its metaphysical inference (**<INPUT$_{11}$>**). Note that the generated change still needs to be one of the seven types we defined in §8.3. We collect one additional change for each metaphysical inference and use it as a candidate data entry for the last task. However, we discard event and inference pairs that ChatGPT deems impossible to render plausible, even with an additional change.

8.4.5 Human Annotations

Annotation: Finally, we carry out large-scale human annotations to label candidate data for each task via Amazon Mechanical Turk (AMT). We provide detailed instructions with examples to qualified workers and task them with annotating (1) the plausibility of the changed events generated in §8.4.2, (2) the plausibility of the plausible/metaphysical inferences produced in §8.4.3, and (3) the plausibility of the transitions generated in §8.4.4. We collect five votes for each entry and the majority vote is used as the final label. The overall inter-annotator agreement (IAA) is 81% in terms of pairwise agreement, and the Fleiss Kappa [285] is 0.56, indicating sufficient agreement.

Expert Verification: To verify the quality of our collected labels, we recruit three postgraduate students with rich experience in NLP to perform a second round annotation. Each of them is asked to annotate a sample of 100 data entries for each task, following the same instructions provided to the AMT annotators. Results in Table 8.1 show that, on average, 93.67% labels collected from human annotations align with the expert's vote, demonstrating the reliability of our collected labels.

8.5 Evaluations and Analysis

8.5.1 MARS Statistics

Table 8.1 presents statistics of the MARS benchmark, which comprises a total of 355,617 annotated data distributed across three tasks. We partition the annotated data into training, development, and testing splits following an 8:1:1 ratio, ensuring there is no overlap of text and events between the different splits to preserve the evaluation's generalizability. On average, 1.04 tokens are generated to describe changes in action for the metaphysical event and transition discrimination tasks, while 10.4 tokens are used for inferences in the metaphysical inference discrimination task. To the best of our knowledge, we are the first in proposing such a triad of tasks concurrently within a single benchmark. To compare MARS with other datasets, we select those with analogous task objectives for each task and compare them individually. We find MARS tends to be significantly larger than other benchmarks, covering a broader range of events and providing training sets for evaluating the performance of fine-tuned models. Each task contains over 90,000 training data and more than 10,000 data for validation and testing. Unannotated data remain unlabeled for future research, such as proposing semi-supervised

Backbone	Training Data	Event			Inference			Transition		
		Acc	AUC	Ma-F1	Acc	AUC	Ma-F1	Acc	AUC	Ma-F1
DeBERTa *435M*	Zero-shot	58.27	49.88	45.87	47.73	49.94	44.44	50.73	46.96	46.15
	CANDLE	57.94	58.22	57.31	59.43	59.03	58.18	62.00	62.19	61.50
	MARS	64.45	64.16	63.27	69.57	71.15	69.33	72.93	74.00	72.01
	CANDLE + MARS	**64.95**	**64.27**	**63.74**	**71.85**	73.32	71.64	**74.39**	77.97	73.30
VERA *11B*	Zero-shot	41.82	50.48	38.52	60.97	62.54	59.09	61.31	66.32	61.17
	CANDLE	57.81	57.24	56.77	56.59	56.08	55.25	59.79	59.88	59.19
	MARS	61.95	61.43	60.81	63.90	66.93	70.84	71.75	74.57	73.27
	CANDLE + MARS	62.21	61.77	61.17	71.45	**74.46**	67.61	73.95	77.35	**78.26**
LLaMa-3 *8B*	Zero-shot	50.62	-	49.12	51.33	-	50.98	51.95	-	51.07
	CANDLE	56.47	56.75	56.07	58.29	57.81	57.00	58.74	58.81	58.19
	MARS	60.06	60.54	59.58	65.76	67.88	65.72	69.83	74.59	68.74
	CANDLE + MARS	60.93	60.80	60.12	69.13	70.84	**72.12**	74.09	**79.38**	71.42

Table 8.2: Evaluation results (%) of transfering knowledge from CANDLE to aid MARS. The best performances among each method is underlined and best ones among all methods are **bold-faced**.

methods [85, 286] to tackle MARS. To further illustrate the diverse coverage of events and changes in MARS, we match each component variation against hypernyms in Probase [62] and plot their distribution according to their number of occurrences in Figure 8.4. Our results indicate that MARS covers over 170,000 hypernyms in Probase, spanning broad categories such as event, activity, concept, unit, etc.

8.5.2 Main Evaluations on MARS

Task Setup and Model Selections

We then experiment with a selection of (L)LMs to investigate their performances on our curated MARS benchmark. Accuracy, AUC, and Macro-F1 scores are used as evaluation metrics. Following the task definitions in §8.3, each task is assessed as a binary classification task. The models are tasked with determining (i) whether a modified event qualifies as a metaphysical event, (ii) whether the inference drawn from the modified event is a metaphysical inference, and (iii) whether changes in the event motivate the transition from a metaphysical inference to a plausible inference. For unbiased evaluations, we use accuracy, ROC-AUC, and Macro-F1 scores as evaluation metrics.

The evaluation of different models are categorized into three types: **(1) Zero-shot:** We first evaluate several (L)LMs in a zero-shot manner. For small-sized Pre-Trained Language Models (PTLMs), we evaluate DeBERTa-v3 [77], GPT2 [91], CAR [111], CANDLE [98], and VERA [102], following the design of zero-shot question answering [35]. For LLMs, we evalu-

ate LLaMa2, LLaMa3, LLaMa3.1 [96, 97, 224], Gemma [110], Falcon [287], and Mistral [213] using direct zero-shot prompting [194]. **(2) FINETUNING:** We then assess the performance of (L)LMs when fine-tuned on the training set of MARS. For PTLMs, we fine-tune DeBERTa, GPT2-xl, and VERA. For LLMs, we fine-tune LLaMa2, LLaMa3, Gemma, and Mistral using LoRA [214]. **(3) LLM API:** Finally, we evaluate the performance of GPT-4 [41] and GPT-4o-mini [223], which represent proprietary LLMs, under zero-shot, five-shots, Chain-of-Thought prompting (COT [196]), and Self-Consistent COT (SC-COT [216]) settings. For LLaMa3.1-70B and GPT-4o-mini, we also test their performances with RAG [288], Multi-agent Calibration [289], and Self Reflection [290]. We also add Random and Majority voting to illustrate the characteristics of MARS.

Results and Analysis

Evaluation results are reported in Table 8.5. From the results, we observe that: **(1) Most models exhibit subpar performance under the zero-shot setting.** Among PTLMs, only VERA delivers acceptable results across all three tasks, while the rest significantly underperform. This indicates the extreme difficulty of our proposed metaphysical reasoning tasks. Though models fine-tuned on commonsense knowledge and conceptualizations, such as CAR and CANDLE, show some improvement compared to their DeBERTa-v3-Large backbone, these performances are still unsatisfactory, even falling below the level of majority voting. For instance, CAR's accuracy improves by 3.63%, 0.6%, and 2.24% on the three tasks. For LLMs, improving training paradigms and increasing the number of parameters can indeed help achieve better performance. Nevertheless, all models perform poorly across all tasks in MARS, emphasizing the difficulty of our tasks. **(2) Fine-tuning only offers limited benefits.** With fine-tuning, all models improve significantly. For example, DeBERTa-Large's accuracy increases by 16.18%, 21.84%, and 22.2% on three tasks, respectively. However, the best results for all tasks are still capped at around 74%, indicating a shared difficulty and significant room for future enhancements. One potential reason for this is that, since we split the data according to the source of text in Wikitext and BookCorpus, the distribution between different splits may differ significantly, as the domain and topics could be diverse from each other. **(3) The GPT series models underperform compared to other LLMs, and COT does not consistently aid performance.** Surprisingly, GPT series models fall short when compared to open LLMs, such as LLaMa-3-70B. One possible explanation is that negative examples in MARS are sourced from ChatGPT's generation and are obtained via post-human annotation. This makes it challenging to discriminate as these negative examples contradict GPT's internal knowledge. Advanced prompting methods only offer

Component Type	Identified				Modified			
	ME.	MI.	MT.	#Avg.	ME.	MI.	MT.	#Avg.
Subject	4,376	3,907	3,507	1.116	3,106	2,950	2,591	1.094
Verb	9,874	8,856	8,061	3.647	4,408	4,146	3,760	3.457
Object	12,645	11,302	9,986	1.760	5,949	5,494	4,865	1.703
Temporal Quantifier	3,003	2,560	2,288	0.472	1,394	1,253	1,110	0.435
Spatial Quantifier	3,866	3,741	3,301	0.459	2,064	1,979	1,718	0.476
Numerical Properties	5,619	4,932	4,355	0.652	3,570	3,353	2,920	0.612
Sub-events	419	385	326	0.040	425	402	332	0.037
Total	39,802	35,683	31,824	8.146	20,916	19,577	17,296	7.814

Table 8.3: Number of unique components by type in annotated splits of MARS. #Avg. refers to the average number of unique identified/modified component per event.

limited improvement in performances.

8.6 Analysis

8.6.1 Transferring from Conceptualization

Improving the performance of LLMs on MARS requires extensive fine-tuning on large-scale human-annotated data, making it non-trivial. Since we observe that approximately 80% of action changes are executed by modifying a component along with its abstracted concepts (see Table 8.3), we first study whether exposing LLMs to more conceptualizations and abstract knowledge can enhance their metaphysical reasoning capabilities. For this purpose, we select CANDLE [98] as the knowledge source, which is an automatically constructed knowledge base containing 382K conceptualizations of events and abstract inferential knowledge. We first convert event-conceptualization pairs into the task format of metaphysical event discrimination and reformat commonsense inferential knowledge to align with the objectives of the metaphysical inference and transition discrimination tasks.

Three backbone models are then fine-tuned separately on CANDLE and MARS. Another group is sequentially fine-tuned on CANDLE and then on MARS. All models are then evaluated on the testing set of MARS, with the results reported in Table 8.2. From the results, a significant improvement is observed across all tasks when the models are sequentially fine-tuned on CANDLE and MARS, compared to solely fine-tuning on CANDLE or MARS.

These findings indicate that the transfer of conceptualizations and abstract knowledge from CANDLE effectively enhances the performance of LMs in metaphysical reasoning tasks. Since CANDLE is constructed by distilling from an LLM without human labor, this opens up a scalable

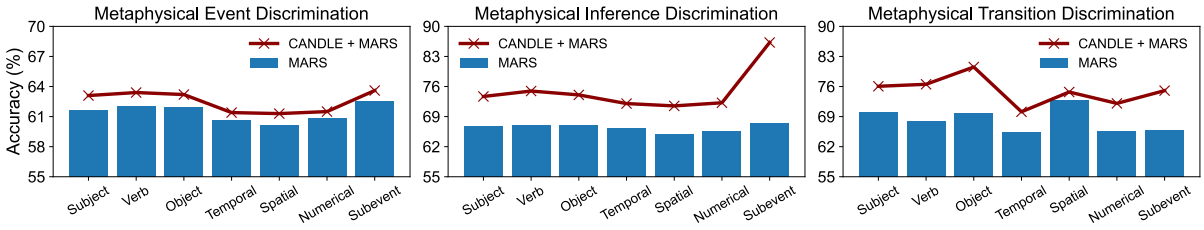


Figure 8.5: Performances by component types of fine-tuned LLaMa3-8B on three tasks of MARS.

and cost-efficient approach to improving the metaphysical reasoning capabilities of LLMs.

8.6.2 Impact of Component Types

We then analyze the performance of LLMs on each component type to understand the reasons for their subpar performance. We select LLaMa-3-8B as the representative model and compare its accuracy on each component type when fine-tuned on MARS and CANDLE + MARS. The results are illustrated in Figure 8.5. We observe that while pre-training the model on CANDLE consistently enhances performance, LLaMa3 still struggles when reasoning with changes in spatial quantifiers, temporal quantifiers, and numerical properties. This is in line with recent studies that demonstrate weaknesses in temporal and numerical reasoning for LLMs [271, 291]. Another possible reason is that since CANDLE only contains conceptualizations for subjects, verbs, objects, and sub-events in social events, pre-training models on it cannot provide benefits for the aforementioned aspects of change. Moreover, we only observe limited improvement for the metaphysical event discrimination task. Future works could focus on how to further enhance LLM's metaphysical reasoning capabilities in these weaker dimensions.

8.6.3 Error Analysis of GPT-Series Models

Finally, we select GPT4 as a representative model and conduct a manual analysis to identify the causes of errors by categorizing the mistakes found in their COT responses. We sample 150 COT responses from each task, all of which result in inconsistent results compared to human annotated labels and present our classifications of these errors as follows:

- **Hallucinations**: 41.7% of errors are caused by factual or metaphysical hallucinations by GPT4, where it creates a context that accommodates changes in actions and inferences that are not mentioned in the original text [292, 293]. For instance, in the event "The poet enjoys writing poems about western festivals," GPT4 incorrectly interprets the poet as Du

Fu, an ancient Chinese poet. This leads to a conflict when reasoning about his life and the subsequent inference "He was famous in the west," resulting in faulty reasoning.

- **Confusion between Concepts and Hypernyms**: 36.3% errors are attributed to GPT4's tendency to perceive abstract components within changed actions as hypernyms that fulfill the change, without considering all potential entities within the original concept. For instance, in a modified event, "He jumps down from *very high altitude* and lands peacefully," GPT4 interprets *very high altitude* as a diving platform, deeming it plausible. However, this concept could also encompass high buildings, which would not be suitable for the event.
- **Internal Conflict**: 17.7% errors are attributed to internal conflicts within GPT4's reasoning rationales, as well as inconsistencies between the binary predictions made and the corresponding reasoning rationales.
- **Annotation Error**: 4.3% errors are erroneously identified due to incorrect labels, potentially caused by spamming or a misunderstanding of the task by human annotators.

8.6.4 Multi-task Fine-tuning on MARS

Setup

To achieve conscious processing, an ideal language model should be capable of performing three tasks uniformly and sequentially. However, fine-tuning each task separately contradicts this objective, as it results in a model that can only perform one task after one training. Therefore, in this section, we investigate the possibility of enabling a language model to master all tasks simultaneously through multitask fine-tuning. Given that all three tasks are binary classification tasks, we adopt a straightforward approach. The language model is trained using a randomly shuffled combination of training data from all three tasks. This anticipates that the model will learn all tasks collectively. The best checkpoint is chosen based on achieving the highest accuracy on the validation sets of all three tasks. After training, the model performance is evaluated separately on the testing sets of each task.

Results and Analysis

The results are presented in Table 8.6. Upon analyzing these results, we observe that LLMs fine-tuned in a multi-task setting generally outperform those simply fine-tuned on the respective training data for each task. This observation is interesting as it suggests that training the model uniformly across all three tasks can enhance the entire process simultaneously, thereby improv-

Data Split	Evaluation Method	Event-ACC	Inference-ACC	Transition-ACC
MARS	Zero-shot	53.90	51.20	49.41
MARS	Few-shot	49.85	51.47	48.88
MARS-Claude	Zero-shot	54.50	54.00	53.50
MARS-Claude	Few-shot	56.00	55.50	54.00
MARS-LLAMA3.1	Zero-shot	52.00	56.50	56.50
MARS-LLAMA3.1	Few-shot	55.50	57.50	58.50

Table 8.4: Evaluation results (%) of GPT-4o on MARS constructed with different backbone LLMs.

ing reasoning with changes in distribution. This implies that LLMs can potentially mimic human learning abilities, which are better equipped to reason with changes by collectively understanding the feasibility, consequence, and necessity of such changes. Such a phenomenon indirectly indicates that our task formulation is indeed interconnected and collectively forms a reasoning pipeline. However, it's important to note that this improvement is only marginal. LLMs still exhibit limited metaphysical reasoning ability, particularly in the metaphysical event discrimination task. More advanced methods are still required to enable LLMs to achieve metaphysical reasoning.

8.6.5 Few-shot Fine-tuning on MARS

Setup

From the main evaluation results in Table 8.5, it is evident that fine-tuning consistently enhances the performance of all models on MARS. In this section, we delve deeper into the impact of fine-tuning in a few-shot setting, with the aim of analyzing the performance of models trained with limited data. More specifically, we aim to examine how models perform with varying sizes of training data. This will enable us to determine whether collecting more data invariably benefits fine-tuning, thereby leading to the development of more robust metaphysical reasoners. To achieve this, we sample the training data for each task in a progressively increasing ratio of 0.2, 0.4, 0.6, 0.8, and 1.0, and use each sampled training data to fine-tune LLMs for each task individually. The models are then evaluated on the complete validation sets to select the optimal checkpoint, and on the full testing set for performance assessment.

Results and Analysis

The results are reported in Table 8.7. From these results, we observe that training the model with a few-shot training data sample generally has a negative impact across all tasks in MARS.

However, this impact is not significant, and on rare occasions, the sampled training data even leads to superior results compared to training on the full sets. When the training data is reduced to different ratios (80%, 60%, 40%, and 20%), the performance of the models is not significantly affected. This suggests that the models are capable of learning from a small amount of training data and that performance is not significantly influenced by the size of the training data. In other words, annotating more data for training does not necessarily result in better performance, indicating that our task cannot be simply resolved by increasing training data. Future research can explore more advanced reasoning paradigms or training methods to further enhance the capabilities of LLMs in metaphysical reasoning.

8.6.6 Fine-tuned PTLMs vs. Fine-tuned LLMs

To validate the reason why fine-tuned PTLMs perform better than fine-tuned LLMs, we first hypothesis that PTLMs have a faster convergence rate to the training data due to their smaller number of parameters and fully fine-tuned paradigm (compared to LoRA when fine-tuning LLMs). This results in better fine-tuned performance than LLMs. Although LLMs have lower performance, they exhibit stronger generalizability to other tasks. We fine-tune a DeBERTa-v3 model with 25% and 50% of the training data and observed their performance in Table 8.7. From the results, we observe that when we reduce the training data for PTLMs, they are hardly comparable to fine-tuned LLMs. However, the last 50% of randomly sampled data brought significant improvements. While we cannot determine the exact reason due to the black box nature of these language models, we believe that PTLMs have a faster rate of fitting into the distribution of the training data or human annotations, resulting in better outcomes on human-annotated evaluation sets. LLMs are more likely to learn how to make correct inferences rather than simply fitting the data. Another possible reason is that we use LoRA to fine-tune LLMs due to limited computational resources; fully fine-tuning LLMs might further enhance their performance.

8.6.7 Inherent Bias in MARS Construction

One concern regarding the MARS benchmark is the potential bias introduced by using GPT-series models, specifically ChatGPT, for dataset construction. Our approach to constructing MARS was guided by the need to balance scalability with quality. In pilot studies evaluating metaphysical reasoning across various models, GPT-series models consistently demonstrated the highest levels of creativity and reliability. Based on these findings, we selected GPT as the primary backbone for data generation. Constructing MARS, however, required extensive manual an-

notation, as LLMs often fail to provide accurate labels for complex reasoning tasks. This manual verification process made it impractical to create multiple versions of MARS using different backbone LLMs due to expensive human labors required. Thus, to address concerns about potential biases arising from reliance on ChatGPT, we conducted additional experiments by constructing two smaller versions of the MARS benchmark. These alternative benchmarks utilized data generated from two different LLMs, Claude-3.5-sonnet [294] and LLAMA 3.1-70B [224], in each step, to obtain 200 evaluation data entries per task in MARS. All samples underwent expert annotation to collect ground-truth labels. We then evaluate GPT-4's zero-shot and few-shot performance on these alternative benchmarks alongside the original MARS.

The results are shown in Table 8.4. We observe that using different LLMs as backbones for MARS construction results in similar performance by GPT-4 across zero-shot and few-shot settings. Overall, the difficulty of the MARS benchmark remains robust and consistent, irrespective of the backbone LLM used during dataset generation. These experiments demonstrate that the reliance on ChatGPT for the original MARS construction does not compromise the benchmark's validity or difficulty. The results reinforce the reliability of MARS as a comprehensive test of metaphysical reasoning, with its complexity surpassing any potential biases introduced by the specific LLM used in data collection.

8.6.8 Binary Task Design in MARS

In MARS, all tasks are designed as a binary prediction task to facilitate automated and easy label collection and evaluation. Here, we discuss the reason and some pilot analysis behind such task design by considering other task formulations, including multiple-choice, open-ended generation, and binary evaluation.

Multiple-choice tasks, while structured and amenable to automated evaluation, posed significant challenges in collecting high-quality negative (distractor) options. Relying on human annotators to create distractors proved labor-intensive and impractical for scaling, as it required drafting multiple plausible but incorrect options for each question. As a result, we adopted open-ended generation and binary evaluation, ultimately choosing a generate-then-annotate paradigm. This approach involved two stages: first, evaluating the performance of LLMs in generating metaphysical cases during the generation phase; second, annotating the generated cases with binary labels (correct/incorrect).

To complement the binary evaluation results, we also included human annotation results for ChatGPT's performance in generating metaphysical data, as indicated in the *Majority* row of Ta-

ble 8.5, which can be regarded as following a generative task paradigm. The results demonstrate that, even when the task is framed as a generation task, ChatGPT struggles with metaphysical reasoning. The low proportion of human-annotated correct generations highlights the difficulty of reasoning about metaphysical changes, regardless of task formulation. While binary evaluation offers clear performance metrics and scalability advantages, the generation task provides complementary insights into the model's creative and reasoning capabilities. Together, these observations underscore the importance of improving LLMs' ability to reason about distributional and situational changes, which is crucial for advancing their metaphysical reasoning capabilities.

8.7 Case Studies

In this section, we present some examples for each of the three tasks in MARS to help readers better understand our benchmark. The examples are displayed in Table 8.8. We observe that examples in MARS typically require careful reasoning and consideration of the plausibility of occurrences in reality or the metaphysical realm to make the correct discrimination.

8.8 Conclusions

In conclusion, this chapter proposes ***Metaphysical Reasoning*** to delineate the process of *reasoning with changes in distribution* and construct MARS as the associated evaluation benchmark in a non-trivial manner. Our extensive experiments show the difficulty of our proposed task, which cannot be easily addressed through advanced prompting and fine-tuning. Further analysis reveals why LMs encounter difficulties with metaphysical reasoning tasks and suggests a potential avenue for improvement. We hope to illuminate the path toward achieving conscious processing in LLMs through System II reasoning by effectively comprehending changes in distribution.

Methods	Backbone	Event			Inference			Transition		
		Acc	AUC	Ma-F1	Acc	AUC	Ma-F1	Acc	AUC	Ma-F1
Random	-	50.00	-	49.56	50.00	-	49.56	50.00	-	49.56
Majority	-	60.98	-	37.99	58.56	-	36.93	50.25	-	33.37
PTLM *(Zero-shot)*	RoBERTa-Base *211M*	38.60	49.40	27.90	44.30	55.11	30.80	51.13	53.37	38.36
	RoBERTa-Large *340M*	38.57	50.94	27.83	44.37	56.49	30.73	50.90	53.08	33.92
	DeBERTa-Base *214M*	60.55	49.41	42.89	50.10	47.57	48.96	49.05	41.32	33.19
	DeBERTa-Large *435M*	48.27	49.88	45.87	47.73	49.94	44.44	50.73	46.96	46.15
	GPT2-XL *1.5B*	38.62	51.12	27.93	44.40	51.88	31.45	49.92	48.35	48.09
	CAR *435M*	54.63	49.34	49.96	48.33	42.85	41.93	52.97	35.05	46.94
	CANDLE *435M*	51.90	49.12	50.30	46.77	44.03	38.48	53.49	34.95	47.95
	VERA *11B*	51.82	50.48	48.52	60.97	62.54	59.09	61.31	66.32	61.17
PTLM *(Fine-tuned)*	RoBERTa-Base *211M*	63.32	62.76	61.76	69.08	70.54	68.90	71.24	72.73	70.65
	RoBERTa-Large *340M*	64.22	63.18	62.62	69.04	70.63	68.90	69.68	71.70	68.73
	DeBERTa-Base *214M*	63.82	63.98	**63.39**	69.50	70.59	69.31	71.96	73.85	71.17
	DeBERTa-Large *435M*	**64.45**	**64.16**	63.27	**69.57**	**71.15**	69.33	**72.93**	74.00	72.01
	GPT2-XL *1.5B*	46.68	47.63	46.96	43.70	44.22	30.41	44.57	45.03	45.89
	VERA *11B*	61.95	61.43	60.81	63.90	66.93	**70.84**	71.75	**74.57**	**73.27**
LLM *(Zero-shot)*	Meta-LLaMa-2-7B	50.64	-	41.41	49.87	-	49.23	50.94	-	50.64
	Meta-LLaMa-2-13B	51.50	-	49.48	50.81	-	50.57	50.81	-	50.80
	Meta-LLaMa-2-70B	52.40	-	49.03	56.13	-	46.81	48.45	-	48.34
	Meta-LLaMa-3-8B	50.62	-	49.12	51.33	-	50.98	51.95	-	51.07
	Meta-LLaMa-3-70B	57.41	-	50.59	63.40	-	61.82	60.15	-	60.01
	Meta-LLaMa-3.1-8B	51.01	-	50.27	52.13	-	51.29	52.35	-	52.09
	Meta-LLaMa-3.1-70B	59.22	-	52.08	63.61	-	61.90	61.28	-	61.03
	+RAG	61.21	-	54.51	66.38	-	65.90	61.53	-	61.22
	+Multi-Agent	56.12	-	51.08	65.06	-	65.01	62.54	-	62.19
	+Self-reflection	57.94	-	53.17	63.91	-	63.51	60.92	-	60.77
	Meta-LLaMa-3.1-405B	60.01	-	52.99	64.52	-	63.23	61.74	-	61.76
	Gemma-2-9B	56.88	-	48.53	51.83	-	51.76	49.41	-	45.01
	Falcon-7B	54.32	-	49.51	51.77	-	50.30	50.42	-	49.02
	Falcon-40B	52.35	-	50.36	49.67	-	49.38	50.27	-	50.22
	Mistral-7B	49.90	-	48.94	50.23	-	50.06	51.75	-	51.75
LLM *(Fine-tuned)*	Meta-LLaMa-2-7B	60.10	59.90	59.00	63.51	66.44	62.55	66.06	70.38	65.12
	Meta-LLaMa-2-13B	60.67	60.64	60.00	64.61	67.67	63.59	68.22	72.19	66.37
	Meta-LLaMa-3-8B	60.06	60.54	59.58	65.76	67.88	65.72	69.83	74.59	68.74
	Gemma-2-9B	61.23	61.25	60.28	69.24	70.76	69.00	73.30	76.91	69.18
	Mistral-7B	60.35	60.77	60.07	66.91	70.06	65.95	71.87	75.47	68.53
LLM *(API)*	ChatGPT	51.00	-	50.35	61.35	-	57.63	60.40	-	60.12
	ChatGPT (5-shots)	53.61	-	53.28	58.05	-	57.42	62.40	-	59.35
	ChatGPT (COT)	53.20	-	52.61	50.40	-	50.32	49.95	-	49.83
	ChatGPT (SC-COT)	53.98	-	53.47	52.47	-	51.99	51.25	-	51.13
	GPT4	53.90	-	53.45	51.20	-	50.95	49.41	-	49.33
	GPT4 (5-shots)	49.85	-	49.58	51.47	-	51.30	48.88	-	48.73
	GPT4 (COT)	51.28	-	50.73	51.49	-	51.35	47.62	-	47.58
	GPT4 (SC-COT)	51.97	-	51.26	52.05	-	52.27	48.24	-	48.11
	GPT-4o-mini	57.94	-	57.91	53.84	-	53.53	48.06	-	48.06
	+RAG	59.99	-	59.97	54.54	-	54.21	49.39	-	49.19
	+Multi-Agent	54.21	-	53.17	52.76	-	52.26	46.94	-	46.70
	+Self-reflection	56.89	-	55.21	53.22	-	53.20	48.51	-	48.45

Table 8.5: Evaluation results (%) of various language models on the testing sets of MARS. The best performances within each method are underlined and the best among all methods are **bold-faced**.

Methods	Backbone	Event			Inference			Transition		
		Acc	AUC	Ma-F1	Acc	AUC	Ma-F1	Acc	AUC	Ma-F1
Random	-	50.00	-	49.56	50.00	-	49.56	50.00	-	49.56
Majority	-	60.98	-	37.99	58.56	-	36.93	50.25	-	33.37
LLM ***(Zero-shot)***	Meta-LLaMa-2-7B	50.64	-	41.41	49.87	-	49.23	50.94	-	50.64
	Meta-LLaMa-2-13B	51.50	-	49.48	50.81	-	50.57	50.81	-	50.80
	Meta-LLaMa-2-70B	52.40	-	49.03	56.13	-	46.81	48.45	-	48.34
	Meta-LLaMa-3-8B	50.62	-	49.12	51.33	-	50.98	51.95	-	51.07
	Meta-LLaMa-3-70B	57.41	-	50.59	63.40	-	61.82	60.15	-	60.01
	Gemma-1.1-7B	56.88	-	48.53	51.83	-	51.76	49.41	-	45.01
	Falcon-7B	54.32	-	49.51	51.77	-	50.30	50.42	-	49.02
	Falcon-40B	52.35	-	50.36	49.67	-	49.38	50.27	-	50.22
	Mistral-7B	49.90	-	48.94	50.23	-	50.06	51.75	-	51.75
LLM ***(Fine-tuned)***	Meta-LLaMa-2-7B	60.10	59.90	59.00	63.51	66.44	62.55	66.06	70.38	65.12
	Meta-LLaMa-2-13B	60.67	60.64	60.00	64.61	67.67	63.59	68.22	72.19	66.37
	Meta-LLaMa-3-8B	60.06	60.54	59.58	65.76	67.88	65.72	69.83	74.59	68.74
	Gemma-1.1-7B	61.23	61.25	60.28	69.24	70.76	69.00	73.30	76.91	69.18
	Mistral-7B	60.35	60.77	60.07	66.91	70.06	65.95	71.87	75.47	68.53
LLM ***(Multi-task)***	Meta-LLaMa-2-7B	60.70	59.88	59.17	66.15	64.67	64.34	70.40	70.89	70.20
	Meta-LLaMa-2-13B	61.36	61.42	60.69	67.07	66.44	65.68	70.44	69.15	68.62
	Meta-LLaMa-3-8B	61.38	61.85	61.02	67.20	67.13	66.60	71.64	72.06	71.12
	Gemma-1.1-7B	61.54	62.36	61.15	67.71	67.60	66.98	73.12	72.82	71.89
	Mistral-7B	61.03	61.16	60.38	67.69	67.20	66.16	72.34	72.52	71.78

Table 8.6: Evaluation results (%) of LLMs fine-tuned on MARS under the multi-task setting.

Backbone	Training Data	Event			Inference			Transition		
		Acc	AUC	Ma-F1	Acc	AUC	Ma-F1	Acc	AUC	Ma-F1
LLaMa-2 ***7B***	20%	58.03	58.24	57.62	62.43	64.47	60.43	63.11	63.08	62.73
	40%	58.81	58.40	57.69	64.03	67.48	61.58	66.44	70.04	64.15
	60%	59.09	59.41	58.62	64.75	68.10	62.79	67.00	70.85	64.15
	80%	59.48	60.54	59.82	64.15	68.01	61.53	66.42	70.64	64.92
	100%	60.10	59.90	59.00	63.51	66.44	62.55	66.06	70.38	65.12
LLaMa-2 ***13B***	20%	59.95	59.75	58.57	63.80	66.86	61.80	64.11	68.73	64.08
	40%	59.45	59.18	58.25	65.49	68.98	63.54	68.52	71.61	64.82
	60%	60.19	59.46	58.92	65.90	69.59	64.18	68.24	72.17	65.59
	80%	60.24	60.05	59.43	65.99	69.70	64.27	68.35	72.43	65.97
	100%	60.67	60.64	60.00	64.61	67.67	63.59	68.22	72.19	66.37
LLaMa-3 ***8B***	20%	60.56	59.91	58.99	63.40	66.77	61.06	65.23	70.50	64.60
	40%	60.68	59.98	59.23	62.35	69.00	61.81	69.43	72.72	65.27
	60%	60.74	60.88	60.49	65.90	69.59	61.81	69.00	72.78	65.55
	80%	60.91	61.03	60.29	66.73	69.71	61.72	68.71	73.15	66.43
	100%	60.06	60.54	59.58	65.76	67.88	65.72	69.83	74.59	68.74
Gemma-v1.1 ***7B***	20%	59.07	59.54	59.18	64.70	70.42	62.43	68.41	73.64	67.08
	40%	60.79	59.93	59.72	62.80	70.57	62.26	69.83	73.91	62.18
	60%	59.26	60.31	59.25	67.83	70.22	60.56	70.68	74.56	66.98
	80%	59.31	59.32	58.73	64.03	70.77	63.73	69.66	73.51	67.05
	100%	61.23	61.25	60.28	69.24	70.76	69.00	73.30	76.91	69.18
Mistral-v1.1 ***7B***	20%	60.67	60.27	59.61	65.28	69.22	63.16	68.37	72.85	66.15
	40%	60.53	60.78	60.03	65.92	70.21	63.96	69.79	72.97	69.46
	60%	61.82	61.86	61.07	67.65	70.46	64.09	67.92	73.38	66.76
	80%	59.35	59.55	58.85	68.07	70.43	66.49	69.84	73.63	65.84
	100%	60.35	60.77	60.07	66.91	70.06	65.95	71.87	75.47	68.53
DeBERTa-v3-Large ***635M***	25%	58.11	57.90	57.64	63.28	64.12	64.70	64.51	67.21	66.54
	50%	61.32	59.71	60.91	65.36	67.12	68.09	67.95	68.21	67.97
	100%	64.45	64.16	63.27	69.57	71.15	69.33	72.93	74.00	72.01

Table 8.7: Evaluation results (%) of LLMs fine-tuned on MARS under the few-shot setting. Training data refers to the ratio of sampled training data from the full training sets of MARS.

Task	Data Examples	Label
ME.	The tax offices were devastation *(burnt down)*	P.
ME.	Keith and Vinnie are running *(competition)* against each other in the sheriff's election	P.
ME.	We worked together environment *(in the marina)* for years	M.
ME.	The sun is melting horizon *(over the landscape)* like an orange popsicle	M.
ME.	Mammal *(human)* seek food for their own survival	P.
MI.	If I perception *(felt)* the tension leave me, then I feel more relaxed now	P.
MI.	If they both reached the excellence *(world top 100)* in 2005, then they both worked hard to achieve their goals	P.
MI.	If Parker and Garbajosa were adaptable *(two very versatile players)* who could both defend and attack, then they were actually terrible basketball players.	M.
MI.	If Stevens success *(won)* his first eight games, then Steven is a skilled player.	P.
MI.	If I communication *(have to talk)* to my insurance company, then my insurance company is not responsive and does not provide good customer service.	M.
MT.	If he was respectful *(overpowering and right intrusion)*, then he will apologize for his actions and make amends.	P.
MT.	If the other guests have just been invited to participate in a karaoke session *(join community on the dance floor)*, then the other guests decline the invitation and choose to sit and watch instead.	P.
MT.	If Australia opposed *(supported)* South Vietnam in that time period, then Australia support South Vietnam during that time period.	M.
MT.	If Churchill has ignoring *(communication)* to the requests for verification in various ways, then Churchill is not interested in verifying the requests and is avoiding them.	P.
MT.	If Tikal has hundreds *(thousands)* of history structures, then archaeologists have not yet discovered the true purpose of Tikal's structures.	M.

Table 8.8: Case studies of three tasks in the MARS benchmark. ME, MI, and MT refer to three tasks in metaphysical reasoning, respectively. P. refers to plausible in reality and M. refers to metaphysical. The original component before the change/transition is marked in *(grey)*.

CHAPTER 9

CONCLUSIONS

In conclusion, this thesis investigated how language models can reason in a way that generalizes beyond memorized surface forms. Our central thesis is that generalization requires abstractions that transfer across situations, grounding that keeps these abstractions faithful to context, and robustness that preserves coherence when the underlying situation changes. We develop this argument through a single mechanism that recurs throughout the dissertation: the *lift-and-ground* loop, in which conceptualization lifts instance-level entities and events into reusable concepts, and instantiation grounds those concepts back into new contexts to support inference.

We begin by establishing this mechanism as a precise vocabulary and scope. Chapter 3 formalizes conceptualization and instantiation and distinguishes four semantic levels, then centers entity- and event-level conceptualization as the main interface between language and reasoning. This provides the connective tissue for everything that follows: it clarifies what it means to abstract, what it means to ground, and what kinds of generalization we expect to obtain.

With the lens fixed, the thesis then asks how the loop can be made operational under realistic constraints. CAT addresses the learning problem: when supervision is scarce and naive abstraction breaks under relational context, conceptualization cannot be treated as a standalone rewriting step. By coupling conceptualization and instantiation into a semi-supervised cycle with contextual verification, CAT provides a practical way to acquire *context-compatible* abstract commonsense knowledge from largely unlabeled CSKB data. CAR then moves from acquisition to application: it shows how conceptual structure can be used to construct cleaner and more transferable supervision for downstream reasoning, expanding coverage while reducing false-negative distractors in CSKB-based QA synthesis, and thereby improving robustness in zero-shot commonsense QA.

Having shown that the loop can be learned and can improve downstream generalization, the next step is scalability. CANDLE turns the lift-and-ground loop from a one-off augmentation into a scalable knowledge acquisition primitive by distilling multi-step conceptualization and instantiation chains from strong LLM teachers, filtering generations with critic models, and feeding accepted instantiations back for iterative expansion. This stage connects the thesis mechanism to the LLM era directly: the loop is no longer only a modeling idea, but a practical recipe

for producing large pools of grounded and abstract commonsense knowledge that can be used for training and evaluation at scale.

The thesis then extends the same mechanism to model maintenance. As LLMs grow, correcting commonsense failures by retraining becomes increasingly expensive, and commonsense errors rarely correspond to isolated facts; they often reflect concept-level gaps that surface across paraphrases and contexts. CONKE uses conceptualization and instantiation to lift an error into a concept-structured neighborhood and ground it into diverse, context-consistent instances before editing, turning editing from a local patch into a more transferable update. This continues the thesis flow from learning and scaling knowledge to maintaining it efficiently while preserving generalization.

Finally, the thesis closes by confronting the boundary condition that motivates the entire progression: in the LLM era, robust intelligence requires not only static generalization but adaptive reasoning under change. We therefore formalize *reasoning with changes in distribution* as a three-step process over feasibility, consequence, and transition, and introduce metaphysical reasoning together with the MARS benchmark to evaluate it at scale under hierarchical, structured perturbations. The empirical difficulty observed across many models clarifies why the earlier chapters matter and where the remaining gaps lie: conceptualization organizes reusable structure, grounding keeps it faithful to context, but coherent reasoning under change remains a stringent test that exposes limitations in how today's models track plausibility and transitions when assumptions drift.

Overall, the thesis advances one integrated message: conceptualization enables transfer, instantiation enforces grounding, and the lift-and-ground loop is a reusable primitive that can be learned under weak supervision (CAT), translated into more robust downstream supervision (CAR), scaled through LLM distillation into iterative acquisition (CANDLE), and used to generalize updates during post-training adaptation (CONKE), culminating in a benchmarked evaluation of reasoning under structured distribution shifts (MARS).

9.1 Strengths and Limitations of Conceptualization

9.1.1 Strengths

Conceptualization remains valuable in the LLM era because it targets a form of generalization that scale alone does not guarantee: invariance under *structured variation* [295]. When surface forms shift through paraphrases, entity swaps, event rewrites, or compositional changes in in-

structions, concept-level structure provides a stable interface for transferring expectations and retrieving the "right kind" of knowledge, rather than relying on brittle lexical overlap [296]. This matters most in precisely the regimes that define real-world robustness: novel contexts, new combinations of familiar parts, and distribution shifts that preserve meaning while altering form.

Conceptualization also improves the quality of supervision and evaluation pipelines. Once instances are grouped by shared concepts, it becomes possible to construct semantically controlled contrast sets, such as harder yet fairer distractors for multiple-choice reasoning or more faithful augmentations for commonsense inference [297]. This reduces spurious training artifacts (e.g., accidental lexical cues or false negatives) and makes the learning signal better aligned with the intended notion of plausibility and inference. In this sense, conceptualization is not only a representational tool, but also a mechanism for shaping data so that models are trained and tested on the right invariances [269].

A further strength is that conceptualization provides *handles* for control and maintenance in post-training. Modern alignment pipelines, including instruction tuning, preference optimization, and RLHF-style objectives, primarily operate at the response level, optimizing helpfulness or preference under coarse feedback [298]. Concept-level representations offer a complementary axis for enforcing consistency across paraphrases and relational realizations, and for localizing failures to reusable abstractions rather than isolated strings. This makes conceptualization useful not only for acquiring new knowledge, but also for auditing and stabilizing behavior: it turns errors into structured neighborhoods in which verification, edits, and evaluations can generalize beyond a single surface form.

9.1.2 Limitations

Despite these benefits, conceptualization faces practical limitations in today's LLM training and deployment pipelines.

First, its gains can become less visible under massive pre-training. Contemporary foundation models are trained on extremely large and diverse corpora, and scaling trends can make models appear to "already know" many broad regularities, especially on in-distribution benchmarks [299]. In such settings, explicit conceptualization may yield smaller headline improvements, not because it is unhelpful, but because standard evaluations under-measure structured transfer and over-reward pattern matching. This can create the impression that conceptualization is redundant when, in fact, its value is concentrated in shift-heavy regimes that are not always

reflected by common leaderboards.

Second, concepts are inherently context-dependent and therefore hard to "pin down" as fixed labels. The same event may admit multiple plausible abstractions, and the usefulness of an abstraction depends on downstream relations, goals, and the intended inference. Naive abstraction can erase precisely the property that supports the original implication, producing concepts that are linguistically plausible but inferentially invalid once re-grounded. This is a structural limitation: conceptualization is not a purely local rewriting operation, but a context-sensitive choice over what information can be compressed without breaking downstream reasoning.

Third, abstraction can encourage over-generalization. Concept-level statements are easy to express fluently and confidently, which increases the risk of producing broad regularities that sound universally valid but fail in grounded situations, especially when the model is optimized for helpful explanations. The tension is that conceptualization is designed to compress away details, yet many commonsense inferences hinge on those very details; without strong grounding pressure, abstractions may drift into plausible-sounding but unreliable generalities.

Fourth, integration with RL-based LLM training remains under-developed. RLHF-style post-training typically optimizes preferences over full responses with reward signals that are coarse, sparse, and dominated by surface-level qualities such as style, verbosity, and perceived helpfulness [298]. In this training regime, it is nontrivial to introduce concept-level objectives, especially when the RL stage is brief compared to pre-training and constrained by stability considerations. Moreover, as you noted, the very premise of pre-training at scale to cover as much knowledge as possible can weaken the perceived role of conceptualization during RL, because the optimization focuses on behavior shaping rather than building structured abstractions.

Finally, evaluation is easy to get wrong. Many "OOD" settings are either too mild, poorly controlled, or confounded by artifacts, which can understate (or overstate) the value of conceptual structure. If shifts are not explicitly structured (what changes, how it changes, and what invariances should hold), it is difficult to isolate whether conceptualization truly improves systematic transfer or merely shifts performance via unrelated factors such as generation length, calibration, or dataset biases [300].

9.2 Future Works

9.2.1 Towards Semantic-Space Analysis of Conceptualization

A natural future direction emerging from this thesis is to study conceptualization not only through downstream performance, but also through the geometry of the learned semantic space. The current thesis evaluates conceptualization primarily by plausibility, transfer utility, editing behavior, and robustness under distributional change. These are the most task-relevant criteria, but they do not fully reveal *how* conceptual structure is reorganized inside a model before and after conceptualization-oriented training or editing. A complementary line of analysis would therefore ask whether conceptually related events become more geometrically aligned in representation space, whether irrelevant surface variation is compressed, and whether the resulting space better separates inferentially distinct abstractions.

Such an analysis could be conducted at multiple levels. At the instance-to-concept level, one may measure whether concrete events that share a valid conceptualization form tighter and more semantically coherent neighborhoods after training. At the concept-to-instance level, one may examine whether a concept embedding or hidden-state representation remains stably connected to diverse but context-compatible instantiations. Across models, one may compare whether improvements in robustness correspond to more organized concept manifolds, better clustering purity, or more reliable local neighborhoods around abstract events. For editing methods such as CONKE, semantic-space analysis may also help distinguish genuine conceptual revision from superficial output calibration by checking whether the representation of an edited concept shifts coherently together with its related instances.

Importantly, this direction is not only diagnostic. A better understanding of the semantic geometry of conceptualization may eventually feed back into model design. For example, it may motivate regularization objectives that encourage concept-level compactness while preserving context-sensitive separability, or it may suggest retrieval strategies that use semantic neighborhoods to select more informative conceptualizations and instantiations. In this sense, semantic-space analysis offers a promising bridge between the functional view of conceptualization adopted in this thesis and a more mechanistic understanding of how language models internally organize reusable knowledge.

9.2.2 Potential Applications in AI for Science

Although this thesis focuses on commonsense reasoning, the conceptualization-centered perspective may also be valuable for AI for Science. Many scientific reasoning problems exhibit a similar structural tension to the one studied here: observations are concrete, noisy, and context-dependent, while effective generalization requires lifting them into abstractions that can be reused across conditions, systems, or experimental settings. In this respect, conceptualization can be viewed as an interface between local observations and reusable scientific regularities, much as it serves in this thesis as an interface between concrete events and transferable commonsense knowledge.

One promising application is scientific knowledge organization and hypothesis transfer. Scientific texts and experimental reports often describe highly specific conditions, materials, procedures, or outcomes. A conceptualization layer could help map these surface-level descriptions into more reusable patterns, allowing models to connect related findings that are phrased differently or instantiated in different domains. For example, a system may abstract over specific experimental interventions to identify broader causal or functional motifs, and then instantiate those motifs in a new setting to propose plausible hypotheses, missing controls, or likely failure points. In this role, conceptualization could support literature synthesis, scientific question answering, and the reuse of procedural knowledge across experiments.

At the same time, scientific domains also expose the limits of the current thesis framework and point to important extensions. Compared with everyday commonsense, AI for Science typically requires stricter adherence to quantitative constraints, domain ontologies, mechanistic consistency, and uncertainty estimation. A useful scientific conceptualization framework would therefore need to combine the flexibility of neural abstraction with stronger external grounding, such as symbolic constraints, simulators, or expert-defined taxonomies. Extending the lift-and-ground loop developed in this thesis to scientific settings is thus an attractive research direction: it could preserve the benefits of reusable concept-level structure while enforcing the tighter correctness requirements that scientific reasoning demands.

9.2.3 Other Possible Directions

A natural direction is to make the benefits of conceptualization more visible under modern scaling by aligning training and evaluation with the kind of transfer conceptualization is meant to support [301]. One promising line is to adopt invariance-focused protocols, such as explicit perturbation families, compositional splits, and controlled distribution shifts, and to report stability-

style metrics (consistency under meaning-preserving changes, calibrated plausibility under shift) alongside accuracy. This reframes conceptualization as a tool for systematic robustness rather than a generic accuracy booster.

To address context-dependence, future work should treat conceptualization as a contextual operator whose output is selected for downstream usefulness rather than surface plausibility. A practical approach is to couple conceptualization with verification and re-grounding as a first-class constraint: abstractions should be tested by whether they preserve (or correctly transform) relational implications when instantiated back into diverse contexts. More broadly, concept selection could be modeled as a decision problem conditioned on relation type, goal, and expected inference, rather than as unconditional generation.

To mitigate over-generalization, future work can strengthen grounding pressure by requiring abstract claims to "pay rent" in concrete instances. For example, a concept-level regularity can be treated as valid only if it supports successful instantiation into multiple diverse, context-consistent cases, or if it can be linked to retrieved evidence/examples that survive adversarial paraphrases [302]. This turns abstraction into a hypothesis that must generalize in grounded tests, reducing the space of fluent but unreliable generalities.

For the RL-based limitation, a key opportunity is to design post-training objectives that explicitly reward concept-level invariances. One concrete direction is to incorporate preference pairs or reward-model training data that differ in surface form but share the same underlying concept, so that the policy is directly optimized for concept-consistent implications across perturbations [303]. Another is to add auxiliary losses during RLHF/DPO-style training that enforce consistency across concept-preserving rewrites or across concept neighborhoods derived from conceptualization and instantiation cycles. Conceptualization can then serve as a generator of structured training pairs for RL, rather than competing with pre-training coverage.

Finally, evaluation should continue moving toward benchmarks that explicitly represent change and test coherence through transitions. Future work can expand beyond discriminative plausibility judgments to interactive or planning-like settings where models must maintain feasibility, predict consequences, and repair trajectories when a change breaks plausibility [304]. In such settings, conceptualization can function as the coordinate system of change: it defines what varies, what remains invariant, and what kinds of repairs preserve meaning while restoring plausibility.

Broader limitation and outlook. Taken together, the chapters of this thesis suggest that conceptualization is a powerful but not all-sufficient organizing principle for generalizable reason-

ing. It helps expose reusable structure, improves transfer across related situations, and provides a more principled unit for acquisition, reasoning, and editing than isolated surface forms alone. At the same time, the results also indicate that abstraction must be carefully controlled by context, that concept-level editing does not automatically guarantee full global consistency, and that reasoning under abstract or distribution-shifted conditions still benefits from stronger feasibility checks and external grounding. In this sense, the main lesson of the thesis is not that reasoning can be reduced to abstraction, but that robust reasoning requires a disciplined interaction between abstraction and grounding. The most promising next step is therefore to combine conceptualization with stronger mechanisms for consistency control, semantic-space diagnosis, and domain-aware feasibility verification, so that the reusable structure captured by concepts can be translated into reasoning that is not only more general, but also more reliable.

REFERENCES

[1] Yiheng Shu et al., "Distribution shifts are bottlenecks: Extensive evaluation for grounding language models to knowledge bases," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024: Student Research Workshop, St. Julian's, Malta, March 21-22, 2024*, Neele Falk et al., Eds., Association for Computational Linguistics, 2024, pp. 71–88. DOI: 10.18653/V1/2024.EACL-SRW.7 (cit. on p. 1).

[2] Myeongjun Jang et al., "BECEL: benchmark for consistency evaluation of language models," in *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, Nicoletta Calzolari et al., Eds., International Committee on Computational Linguistics, 2022, pp. 3680–3696 (cit. on p. 1).

[3] Guozheng Li et al., "On the consistency of commonsense in large language models," in *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Wanxiang Che et al., Eds., Association for Computational Linguistics, 2025, pp. 16 205–16 225 (cit. on p. 1).

[4] Xingxuan Zhang et al., "Understanding the generalization of in-context learning in transformers: An empirical study," in *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*, OpenReview.net, 2025 (cit. on p. 1).

[5] Erik T Mueller, *Commonsense reasoning: an event calculus based approach*. Morgan Kaufmann, 2014 (cit. on pp. 2, 66).

[6] Shuai Yang et al., "Word-level commonsense knowledge selection for event detection," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, Nicoletta Calzolari et al., Eds., ELRA and ICCL, 2024, pp. 17 675–17 682 (cit. on p. 2).

[7] Ernest Davis, *Representations of commonsense knowledge*. Morgan Kaufmann, 2014 (cit. on pp. 2, 66).

[8] Xiang Li et al., "Commonsense knowledge base completion," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*, The Association for Computer Linguistics, 2016. DOI: 10.18653/v1/p16-1137 (cit. on pp. 3, 45).

[9] Shreshth Tuli et al., "TANGO: commonsense generalization in predicting tool interactions for mobile manipulators," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, Zhi-Hua Zhou, Ed., ijcai.org, 2021, pp. 4197–4205. DOI: 10.24963/IJCAI.2021/577 (cit. on p. 3).

[10] Alexandre Drouin et al., "Workarena: How capable are web agents at solving common knowledge work tasks?" In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, OpenReview.net, 2024 (cit. on p. 3).

[11] Tianqing Fang et al., "Ckbp v2: An expert-annotated evaluation set for commonsense knowledge base population," *CoRR*, vol. abs/2304.10392, 2023. DOI: 10.48550/arXiv.2304.10392 (cit. on pp. 3, 88).

[12] Chaitanya Malaviya et al., "Commonsense knowledge base completion with structural and semantic context," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, AAAI Press, 2020, pp. 2925–2933. DOI: 10.1609/AAAI.V34I03.5684 (cit. on pp. 3, 33, 45, 48).

[13] Bibo Cai et al., "Mitigating reporting bias in semi-supervised temporal commonsense inference with probabilistic soft logic," in *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, AAAI Press, 2022, pp. 10 454–10 462. DOI: 10.1609/AAAI.V36I10.21288 (cit. on p. 6).

[14] Xiaoyuan Li et al., "Hellaswag-pro: A large-scale bilingual benchmark for evaluating the robustness of llms in commonsense reasoning," in *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Wanxiang Che et al., Eds., Association for Computational Linguistics, 2025, pp. 9038–9072 (cit. on p. 7).

[15] Ke Shen, "The generalization and robustness of transformer-based language models on commonsense reasoning," in *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, Michael J. Wooldridge et al., Eds., AAAI Press, 2024, pp. 23 419–23 420. DOI: 10.1609/AAAI.V38I21.30410 (cit. on p. 7).

[16] Qizhou Chen et al., "Attribution analysis meets model editing: Advancing knowledge correction in vision language models with visedit," in *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, Toby Walsh et al., Eds., AAAI Press, 2025, pp. 2168–2176. DOI: 10.1609/AAAI.V39I2.32215 (cit. on p. 8).

[17] Chenhui Hu et al., "Knowledge in superposition: Unveiling the failures of lifelong knowledge editing for large language models," in *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, Toby Walsh et al., Eds., AAAI Press, 2025, pp. 24 086–24 094. DOI: 10.1609/AAAI.V39I22.34583 (cit. on p. 8).

[18] Yifan Lu et al., "Knowledge editing with dynamic knowledge graphs for multi-hop question answering," in *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, Toby Walsh et al., Eds., AAAI Press, 2025, pp. 24 741–24 749. DOI: 10.1609/AAAI.V39I23.34655 (cit. on p. 8).

[19] Xiaohang Tang et al., "A word sense distribution-based approach for semantic change prediction," in *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, Houda Bouamor et al., Eds., Association for Computational Linguistics, 2023, pp. 3575–3590. DOI: 10.18653/V1/2023.FINDINGS-EMNLP.231 (cit. on p. 8).

[20] Maarten Sap et al., "Commonsense reasoning for natural language processing," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, ACL 2020, Online, July 5, 2020*, Agata Savary et al., Eds., Association for Computational Linguistics, 2020, pp. 27–33. DOI: 10.18653/v1/2020.acl-tutorials.7 (cit. on pp. 12, 29).

[21] Alon Talmor et al., "Commonsenseqa: A question answering challenge targeting commonsense knowledge," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein et al., Eds., Association for Computational Linguistics, 2019, pp. 4149–4158. DOI: 10.18653/v1/n19-1421 (cit. on pp. 12, 13, 29, 47, 49, 81, 94).

[22] Kazumasa Omura et al., "A method for building a commonsense inference dataset based on basic events," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber et al., Eds., Association for Computational Linguistics, 2020, pp. 2450–2460. DOI: 10.18653/v1/2020.emnlp-main.192 (cit. on pp. 12, 29).

[23] Edoardo Maria Ponti et al., "XCOPA: A multilingual dataset for causal commonsense reasoning," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber et al., Eds., Association for Computational Linguistics, 2020, pp. 2362–2376. DOI: 10.18653/v1/2020.emnlp-main.185 (cit. on pp. 12, 29).

[24] Tianqing Fang et al., "Benchmarking commonsense knowledge base population with an effective evaluation dataset," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, Marie-Francine Moens et al., Eds., Association for Computational Linguistics, 2021, pp. 8949–8964. DOI: 10.18653/v1/2021.emnlp-main.705 (cit. on pp. 12, 29, 67, 88).

[25] Trieu H. Trinh et al., "A simple method for commonsense reasoning," *CoRR*, vol. abs/1806.02847, 2018 (cit. on pp. 12, 47).

[26] Xiang Lorraine Li et al., "A systematic investigation of commonsense knowledge in large language models," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Yoav Goldberg et al., Eds., Association for Computational Linguistics, 2022, pp. 11 838–11 855 (cit. on pp. 12, 47).

[27] Vered Shwartz et al., "Unsupervised commonsense question answering with self-talk," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber et al., Eds., Association for Computational Linguistics, 2020, pp. 4615–4629. DOI: 10.18653/v1/2020.emnlp-main.373 (cit. on pp. 12, 33, 44, 47, 52, 64, 81, 85).

[28] Zi-Yi Dou et al., "Zero-shot commonsense question answering with cloze translation and consistency optimization," in *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, AAAI Press, 2022, pp. 10 572–10 580 (cit. on pp. 12, 47).

[29] Antoine Bosselut et al., "Dynamic neuro-symbolic knowledge graph construction for zero-shot commonsense question answering," in *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, AAAI Press, 2021, pp. 4923–4931 (cit. on pp. 12, 47, 52, 64, 81, 85).

[30] Zhenzhong Lan et al., "ALBERT: A lite BERT for self-supervised learning of language representations," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, OpenReview.net, 2020 (cit. on p. 12).

[31] Antoine Bosselut et al., "COMET: commonsense transformers for automatic knowledge graph construction," in *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen et al., Eds., Association for Computational Linguistics, 2019, pp. 4762–4779. DOI: 10.18653/v1/p19-1470 (cit. on pp. 12, 13, 23, 25, 29, 33, 35, 36, 47, 76, 79).

[32] Tom B. Brown et al., "Language models are few-shot learners," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle et al., Eds., 2020 (cit. on pp. 12, 16, 53, 57, 70, 81, 106).

[33] Jiawei Wang et al., "Art: All-round thinker for unsupervised commonsense question answering," in *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, Nicoletta Calzolari et al., Eds., International Committee on Computational Linguistics, 2022, pp. 1490–1501 (cit. on p. 12).

[34] Pratyay Banerjee et al., "Self-supervised knowledge triplet learning for zero-shot question answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber et al., Eds., Association for Computational Linguistics, 2020, pp. 151–162. DOI: 10.18653/v1/2020.emnlp-main.11 (cit. on pp. 13, 47, 52, 64, 81, 85).

[35] Kaixin Ma et al., "Knowledge-driven data construction for zero-shot evaluation in commonsense question answering," in *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, AAAI Press, 2021, pp. 13 507–13 515 (cit. on pp. 13, 44, 47, 48, 53, 55–57, 59–62, 64, 76, 81, 85, 94, 110).

[36] Ying Su et al., "MICO: A multi-alternative contrastive learning framework for commonsense knowledge representation," in *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Yoav Goldberg et al., Eds., Association for Computational Linguistics, 2022, pp. 1339–1351 (cit. on pp. 13, 23, 47, 52, 64, 81, 85).

[37] Yu Jin Kim et al., "Modularized transfer learning with multiple knowledge graphs for zero-shot commonsense reasoning," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, Marine Carpuat et al., Eds., Association for Computational Linguistics, 2022, pp. 2244–2257. DOI: 10.18653/v1/2022.naacl-main.163 (cit. on pp. 13, 47, 49, 53, 62, 64, 81, 85).

[38] Haochen Shi et al., "QADYNAMICS: training dynamics-driven synthetic QA diagnostic for zero-shot commonsense question answering," in *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, Houda Bouamor et al., Eds., Association for Computational Linguistics, 2023, pp. 15 329–15 341 (cit. on pp. 13, 47, 94).

[39] Xin Guan et al., "Multi-hop commonsense knowledge injection framework for zero-shot commonsense question answering," *Expert Syst. Appl.*, vol. 298, p. 129 806, 2026. DOI: 10.1016/J.ESWA.2025.129806 (cit. on pp. 13, 47, 64, 85).

[40] OpenAI, "Chatgpt: Optimizing language models for dialogue," *OpenAI*, 2022 (cit. on pp. 13, 16, 17, 53, 61, 68, 69, 79, 103, 106).

[41] OpenAI, "GPT-4 technical report," *CoRR*, vol. abs/2303.08774, 2023. DOI: 10.48550/arXiv.2303.08774 (cit. on pp. 13, 16, 17, 69, 81, 103, 111).

[42] Peter West et al., "Symbolic knowledge distillation: From general language models to commonsense models," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, Marine Carpuat et al., Eds., Association for Computational Linguistics, 2022, pp. 4602–4625. DOI: 10.18653/v1/2022.naacl-main.341 (cit. on pp. 13, 45, 47, 53, 55, 57, 63, 64, 69–71, 75, 80, 85, 106).

[43] Melanie Sclar et al., "Referee: Reference-free sentence summarization with sharper controllability through symbolic knowledge distillation," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Yoav Goldberg et al., Eds., Association for Computational Linguistics, 2022, pp. 9649–9668. DOI: 10.18653/V1/2022.EMNLP-MAIN.655 (cit. on pp. 13, 69).

[44] Chandra Bhagavatula et al., "I2D2: inductive knowledge distillation with neurologic and self-imitation," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, Anna Rogers et al., Eds., Association for Computational Linguistics, 2023, pp. 9614–9630. DOI: 10.18653/V1/2023.ACL-LONG.535 (cit. on pp. 13, 69).

[45] Peter West et al., "Novacomet: Open commonsense foundation models with symbolic knowledge distillation," in *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, Houda Bouamor et al., Eds., Association for Computational Linguistics, 2023, pp. 1127–1149 (cit. on pp. 13, 69, 88).

[46] Xingwei He et al., "Metric-guided distillation: Distilling knowledge from the metric to ranker and retriever for generative commonsense reasoning," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Yoav Goldberg et al., Eds., Association for Computational Linguistics, 2022, pp. 839–852. DOI: 10.18653/V1/2022.EMNLP-MAIN.53 (cit. on pp. 13, 69).

[47] Hyungjoo Chae et al., "Dialogue chain-of-thought distillation for commonsense-aware conversational agents," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, Houda Bouamor et al., Eds., Association for Computational Linguistics, 2023, pp. 5606–5632 (cit. on pp. 13, 69).

[48] Hyunwoo Kim et al., "SODA: million-scale dialogue distillation with social commonsense contextualization," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, Houda Bouamor et al., Eds., Association for Computational Linguistics, 2023, pp. 12 930–12 949 (cit. on pp. 13, 69).

[49] Christine A Montgomery, "Concept extraction," *American journal of computational linguistics*, vol. 8, no. 2, pp. 70–73, 1982 (cit. on p. 14).

[50] Boris Gelfand et al., "Automated concept extraction from plain text," in *AAAI 1998 Workshop on Text Categorization*, 1998, pp. 13–17 (cit. on p. 14).

[51] Shuting Wang et al., "Using prerequisites to extract concept maps fromtextbooks," in *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*, Snehasis Mukhopadhyay et al., Eds., ACM, 2016, pp. 317–326. DOI: 10.1145/2983323.2983725 (cit. on p. 14).

[52] Aditya G. Parameswaran et al., "Towards the web of concepts: Extracting concepts from large datasets," *Proc. VLDB Endow.*, vol. 3, no. 1, pp. 566–577, 2010. DOI: 10.14778/1920841.1920914 (cit. on p. 14).

[53] Dheeraj Rajagopal et al., "A graph-based approach to commonsense concept extraction and semantic similarity detection," in *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013, Companion Volume*, Leslie Carr et al., Eds., International World Wide Web Conferences Steering Committee / ACM, 2013, pp. 565–570. DOI: 10.1145/2487788.2487995 (cit. on p. 14).

[54] Eduard H. Hovy et al., "Toward completeness in concept extraction and classification," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, 6-7 August 2009, Singapore, A meeting of SIGDAT, a Special Interest Group of the ACL*, ACL, 2009, pp. 948–957 (cit. on p. 14).

[55] Adit Krishnan et al., "Unsupervised concept categorization and extraction from scientific document titles," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, Ee-Peng Lim et al., Eds., ACM, 2017, pp. 1339–1348. DOI: 10.1145/3132847.3133023 (cit. on p. 14).

[56] Marius Pasca, "Outclassing wikipedia in open-domain information extraction: Weakly-supervised acquisition of attributes over conceptual hierarchies," in *EACL 2009, 12th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, Athens, Greece, March 30 - April 3, 2009*, Alex Lascarides et al., Eds., The Association for Computer Linguistics, 2009, pp. 639–647 (cit. on p. 14).

[57] George A. Miller, "Wordnet: A lexical database for english," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995. DOI: 10.1145/219717.219748 (cit. on pp. 14, 28, 49, 50, 55, 61, 64, 67, 68, 89).

[58] Robyn Speer et al., "Conceptnet 5.5: An open multilingual graph of general knowledge," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, Satinder Singh et al., Eds., AAAI Press, 2017, pp. 4444–4451 (cit. on pp. 14, 23, 61, 64, 67).

[59] Robyn Speer et al., "Conceptnet 5: A large semantic network for relational knowledge," in *The People's Web Meets NLP, Collaboratively Constructed Language Resources*, ser. Theory and Applications of Natural Language Processing, Iryna Gurevych et al., Eds., Springer, 2013, pp. 161–176. DOI: 10.1007/978-3-642-35085-6_6 (cit. on p. 14).

[60] Robyn Speer et al., "Representing general relational knowledge in conceptnet 5," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, Nicoletta Calzolari et al., Eds., European Language Resources Association (ELRA), 2012, pp. 3679–3686 (cit. on p. 14).

[61] Robyn Speer et al., "Conceptnet 5," *Tiny Trans. Comput. Sci.*, vol. 1, 2012 (cit. on p. 14).

[62] Wentao Wu et al., "Probase: A probabilistic taxonomy for text understanding," in *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2012, Scottsdale, AZ, USA, May 20-24, 2012*, K. Selçuk Candan et al., Eds., ACM, 2012, pp. 481–492. DOI: 10.1145/2213836.2213891 (cit. on pp. 14, 28, 49, 50, 67, 68, 76, 89, 110).

[63] Jiaqing Liang et al., "Probase+: Inferring missing links in conceptual taxonomies," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 6, pp. 1281–1295, 2017. DOI: 10.1109/TKDE.2017.2653115 (cit. on p. 14).

[64] Sören Auer et al., "Dbpedia: A nucleus for a web of open data," in *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007*, Karl Aberer et al., Eds., ser. Lecture Notes in Computer Science, vol. 4825, Springer, 2007, pp. 722–735. DOI: 10.1007/978-3-540-76298-0_52 (cit. on p. 14).

[65] Christian Bizer et al., "Dbpedia - A crystallization point for the web of data," *J. Web Semant.*, vol. 7, no. 3, pp. 154–165, 2009. DOI: 10.1016/J.WEBSEM.2009.07.002 (cit. on p. 14).

[66] Shuang Liu et al., "An effective approach to document retrieval via utilizing wordnet and recognizing phrases," in *SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, July 25-29, 2004*, Mark Sanderson et al., Eds., ACM, 2004, pp. 266–272. DOI: 10.1145/1008992.1009039 (cit. on p. 14).

[67] Apostol Natsev et al., "Semantic concept-based query expansion and re-ranking for multimedia retrieval," in *Proceedings of the 15th International Conference on Multimedia 2007, Augsburg, Germany, September 24-29, 2007*, Rainer Lienhart et al., Eds., ACM, 2007, pp. 991–1000. DOI: 10.1145/1291233.1291448 (cit. on p. 14).

[68] Yangqiu Song et al., "Short text conceptualization using a probabilistic knowledgebase," in *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*, Toby Walsh, Ed., IJCAI/AAAI, 2011, pp. 2330–2336. DOI: 10.5591/978-1-57735-516-8/IJCAI11-388 (cit. on pp. 14, 24, 28, 47, 67, 68, 89).

[69] Yangqiu Song et al., "Open domain short text conceptualization: A generative + descriptive modeling approach," in *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, Qiang Yang et al., Eds., AAAI Press, 2015, pp. 3820–3826 (cit. on pp. 14, 24, 28, 47, 67, 68, 89).

[70] Bevan Koopman et al., "An evaluation of corpus-driven measures of medical concept similarity for information retrieval," in *21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29 - November 02, 2012*, Xue-wen Chen et al., Eds., ACM, 2012, pp. 2439–2442. DOI: 10.1145/2396761.2398661 (cit. on p. 14).

[71] Jiaping Zheng et al., "Key concept identification for medical information retrieval," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, Lluís Màrquez et al., Eds., The Association for Computational Linguistics, 2015, pp. 579–584. DOI: 10.18653/V1/D15-1069 (cit. on p. 14).

[72] Lihan Chen et al., "Short text entity linking with fine-grained topics," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*, Alfredo Cuzzocrea et al., Eds., ACM, 2018, pp. 457–466. DOI: 10.1145/3269206.3271809 (cit. on p. 14).

[73] Wen Hua et al., "Short text understanding through lexical-semantic analysis," in *31st IEEE International Conference on Data Engineering, ICDE 2015, Seoul, South Korea, April 13-17, 2015*, Johannes Gehrke et al., Eds., IEEE Computer Society, 2015, pp. 495–506. DOI: 10.1109/ICDE.2015.7113309 (cit. on p. 14).

[74] Jacob Devlin et al., "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein et al., Eds., Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/v1/n19-1423 (cit. on pp. 15, 33, 44, 55).

[75] Yinhan Liu et al., "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019 (cit. on pp. 15, 33, 52, 64, 78, 85).

[76] Pengcheng He et al., "Deberta: Decoding-enhanced bert with disentangled attention," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, OpenReview.net, 2021 (cit. on pp. 15, 33).

[77] Pengcheng He et al., "Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing," in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, OpenReview.net, 2023 (cit. on pp. 15, 33, 52, 64, 78, 85, 110).

[78] Mutian He et al., "Acquiring and modeling abstract commonsense knowledge via conceptualization," *Artif. Intell.*, vol. 333, p. 104 149, 2024. DOI: 10.1016/J.ARTINT.2024.104149 (cit. on pp. 15, 20, 21, 23–25, 27–29, 34, 35, 45, 47, 48, 67–69, 74, 75, 77, 85, 89, 90, 107, 163).

[79] Luyao Huang et al., "Glossbert: BERT for word sense disambiguation with gloss knowledge," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui et al., Eds., Association for Computational Linguistics, 2019, pp. 3507–3512. DOI: 10.18653/v1/D19-1355 (cit. on p. 15).

[80] Maarten Sap et al., "Social iqa: Commonsense reasoning about social interactions," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui et al., Eds., Association for Computational Linguistics, 2019, pp. 4462–4472. DOI: 10.18653/v1/D19-1454 (cit. on pp. 15, 33, 44, 48, 49, 81, 94).

[81] Zhaowei Wang et al., "AbsPyramid: Benchmarking the abstraction ability of language models with a unified entailment graph," in *Findings of the Association for Computational Linguistics: NAACL 2024*, Kevin Duh et al., Eds., Mexico City, Mexico: Association for Computational Linguistics, 2024, pp. 3991–4010 (cit. on pp. 15, 16, 20, 21, 68, 99, 107).

[82] Hongming Zhang et al., "ASER: A large-scale eventuality knowledge graph," in *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, Yennun Huang et al., Eds., ACM / IW3C2, 2020, pp. 201–211. DOI: 10.1145/3366423.3380107 (cit. on p. 15).

[83] Hongming Zhang et al., "ASER: towards large-scale commonsense knowledge acquisition via higher-order selectional preference over eventualities," *Artif. Intell.*, vol. 309, p. 103 740, 2022. DOI: 10.1016/j.artint.2022.103740 (cit. on pp. 15, 68).

[84] Mengying Lu et al., "Distantly supervised course concept extraction in moocs with academic discipline," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, Anna Rogers et al., Eds., Association for Computational Linguistics, 2023, pp. 13 044–13 059. DOI: 10.18653/V1/2023.ACL-LONG.729 (cit. on p. 15).

[85] Weiqi Wang et al., "CAT: A contextualized conceptualization and instantiation framework for commonsense reasoning," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, Anna Rogers et al., Eds., Association for Computational Linguistics, 2023, pp. 13 111–13 140. DOI: 10.18653/V1/2023.ACL-LONG.733 (cit. on pp. 15, 17, 47, 50, 53, 67–69, 71, 74, 76–78, 88, 89, 110, 163).

[86] Silin Gao et al., "Comfact: A benchmark for linking contextual commonsense knowledge," in *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Yoav Goldberg et al., Eds., Association for Computational Linguistics, 2022, pp. 1656–1675. DOI: 10.18653/V1/2022.FINDINGS-EMNLP.120 (cit. on p. 15).

[87] Maria Becker et al., "COCO-EX: A tool for linking concepts from texts to conceptnet," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, EACL 2021, Online, April 19-23, 2021*, Dimitra Gkatzia et al., Eds., Association for Computational Linguistics, 2021, pp. 119–126. DOI: 10.18653/V1/2021.EACL-DEMOS.15 (cit. on p. 15).

[88] Hao Peng et al., "COPEN: probing conceptual knowledge in pre-trained language models," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Yoav Goldberg et al., Eds., Association for Computational Linguistics, 2022, pp. 5015–5035. DOI: 10.18653/V1/2022.EMNLP-MAIN.335 (cit. on pp. 15, 20, 24, 28, 47, 67, 68, 89).

[89] Siyu Yuan et al., "Causality-aware concept extraction based on knowledge-guided prompting," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, Anna Rogers et al., Eds., Association for Computational Linguistics, 2023, pp. 9255–9272. DOI: 10.18653/V1/2023.ACL-LONG.514 (cit. on p. 15).

[90] Zhaowei Wang et al., "Absinstruct: Eliciting abstraction ability from llms through explanation tuning with plausibility estimation," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, Lun-Wei Ku et al., Eds., Association for Computational Linguistics, 2024, pp. 973–994. DOI: 10.18653/V1/2024.ACL-LONG.55 (cit. on p. 15).

[91] Alec Radford et al., "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019 (cit. on pp. 15, 33, 61, 64, 78, 79, 85, 110).

[92] Mike Lewis et al., "BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky et al., Eds., Association for Computational Linguistics, 2020, pp. 7871–7880. DOI: 10.18653/v1/2020.acl-main.703 (cit. on pp. 15, 33).

[93] Colin Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, 140:1–140:67, 2020 (cit. on pp. 15, 81).

[94] Jiayu Liu et al., "Revisiting epistemic markers in confidence estimation: Can markers accurately reflect large language models' uncertainty?" In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Wanxiang Che et al., Eds., Vienna, Austria: Association for Computational Linguistics, 2025, pp. 206–221. DOI: 10.18653/v1/2025.acl-short.18 (cit. on p. 15).

[95] Machel Reid et al., "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," *CoRR*, vol. abs/2403.05530, 2024. DOI: 10.48550/ARXIV.2403.05530 (cit. on pp. 16, 17, 103).

[96] Hugo Touvron et al., "Llama: Open and efficient foundation language models," *CoRR*, vol. abs/2302.13971, 2023. DOI: 10.48550/arXiv.2302.13971 (cit. on pp. 16, 17, 103, 111).

[97] Hugo Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," *CoRR*, vol. abs/2307.09288, 2023. DOI: 10.48550/ARXIV.2307.09288 (cit. on pp. 16, 17, 68, 78, 79, 85, 103, 111).

[98] Weiqi Wang et al., "CANDLE: iterative conceptualization and instantiation distillation from large language models for commonsense reasoning," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, Association for Computational Linguistics, 2024 (cit. on pp. 16, 17, 20, 21, 89–91, 106, 110, 112).

[99] Huaixiu Steven Zheng et al., "Take a step back: Evoking reasoning via abstraction in large language models," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, OpenReview.net, 2024 (cit. on p. 16).

[100] Ruochen Zhao et al., "Explaining language model predictions with high-impact concepts," in *Findings of the Association for Computational Linguistics: EACL 2024, St. Julian's, Malta, March 17-22, 2024*, Yvette Graham et al., Eds., Association for Computational Linguistics, 2024, pp. 995–1012 (cit. on p. 16).

[101] Tianqing Fang et al., "Complex reasoning over logical queries on commonsense knowledge graphs," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, Lun-Wei Ku et al., Eds., Association for Computational Linguistics, 2024, pp. 11 365–11 384. DOI: 10.18653/V1/2024.ACL-LONG.613 (cit. on pp. 16, 84).

[102] Jiacheng Liu et al., "Vera: A general-purpose plausibility estimation model for commonsense statements," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, Houda Bouamor et al., Eds., Association for Computational Linguistics, 2023, pp. 1264–1287. DOI: 10.18653/V1/2023.EMNLP-MAIN.81 (cit. on pp. 16, 71, 74, 78, 85, 88, 90, 110).

[103] Gregory Murphy, *The big book of concepts*. MIT press, 2004 (cit. on pp. 17, 20, 21, 23, 31, 45, 47, 67, 68, 89).

[104] Daniel Kahneman, *Thinking, fast and slow*. macmillan, 2011 (cit. on pp. 17, 98).

[105] Jonathan St BT Evans, "In two minds: Dual-process accounts of reasoning," *Trends in cognitive sciences*, vol. 7, no. 10, pp. 454–459, 2003 (cit. on p. 17).

[106] John D Bransford et al., "The abstraction of linguistic ideas," *Cognitive psychology*, vol. 2, no. 4, pp. 331–350, 1971 (cit. on p. 17).

[107] Yoshua Bengio et al., "Deep learning for AI," *Commun. ACM*, vol. 64, no. 7, pp. 58–65, 2021. DOI: 10.1145/3448250 (cit. on pp. 17, 98, 100, 102).

[108] Joshua B Tenenbaum et al., "How to grow a mind: Statistics, structure, and abstraction," *science*, vol. 331, no. 6022, pp. 1279–1285, 2011 (cit. on pp. 17, 19, 21, 23, 47, 67, 103).

[109] Royi Lachmy et al., "Draw me a flower: Processing and grounding abstraction in natural language," *Trans. Assoc. Comput. Linguistics*, vol. 10, pp. 1341–1356, 2022 (cit. on p. 17).

[110] Thomas Mesnard et al., "Gemma: Open models based on gemini research and technology," *CoRR*, vol. abs/2403.08295, 2024. DOI: 10.48550/ARXIV.2403.08295 (cit. on pp. 17, 111).

[111] Weiqi Wang et al., "CAR: conceptualization-augmented reasoner for zero-shot commonsense question answering," in *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, Houda Bouamor et al., Eds., Association for Computational Linguistics, 2023, pp. 13 520–13 545. DOI: 10.18653/V1/2023.FINDINGS-EMNLP.902 (cit. on pp. 17, 68, 81, 85, 88, 89, 110).

[112] Amir Feder et al., "Causalm: Causal model explanation through counterfactual language models," *Comput. Linguistics*, vol. 47, no. 2, pp. 333–386, 2021. DOI: 10.1162/COLI_A_00404 (cit. on p. 17).

[113] Ziva Kunda et al., "Combining social concepts: The role of causal reasoning," *Cogn. Sci.*, vol. 14, no. 4, pp. 551–577, 1990. DOI: 10.1207/S15516709COG1404_3 (cit. on p. 17).

[114] Yonatan Bisk et al., "PIQA: reasoning about physical commonsense in natural language," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, AAAI Press, 2020, pp. 7432–7439 (cit. on pp. 17, 44, 49, 81, 94).

[115] Yi Ru Wang et al., "NEWTON: are large language models capable of physical reasoning?" In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, Houda Bouamor et al., Eds., Association for Computational Linguistics, 2023, pp. 9743–9758. DOI: 10.18653/V1/2023.FINDINGS-EMNLP.652 (cit. on p. 17).

[116] Yining Hong et al., "PTR: A benchmark for part-based conceptual, relational, and physical reasoning," in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, Marc'Aurelio Ranzato et al., Eds., 2021, pp. 17 427–17 440 (cit. on p. 17).

[117] Shan Yang et al., "Entity concept-enhanced few-shot relation extraction," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, Chengqing Zong et al., Eds., Association for Computational Linguistics, 2021, pp. 987–991. DOI: 10.18653/V1/2021.ACL-SHORT.124 (cit. on p. 20).

[118] Susan Carey, "Knowledge acquisition: Enrichment or conceptual change," *The epigenesis of mind: Essays on biology and cognition*, pp. 257–291, 1991 (cit. on pp. 20, 21).

[119] Faisal Ladhak et al., "Faithful or extractive? on mitigating the faithfulness-abstractiveness tradeoff in abstractive summarization," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, Smaranda Muresan et al., Eds., Association for Computational Linguistics, 2022, pp. 1410–1421. DOI: 10.18653/V1/2022.ACL-LONG.100 (cit. on p. 20).

[120] Wenbo Wang et al., "Concept pointer network for abstractive summarization," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui et al., Eds., Association for Computational Linguistics, 2019, pp. 3074–3083. DOI: 10.18653/V1/D19-1304 (cit. on p. 20).

[121] Virgile Rennard et al., "Abstractive meeting summarization: A survey," *Trans. Assoc. Comput. Linguistics*, vol. 11, pp. 861–884, 2023. DOI: 10.1162/TACL_A_00578 (cit. on p. 21).

[122] Hui Lin et al., "Abstractive summarization: A survey of the state of the art," in *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, AAAI Press, 2019, pp. 9815–9822. DOI: 10.1609/AAAI.V33I01.33019815 (cit. on p. 21).

[123] Ran Liu et al., "Sumsurvey: An abstractive dataset of scientific survey papers for long document summarization," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Lun-Wei Ku et al., Eds., Association for Computational Linguistics, 2024, pp. 9632–9651. DOI: 10.18653/V1/2024.FINDINGS-ACL.574 (cit. on p. 21).

[124] Stephanie M. Doane et al., "Expertise in a computer operating system: Conceptualization and performance," *Hum. Comput. Interact.*, vol. 5, no. 2-3, pp. 267–304, 1990. DOI: 10.1080/07370024.1990.9667156 (cit. on p. 21).

[125] Arjun Subramonian et al., "It takes two to tango: Navigating conceptualizations of NLP tasks and measurements of performance," in *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, Anna Rogers et al., Eds., Association for Computational Linguistics, 2023, pp. 3234–3279. DOI: 10.18653/V1/2023.FINDINGS-ACL.202 (cit. on p. 21).

[126] Ernest Davis, *Representations of commonsense knowledge* (notThenot Morgan Kaufmann series in representation and reasoning). Morgan Kaufmann, 1990 (cit. on pp. 23, 44).

[127] Ernest Davis et al., "Commonsense reasoning and commonsense knowledge in artificial intelligence," *Commun. ACM*, vol. 58, no. 9, pp. 92–103, 2015. DOI: 10.1145/2701413 (cit. on pp. 23, 88).

[128] Maarten Sap et al., "ATOMIC: an atlas of machine commonsense for if-then reasoning," in *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, AAAI Press, 2019, pp. 3027–3035. DOI: 10.1609/aaai.v33i01.33013027 (cit. on pp. 23, 27, 28, 44, 47, 64, 67–69, 99).

[129] Jena D. Hwang et al., "(comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs," in *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, AAAI Press, 2021, pp. 6384–6392 (cit. on pp. 23, 44, 57, 79).

[130] Nazneen Fatema Rajani et al., "Explain yourself! leveraging language models for commonsense reasoning," in *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen et al., Eds., Association for Computational Linguistics, 2019, pp. 4932–4942. DOI: 10.18653/v1/p19-1487 (cit. on p. 23).

[131] Jiacheng Liu et al., "Generated knowledge prompting for commonsense reasoning," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, Smaranda Muresan et al., Eds., Association for Computational Linguistics, 2022, pp. 3154–3169. DOI: 10.18653/v1/2022.acl-long.225 (cit. on p. 23).

[132] Changlong Yu et al., "Cocolm: Complex commonsense enhanced language model with discourse relations," in *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, Smaranda Muresan et al., Eds., Association for Computational Linguistics, 2022, pp. 1175–1187. DOI: 10.18653/v1/2022.findings-acl.93 (cit. on p. 23).

[133] Jan Nuyts et al., *Language and conceptualization*. Cambridge University Press, 1999 (cit. on p. 23).

[134] Richard C Anderson et al., "Instantiation of general terms," *Journal of Verbal Learning and Verbal Behavior*, vol. 15, no. 6, pp. 667–679, 1976 (cit. on pp. 23, 67).

[135] Emily Allaway et al., "Penguins don't fly: Reasoning about generics through instantiations and exceptions," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, Andreas Vlachos et al., Eds., Association for Computational Linguistics, 2023, pp. 2610–2627. DOI: 10.18653/V1/2023.EACL-MAIN.192 (cit. on pp. 23, 24, 28, 67, 68, 75, 89).

[136] Benjamin Van Durme et al., "Deriving generalized knowledge from corpora using wordnet abstraction," in *EACL 2009, 12th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, Athens, Greece, March 30 - April 3, 2009*, Alex Lascarides et al., Eds., The Association for Computer Linguistics, 2009, pp. 808–816 (cit. on pp. 24, 28, 47, 67, 68, 89).

[137] Yu Gong et al., "Representing verbs as argument concepts," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, Dale Schuurmans et al., Eds., AAAI Press, 2016, pp. 2615–2621 (cit. on pp. 24, 28, 47, 67).

[138] Mutian He et al., "On the role of conceptualization in commonsense knowledge graph construction," *CoRR*, vol. abs/2003.03239, 2020 (cit. on pp. 24, 67).

[139] Muhao Chen et al., "What are you trying to do? semantic typing of event processes," in *Proceedings of the 24th Conference on Computational Natural Language Learning, CoNLL 2020, Online, November 19-20, 2020*, Raquel Fernández et al., Eds., Association for Computational Linguistics, 2020, pp. 531–542. DOI: 10.18653/v1/2020.conll-1.43 (cit. on pp. 24, 67).

[140] Susan Carey, "Bootstrapping & the origin of concepts," *Daedalus*, vol. 133, no. 1, pp. 59–68, 2004 (cit. on pp. 25, 32).

[141] Steven Pinker et al., "The bootstrapping problem in language acquisition," *Mechanisms of language acquisition*, pp. 399–441, 1987 (cit. on p. 25).

[142] Jingping Liu et al., "Vocsk: Verb-oriented commonsense knowledge mining with taxonomy-guided induction," *Artif. Intell.*, vol. 310, p. 103 744, 2022. DOI: 10.1016/j.artint.2022.103744 (cit. on pp. 28, 47, 68, 89).

[143] Eunsol Choi et al., "Ultra-fine entity typing," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, Iryna Gurevych et al., Eds., Association for Computational Linguistics, 2018, pp. 87–96. DOI: 10.18653/v1/P18-1009 (cit. on p. 28).

[144] Hongliang Dai et al., "Ultra-fine entity typing with weak supervision from a masked language model," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, Chengqing Zong et al., Eds., Association for Computational Linguistics, 2021, pp. 1790–1799. DOI: 10.18653/v1/2021.acl-long.141 (cit. on p. 28).

[145] Bangzheng Li et al., "Ultra-fine entity typing with indirect supervision from natural language inference," *Trans. Assoc. Comput. Linguistics*, vol. 10, pp. 607–622, 2022. DOI: 10.1162/tacl_a_00479 (cit. on p. 28).

[146] Ian Porada et al., "Modeling event plausibility with consistent conceptual abstraction," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, Kristina Toutanova et al., Eds., Association for Computational Linguistics, 2021, pp. 1732–1743. DOI: 10.18653/v1/2021.naacl-main.138 (cit. on p. 28).

[147] Jesper E. van Engelen et al., "A survey on semi-supervised learning," *Mach. Learn.*, vol. 109, no. 2, pp. 373–440, 2020. DOI: 10.1007/s10994-019-05855-6 (cit. on p. 29).

[148] Ahmet Iscen et al., "Label propagation for deep semi-supervised learning," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, Computer Vision Foundation / IEEE, 2019, pp. 5070–5079. DOI: 10.1109/CVPR.2019.00521 (cit. on p. 29).

[149] Kexin Wang et al., "GPL: generative pseudo labeling for unsupervised domain adaptation of dense retrieval," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, Marine Carpuat et al., Eds., Association for Computational Linguistics, 2022, pp. 2345–2360. DOI: 10.18653/v1/2022.naacl-main.168 (cit. on p. 29).

[150] Fengbei Liu et al., "ACPL: anti-curriculum pseudo-labelling for semi-supervised medical image classification," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, IEEE, 2022, pp. 20 665–20 674. DOI: 10.1109/CVPR52688.2022.02004 (cit. on p. 29).

[151] Zijian Hu et al., "Simple: Similar pseudo label exploitation for semi-supervised classification," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, Computer Vision Foundation / IEEE, 2021, pp. 15 099–15 108. DOI: 10.1109/CVPR46437.2021.01485 (cit. on p. 29).

[152] Zheng Li et al., "Metats: Meta teacher-student network for multilingual sequence labeling with minimal supervision," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, Marie-Francine Moens et al., Eds., Association for Computational Linguistics, 2021, pp. 3183–3196. DOI: 10.18653/v1/2021.emnlp-main.255 (cit. on p. 29).

[153] Yu Meng et al., "Weakly-supervised hierarchical text classification," in *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, AAAI Press, 2019, pp. 6826–6833. DOI: 10.1609/aaai.v33i01.33016826 (cit. on p. 29).

[154] Huiru Xiao et al., "Efficient path prediction for semi-supervised and weakly supervised hierarchical text classification," in *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, Ling Liu et al., Eds., ACM, 2019, pp. 3370–3376. DOI: 10.1145/3308558.3313658 (cit. on p. 29).

[155] Tianqing Fang et al., "Pseudoreasoner: Leveraging pseudo labels for commonsense knowledge base population," in *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Yoav Goldberg et al., Eds., Association for Computational Linguistics, 2022, pp. 3379–3394. DOI: 10.18653/V1/2022.FINDINGS-EMNLP.246 (cit. on pp. 29, 33, 41, 78).

[156] Kun Liu et al., "Noisy-labeled NER with confidence estimation," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, Kristina Toutanova et al., Eds., Association for Computational Linguistics, 2021, pp. 3437–3445. DOI: 10.18653/v1/2021.naacl-main.269 (cit. on p. 29).

[157] Weile Chen et al., "Advpicker: Effectively leveraging unlabeled data via adversarial discriminator for cross-lingual NER," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, Chengqing Zong et al., Eds., Association for Computational Linguistics, 2021, pp. 743–753. DOI: 10.18653/v1/2021.acl-long.61 (cit. on p. 29).

[158] Liang Yao et al., "KG-BERT: BERT for knowledge graph completion," *CoRR*, vol. abs/1909.03193, 2019 (cit. on pp. 30, 33).

[159] Kevin Clark et al., "ELECTRA: pre-training text encoders as discriminators rather than generators," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, OpenReview.net, 2020 (cit. on pp. 33, 44).

[160] Qizhe Xie et al., "Unsupervised data augmentation for consistency training," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle et al., Eds., 2020 (cit. on p. 33).

[161] Qizhe Xie et al., "Self-training with noisy student improves imagenet classification," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, Computer Vision Foundation / IEEE, 2020, pp. 10 684–10 695. DOI: 10.1109/CVPR42600.2020.01070 (cit. on p. 33).

[162] Kishore Papineni et al., "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, ACL, 2002, pp. 311–318. DOI: 10.3115/1073083.1073135 (cit. on pp. 35, 79).

[163] Alon Lavie et al., "METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments," in *Proceedings of the Second Workshop on Statistical Machine Translation, WMT@ACL 2007, Prague, Czech Republic, June 23, 2007*, Chris Callison-Burch et al., Eds., Association for Computational Linguistics, 2007, pp. 228–231 (cit. on pp. 35, 79).

[164] Chin-Yew Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, Barcelona, Spain: Association for Computational Linguistics, 2004, pp. 74–81 (cit. on pp. 35, 79).

[165] Ramakrishna Vedantam et al., "Cider: Consensus-based image description evaluation," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, IEEE Computer Society, 2015, pp. 4566–4575. DOI: 10.1109/CVPR.2015.7299087 (cit. on pp. 35, 79).

[166] Margaret Li et al., "Don't say that! making inconsistent dialogue unlikely with unlikelihood training," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky et al., Eds., Association for Computational Linguistics, 2020, pp. 4715–4728. DOI: 10.18653/v1/2020.acl-main.428 (cit. on p. 35).

[167] Ari Holtzman et al., "The curious case of neural text degeneration," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, OpenReview.net, 2020 (cit. on p. 37).

[168] Tianyi Zhang et al., "Bertscore: Evaluating text generation with BERT," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, OpenReview.net, 2020 (cit. on pp. 37, 60, 79).

[169] Daniel Deutsch et al., "Understanding the extent to which content quality metrics measure the information quality of summaries," in *Proceedings of the 25th Conference on Computational Natural Language Learning, CoNLL 2021, Online, November 10-11, 2021*, Arianna Bisazza et al., Eds., Association for Computational Linguistics, 2021, pp. 300–309. DOI: 10.18653/v1/2021.conll-1.24 (cit. on p. 37).

[170] Yen-Chang Hsu et al., "Generalized ODIN: detecting out-of-distribution image without learning from out-of-distribution data," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, Computer Vision Foundation / IEEE, 2020, pp. 10 948–10 957. DOI: 10.1109/CVPR42600.2020.01096 (cit. on p. 37).

[171] Douglas B. Lenat et al., "CYC: toward programs with common sense," *Commun. ACM*, vol. 33, no. 8, pp. 30–49, 1990. DOI: 10.1145/79173.79176 (cit. on p. 44).

[172] Tom McCoy et al., "Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference," in *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen et al., Eds., Association for Computational Linguistics, 2019, pp. 3428–3448. DOI: 10.18653/v1/p19-1334 (cit. on p. 44).

[173] Kaixin Ma et al., "Towards generalizable neuro-symbolic systems for commonsense question answering," *CoRR*, vol. abs/1910.14087, 2019 (cit. on p. 44).

[174] Pei Zhou et al., "RICA: evaluating robust inference capabilities based on commonsense axioms," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, Marie-Francine Moens et al., Eds., Association for Computational Linguistics, 2021, pp. 7560–7579. DOI: 10.18653/v1/2021.emnlp-main.598 (cit. on p. 44).

[175] Peifeng Wang et al., "Do language models perform generalizable commonsense inference?" In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, Chengqing Zong et al., Eds., ser. Findings of ACL, vol. ACL/IJCNLP 2021, Association for Computational Linguistics, 2021, pp. 3681–3688. DOI: 10.18653/v1/2021.findings-acl.322 (cit. on p. 44).

[176] Ruben Branco et al., "Shortcutted commonsense: Data spuriousness in deep learning of commonsense reasoning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, Marie-Francine Moens et al., Eds., Association for Computational Linguistics, 2021, pp. 1504–1521. DOI: 10.18653/v1/2021.emnlp-main.113 (cit. on p. 44).

[177] Zhongli Li et al., "Harvesting and refining question-answer pairs for unsupervised QA," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Dan Jurafsky et al., Eds., Association for Computational Linguistics, 2020, pp. 6719–6728. DOI: 10.18653/v1/2020.acl-main.600 (cit. on p. 44).

[178] Liwei Jiang et al., ""i'm not mad": Commonsense implications of negation and contradiction," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, Kristina Toutanova et al., Eds., Association for Computational Linguistics, 2021. DOI: 10.18653/v1/2021.naacl-main.346 (cit. on p. 44).

[179] Yen-Ling Kuo et al., "Bridging common sense knowledge bases with analogy by graph similarity," in *Collaboratively-Built Knowledge Sources and Artificial Intelligence, Papers from the 2010 AAAI Workshop, Atlanta, Georgia, USA, July 11, 2010*, ser. AAAI Technical Report, vol. WS-10-02, AAAI, 2010 (cit. on p. 45).

[180] Silin Gao et al., "Peacok: Persona commonsense knowledge for consistent and engaging narratives," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, Anna Rogers et al., Eds., Association for Computational Linguistics, 2023. DOI: 10.18653/v1/2023.acl-long.362 (cit. on pp. 45, 47).

[181] Jiangjie Chen et al., "Say what you mean! large language models speak too positively about negative commonsense knowledge," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada: Association for Computational Linguistics, 2023, pp. 9890–9908 (cit. on p. 45).

[182] Swabha Swayamdipta et al., "Dataset cartography: Mapping and diagnosing datasets with training dynamics," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber et al., Eds., Association for Computational Linguistics, 2020, pp. 9275–9293. DOI: 10.18653/v1/2020.emnlp-main.746 (cit. on pp. 46, 54, 58, 59).

[183] Jeff Da et al., "Analyzing commonsense emergence in few-shot knowledge models," in *3rd Conference on Automated Knowledge Base Construction, AKBC 2021, Virtual, October 4-8, 2021*, Danqi Chen et al., Eds., 2021. DOI: 10.24432/C5NK5J (cit. on p. 47).

[184] Tri Nguyen et al., "MS MARCO: A human generated machine reading comprehension dataset," in *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*, Tarek Richard Besold et al., Eds., ser. CEUR Workshop Proceedings, vol. 1773, CEUR-WS.org, 2016 (cit. on p. 47).

[185] Jason W. Wei et al., "EDA: easy data augmentation techniques for boosting performance on text classification tasks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui et al., Eds., Association for Computational Linguistics, 2019, pp. 6381–6387. DOI: 10.18653/v1/D19-1670 (cit. on pp. 47, 53).

[186] William Yang Wang et al., "That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, Lluís Màrquez et al., Eds., The Association for Computational Linguistics, 2015, pp. 2557–2563. DOI: 10.18653/v1/d15-1306 (cit. on pp. 47, 53).

[187] Tong Niu et al., "Adversarial over-sensitivity and over-stability strategies for dialogue models," in *Proceedings of the 22nd Conference on Computational Natural Language Learning, CoNLL 2018, Brussels, Belgium, October 31 - November 1, 2018*, Anna Korhonen et al., Eds., Association for Computational Linguistics, 2018, pp. 486–496. DOI: 10.18653/v1/k18-1047 (cit. on pp. 47, 53).

[188] Rico Sennrich et al., "Improving neural machine translation models with monolingual data," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*, The Association for Computer Linguistics, 2016. DOI: 10.18653/v1/p16-1009 (cit. on p. 47).

[189] Mete Ismayilzada et al., "Kogito: A commonsense knowledge inference toolkit," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. EACL 2023 - System Demonstrations, Dubrovnik, Croatia, May 2-4, 2023*, Danilo Croce et al., Eds., Association for Computational Linguistics, 2023. DOI: 10.18653/v1/2023.eacl-demo.12 (cit. on p. 47).

[190] Chandra Bhagavatula et al., "Abductive commonsense reasoning," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, OpenReview.net, 2020 (cit. on pp. 49, 81, 94).

[191] Keisuke Sakaguchi et al., "Winogrande: An adversarial winograd schema challenge at scale," *Commun. ACM*, vol. 64, no. 9, pp. 99–106, 2021. DOI: 10.1145/3474381 (cit. on pp. 49, 81, 94).

[192] Yejin Bang et al., "A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity," in *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, IJCNLP 2023 -Volume 1: Long Papers, Nusa Dua, Bali, November 1 - 4, 2023*, Jong C. Park et al., Eds., Association for Computational Linguistics, 2023, pp. 675–718. DOI: 10.18653/V1/2023.IJCNLP-MAIN.45 (cit. on p. 53).

[193] Chunkit Chan et al., "Exploring the potential of chatgpt on sentence level relations: A focus on temporal, causal, and discourse relations," in *Findings of the Association for Computational Linguistics: EACL 2024, St. Julian's, Malta, March 17-22, 2024*, Yvette Graham et al., Eds., Association for Computational Linguistics, 2024, pp. 684–721 (cit. on pp. 53, 88, 98, 103).

[194] Chengwei Qin et al., "Is chatgpt a general-purpose natural language processing task solver?" In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, Houda Bouamor et al., Eds., Association for Computational Linguistics, 2023, pp. 1339–1384. DOI: 10.18653/V1/2023.EMNLP-MAIN.85 (cit. on pp. 53, 98, 103, 111).

[195] Sewon Min et al., "Rethinking the role of demonstrations: What makes in-context learning work?" In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Yoav Goldberg et al., Eds., Association for Computational Linguistics, 2022, pp. 11 048–11 064 (cit. on p. 53).

[196] Jason Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," in *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, Sanmi Koyejo et al., Eds., 2022 (cit. on pp. 53, 78, 81, 111).

[197] Joshua Robinson et al., "Leveraging large language models for multiple choice question answering," in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, OpenReview.net, 2023 (cit. on p. 53).

[198] Sosuke Kobayashi, "Contextual augmentation: Data augmentation by words with paradigmatic relations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, Marilyn A. Walker et al., Eds., Association for Computational Linguistics, 2018, pp. 452–457. DOI: 10.18653/v1/n18-2072 (cit. on p. 53).

[199] Ilya Loshchilov et al., "Decoupled weight decay regularization," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, OpenReview.net, 2019 (cit. on p. 53).

[200] Tomás Mikolov et al., "Efficient estimation of word representations in vector space," in *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, Yoshua Bengio et al., Eds., 2013 (cit. on p. 55).

[201] Jeffrey Pennington et al., "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, Alessandro Moschitti et al., Eds., Association for Computational Linguistics, 2014, pp. 1532–1543. DOI: 10.3115/v1/d14-1162 (cit. on p. 55).

[202] Mary L McHugh, "Interrater reliability: The kappa statistic," *Biochemia medica*, vol. 22, no. 3, 2012 (cit. on p. 55).

[203] Nils Reimers et al., "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui et al., Eds., Association for Computational Linguistics, 2019, pp. 3980–3990. DOI: 10.18653/V1/D19-1410 (cit. on pp. 55, 82).

[204] Zheye Deng et al., "Gold: A global and local-aware denoising framework for commonsense knowledge graph noise detection," in *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, Houda Bouamor et al., Eds., Association for Computational Linguistics, 2023, pp. 3591–3608. DOI: 10.18653/V1/2023.FINDINGS-EMNLP.232 (cit. on pp. 57, 84).

[205] Denny Vrandecic et al., "Wikidata: A free collaborative knowledgebase," *Commun. ACM*, vol. 57, no. 10, pp. 78–85, 2014. DOI: 10.1145/2629489 (cit. on pp. 61, 64).

[206] Ranjay Krishna et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, 2017. DOI: 10.1007/s11263-016-0981-7 (cit. on p. 64).

[207] Filip Ilievski et al., "CSKG: the commonsense knowledge graph," in *The Semantic Web - 18th International Conference, ESWC 2021, Virtual Event, June 6-10, 2021, Proceedings*, Ruben Verborgh et al., Eds., ser. Lecture Notes in Computer Science, vol. 12731, Springer, 2021, pp. 680–696. DOI: 10.1007/978-3-030-77385-4_41 (cit. on p. 64).

[208] Eduardo F Mortimer, "Conceptual change or conceptual profile change?" *Science & Education*, vol. 4, pp. 267–285, 1995 (cit. on p. 66).

[209] Mahzarin R Banaji et al., "The bankruptcy of everyday memory.," *American Psychologist*, vol. 44, no. 9, p. 1185, 1989 (cit. on p. 66).

[210] Tianqing Fang et al., "DISCOS: bridging the gap between discourse knowledge and commonsense knowledge," in *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, Jure Leskovec et al., Eds., ACM / IW3C2, 2021, pp. 2648–2659. DOI: 10.1145/3442381.3450117 (cit. on pp. 67, 88).

[211] Thomas Wolf et al., "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, Qun Liu et al., Eds., Association for Computational Linguistics, 2020, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6 (cit. on p. 73).

[212] Yaoming Zhu et al., "Texygen: A benchmarking platform for text generation models," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, Kevyn Collins-Thompson et al., Eds., ACM, 2018, pp. 1097–1100. DOI: 10.1145/3209978.3210080 (cit. on p. 75).

[213] Albert Q. Jiang et al., "Mistral 7b," *CoRR*, vol. abs/2310.06825, 2023. DOI: 10.48550/ARXIV.2310.06825 (cit. on pp. 78, 85, 92, 111).

[214] Edward J. Hu et al., "Lora: Low-rank adaptation of large language models," in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, OpenReview.net, 2022 (cit. on pp. 78, 111).

[215] Zhenmei Shi et al., "Why larger language models do in-context learning differently?" In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, OpenReview.net, 2024 (cit. on p. 78).

[216] Xuezhi Wang et al., "Self-consistency improves chain of thought reasoning in language models," in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, OpenReview.net, 2023 (cit. on pp. 81, 111).

[217] Weiqi Wang et al., "MARS: benchmarking the metaphysical reasoning abilities of language models with a multi-task evaluation dataset," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Wanxiang Che et al., Eds., Association for Computational Linguistics, 2025, pp. 1568–1596 (cit. on pp. 84, 94).

[218] Jiaxin Bai et al., "Complex query answering on eventuality knowledge graph with implicit logical constraints," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, Alice Oh et al., Eds., 2023 (cit. on p. 84).

[219] Wenxuan Ding et al., "Knowledge crosswords: Geometric knowledge reasoning with large language models," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Lun-Wei Ku et al., Eds., Association for Computational Linguistics, 2024, pp. 2609–2636. DOI: 10.18653/V1/2024.FINDINGS-ACL.154 (cit. on p. 84).

[220] Chunkit Chan et al., "Negotiationtom: A benchmark for stress-testing machine theory of mind on negotiation surrounding," in *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, Yaser Al-Onaizan et al., Eds., Association for Computational Linguistics, 2024, pp. 4211–4241. DOI: 10.18653/V1/2024.FINDINGS-EMNLP.244 (cit. on p. 84).

[221] Zheye Deng et al., "Text-tuple-table: Towards information integration in text-to-table generation via global tuple extraction," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, Yaser Al-Onaizan et al., Eds., Association for Computational Linguistics, 2024, pp. 9300–9322. DOI: 10.18653/V1/2024.EMNLP-MAIN.523 (cit. on p. 84).

[222] OpenAI, "Hello gpt-4o," *OpenAI*, 2024 (cit. on p. 88).

[223] OpenAI, "Gpt-4o mini: Advancing cost-efficient intelligence," *OpenAI*, 2024 (cit. on pp. 88, 111).

[224] Abhimanyu Dubey et al., "The llama 3 herd of models," *CoRR*, vol. abs/2407.21783, 2024. DOI: 10.48550/ARXIV.2407.21783 (cit. on pp. 88, 92, 111, 117).

[225] Ningyu Zhang et al., "A comprehensive study of knowledge editing for large language models," *CoRR*, vol. abs/2401.01286, 2024. DOI: 10.48550/ARXIV.2401.01286 (cit. on p. 88).

[226] Song Wang et al., "Knowledge editing for large language models: A survey," *ACM Comput. Surv.*, vol. 57, no. 3, 59:1–59:37, 2025. DOI: 10.1145/3698590 (cit. on pp. 88, 91).

[227] Ching Ming Samuel Lau et al., *Ecomedit: An automated e-commerce knowledge editing framework for enhanced product and purchase intention understanding*, 2024. [Online]. Available: https://arxiv.org/abs/2410.14276 (cit. on p. 88).

[228] Tianjie Ju et al., "Investigating multi-hop factual shortcuts in knowledge editing of large language models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, Lun-Wei Ku et al., Eds., Association for Computational Linguistics, 2024, pp. 8987–9001. DOI: 10.18653/V1/2024.ACL-LONG.486 (cit. on p. 88).

[229] Jiaan Wang et al., "Cross-lingual knowledge editing in large language models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, Lun-Wei Ku et al., Eds., Association for Computational Linguistics, 2024, pp. 11 676–11 686. DOI: 10.18653/V1/2024.ACL-LONG.627 (cit. on p. 88).

[230] Derong Xu et al., "Editing factual knowledge and explanatory ability of medical large language models," in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM 2024, Boise, ID, USA, October 21-25, 2024*, Edoardo Serra et al., Eds., ACM, 2024, pp. 2660–2670. DOI: 10.1145/3627673.3679673 (cit. on p. 88).

[231] Xiusheng Huang et al., "Commonsense knowledge editing based on free-text in llms," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, Yaser Al-Onaizan et al., Eds., Association for Computational Linguistics, 2024, pp. 14 870–14 880 (cit. on p. 88).

[232] Zonglin Yang et al., "End-to-end case-based reasoning for commonsense knowledge base completion," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, Andreas Vlachos et al., Eds., Association for Computational Linguistics, 2023, pp. 3491–3504. DOI: 10.18653/V1/2023.EACL-MAIN.255 (cit. on p. 88).

[233] Wenxuan Ding et al., "Intentionqa: A benchmark for evaluating purchase intention comprehension abilities of language models in e-commerce," in *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, Yaser Al-Onaizan et al., Eds., Association for Computational Linguistics, 2024, pp. 2247–2266. DOI: 10.18653/V1/2024.FINDINGS-EMNLP.123 (cit. on p. 88).

[234] Baixuan Xu et al., "MIND: multimodal shopping intention distillation from large vision-language models for e-commerce purchase understanding," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, Yaser Al-Onaizan et al., Eds., Association for Computational Linguistics, 2024, pp. 7800–7815. DOI: 10.18653/V1/2024.EMNLP-MAIN.446 (cit. on p. 88).

[235] Kaixin Ma et al., "Exploring strategies for generalizable commonsense reasoning with pre-trained models," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, Marie-Francine Moens et al., Eds., Association for Computational Linguistics, 2021, pp. 5474–5483. DOI: 10.18653/v1/2021.emnlp-main.445 (cit. on p. 88).

[236] Weiqi Wang et al., "On the role of entity and event level conceptualization in generalizable reasoning: A survey of tasks, methods, applications, and future directions," in *Findings of the Association for Computational Linguistics: EMNLP 2025*, Christos Christodoulopoulos et al., Eds., Suzhou, China: Association for Computational Linguistics, 2025, pp. 2260–2281. DOI: 10.18653/v1/2025.findings-emnlp.122 (cit. on p. 88).

[237] Nicola De Cao et al., "Editing factual knowledge in language models," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, Marie-Francine Moens et al., Eds., Association for Computational Linguistics, 2021, pp. 6491–6506. DOI: 10.18653/V1/2021.EMNLP-MAIN.522 (cit. on p. 89).

[238] Kevin Meng et al., "Locating and editing factual associations in GPT," in *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, Sanmi Koyejo et al., Eds., 2022 (cit. on pp. 89, 92).

[239] Kevin Meng et al., "Mass-editing memory in a transformer," in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, OpenReview.net, 2023 (cit. on pp. 89, 92).

[240] Tom Hartvigsen et al., "Aging with GRACE: lifelong model editing with discrete key-value adaptors," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, Alice Oh et al., Eds., 2023 (cit. on pp. 89, 92).

[241] Vyas Raina et al., "Is llm-as-a-judge robust? investigating universal adversarial attacks on zero-shot LLM assessment," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, Yaser Al-Onaizan et al., Eds., Association for Computational Linguistics, 2024, pp. 7499–7517 (cit. on p. 90).

[242] Weiqi Wang et al., "Ecomscriptbench: A multi-task benchmark for e-commerce script planning via step-wise intention-driven product association," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Wanxiang Che et al., Eds., Association for Computational Linguistics, 2025, pp. 1–22 (cit. on pp. 90, 160).

[243] Aohan Zeng et al., "Chatglm: A family of large language models from GLM-130B to GLM-4 all tools," *CoRR*, vol. abs/2406.12793, 2024. DOI: 10.48550/ARXIV.2406.12793 (cit. on p. 92).

[244] Ben Wang et al., *GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model*, https://github.com/kingoflolz/mesh-transformer-jax, 2021 (cit. on p. 92).

[245] Haochen Shi et al., *Inferencedynamics: Efficient routing across llms through structured capability and knowledge profiling*, 2025. [Online]. Available: https://arxiv.org/abs/2505.16303 (cit. on p. 94).

[246] Chunyang Li et al., "Patterns over principles: The fragility of inductive reasoning in llms under noisy observations," in *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Wanxiang Che et al., Eds., Association for Computational Linguistics, 2025, pp. 19 608–19 626 (cit. on p. 94).

[247] Baixuan Xu et al., *Towards multi-agent reasoning systems for collaborative expertise delegation: An exploratory design study*, 2025. [Online]. Available: https://arxiv.org/abs/2505.07313 (cit. on p. 95).

[248] Xiao Liu et al., "The magic of IF: investigating causal reasoning abilities in large language models of code," in *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, Anna Rogers et al., Eds., Association for Computational Linguistics, 2023, pp. 9009–9022. DOI: 10.18653/V1/2023.FINDINGS-ACL.574 (cit. on p. 98).

[249] Dohwan Ko et al., "Large language models are temporal and causal reasoners for video question answering," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, Houda Bouamor et al., Eds., Association for Computational Linguistics, 2023, pp. 4300–4316. DOI: 10.18653/V1/2023.EMNLP-MAIN.261 (cit. on p. 98).

[250] Raghav Jain et al., "Do language models have a common sense regarding time? revisiting temporal commonsense reasoning in the era of large language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, Houda Bouamor et al., Eds., Association for Computational Linguistics, 2023, pp. 6750–6774. DOI: 10.18653/V1/2023.EMNLP-MAIN.418 (cit. on pp. 98, 103).

[251] Jacob Andreas, "Language models as agent models," in *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Yoav Goldberg et al., Eds., Association for Computational Linguistics, 2022, pp. 5769–5779. DOI: 10.18653/V1/2022.FINDINGS-EMNLP.423 (cit. on p. 98).

[252] Steven A Sloman, "The empirical case for two systems of reasoning.," *Psychological bulletin*, vol. 119, no. 1, p. 3, 1996 (cit. on p. 98).

[253] Yoshua Bengio, "The consciousness prior," *CoRR*, vol. abs/1709.08568, 2017 (cit. on p. 98).

[254] Xu Huang et al., "Understanding the planning of LLM agents: A survey," *CoRR*, vol. abs/2402.02716, 2024. DOI: 10.48550/ARXIV.2402.02716 (cit. on p. 98).

[255] Brenden M. Lake et al., "Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, Jennifer G. Dy et al., Eds., ser. Proceedings of Machine Learning Research, vol. 80, PMLR, 2018, pp. 2879–2888 (cit. on p. 98).

[256] Dzmitry Bahdanau et al., "Systematic generalization: What is required and can it be learned?" In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, OpenReview.net, 2019 (cit. on p. 98).

[257] Harm de Vries et al., "CLOSURE: assessing systematic generalization of CLEVR models," in *Visually Grounded Interaction and Language (ViGIL), NeurIPS 2019 Workshop, Vancouver, Canada, December 13, 2019*, 2019 (cit. on p. 98).

[258] Karthik Valmeekam et al., "Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, Alice Oh et al., Eds., 2023 (cit. on pp. 100, 102, 107).

[259] Weinan He et al., "Exploring the capacity of pretrained language models for reasoning about actions and change," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, Anna Rogers et al., Eds., Association for Computational Linguistics, 2023, pp. 4629–4643. DOI: 10.18653/V1/2023.ACL-LONG.255 (cit. on pp. 100, 102, 107).

[260] Martin Heidegger, *Introduction to metaphysics*. Yale University Press, 2014 (cit. on p. 100).

[261] Stephen Merity et al., "Pointer sentinel mixture models," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, OpenReview.net, 2017 (cit. on pp. 100, 105).

[262] Yukun Zhu et al., "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, IEEE Computer Society, 2015, pp. 19–27. DOI: 10.1109/ICCV.2015.11 (cit. on pp. 100, 105).

[263] Yoshua Bengio et al., "From system 1 deep learning to system 2 deep learning," in *Neural Information Processing Systems*, 2019 (cit. on p. 102).

[264] Niket Tandon et al., "Reasoning about actions and state changes by injecting commonsense knowledge," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, Ellen Riloff et al., Eds., Association for Computational Linguistics, 2018, pp. 57–66. DOI: 10.18653/V1/D18-1006 (cit. on p. 102).

[265] Nena Basina et al., "ECAVI: an assistant for reasoning about actions and change with the event calculus," in *Domain-Specific Conceptual Modeling - Concepts, Methods and ADOxx Tools*, Dimitris Karagiannis et al., Eds., Springer, 2022, pp. 457–477. DOI: 10.1007/978-3-030-93547-4_20 (cit. on p. 102).

[266] Youzhi Zhang et al., "A fuzzy reasoning model for action and change in timed domains," *Int. J. Intell. Syst.*, vol. 28, no. 8, pp. 787–805, 2013. DOI: 10.1002/INT.21602 (cit. on p. 102).

[267] Bo Wu et al., "STAR: A benchmark for situated reasoning in real-world videos," in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, Joaquin Vanschoren et al., Eds., 2021 (cit. on p. 102).

[268] Joshua Maynez et al., "Benchmarking large language model capabilities for conditional generation," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, Anna Rogers et al., Eds., Association for Computational Linguistics, 2023, pp. 9194–9213. DOI: 10.18653/V1/2023.ACL-LONG.511 (cit. on p. 103).

[269] Yihan Chen et al., "Benchmarking large language models on controllable generation under diversified instructions," in *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, Michael J. Wooldridge et al., Eds., AAAI Press, 2024, pp. 17 808–17 816. DOI: 10.1609/AAAI.V38I16.29734 (cit. on pp. 103, 125).

[270] Jiawei Chen et al., "Benchmarking large language models in retrieval-augmented generation," in *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, Michael J. Wooldridge et al., Eds., AAAI Press, 2024, pp. 17 754–17 762. DOI: 10.1609/AAAI.V38I16.29728 (cit. on p. 103).

[271] Qingyu Tan et al., "Towards benchmarking and improving the temporal reasoning capability of large language models," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, Anna Rogers et al., Eds., Association for Computational Linguistics, 2023, pp. 14 820–14 835. DOI: 10.18653/V1/2023.ACL-LONG.828 (cit. on pp. 103, 113).

[272] Chenhan Yuan et al., "Back to the future: Towards explainable temporal reasoning with large language models," in *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, Tat-Seng Chua et al., Eds., ACM, 2024, pp. 1963–1974. DOI: 10.1145/3589334.3645376 (cit. on p. 103).

[273] Dhairya Dalal et al., "Calm-bench: A multi-task benchmark for evaluating causality-aware language models," in *Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, Andreas Vlachos et al., Eds., Association for Computational Linguistics, 2023, pp. 296–311. DOI: 10.18653/V1/2023.FINDINGS-EACL.23 (cit. on p. 103).

[274] Zhijing Jin et al., "Cladder: A benchmark to assess causal reasoning capabilities of language models," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, Alice Oh et al., Eds., 2023 (cit. on p. 103).

[275] Ning Bian et al., "Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, Nicoletta Calzolari et al., Eds., ELRA and ICCL, 2024, pp. 3098–3110 (cit. on p. 103).

[276] Shuofei Qiao et al., "Reasoning with language model prompting: A survey," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, Anna Rogers et al., Eds., Association for Computational Linguistics, 2023, pp. 5368–5393. DOI: 10.18653/V1/2023.ACL-LONG.294 (cit. on p. 103).

[277] Fausto Giunchiglia et al., "A theory of abstraction," *Artificial intelligence*, vol. 57, no. 2-3, pp. 323–389, 1992 (cit. on p. 103).

[278] Aristotle Aristotle et al., *Metaphysics*. Harvard University Press Cambridge, MA, 1933, vol. 1 (cit. on p. 104).

[279] Henri Bergson, *An introduction to metaphysics*. Hackett Publishing Company, 1999 (cit. on p. 104).

[280] Jiaxuan Li et al., "Counterfactual reasoning: Testing language models' understanding of hypothetical scenarios," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, Anna Rogers et al., Eds., Association for Computational Linguistics, 2023, pp. 804–815. DOI: 10.18653/V1/2023.ACL-SHORT.70 (cit. on p. 105).

[281] Wenyue Hua et al., "Propagation and pitfalls: Reasoning-based assessment of knowledge editing through counterfactual tasks," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Lun-Wei Ku et al., Eds., Association for Computational Linguistics, 2024, pp. 12 503–12 525. DOI: 10.18653/V1/2024.FINDINGS-ACL.743 (cit. on p. 105).

[282] Yunhu Ye et al., "Large language models are versatile decomposers: Decomposing evidence and questions for table-based reasoning," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, Hsin-Hsi Chen et al., Eds., ACM, 2023, pp. 174–184. DOI: 10.1145/3539618.3591708 (cit. on p. 105).

[283] Harsh Jhamtani et al., "Natural language decomposition and interpretation of complex utterances," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, ijcai.org, 2024, pp. 6306–6314 (cit. on p. 105).

[284] Bhavana Dalvi et al., "Tracking state changes in procedural text: A challenge dataset and models for process paragraph comprehension," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, Marilyn A. Walker et al., Eds., Association for Computational Linguistics, 2018, pp. 1595–1604. DOI: 10.18653/V1/N18-1144 (cit. on p. 107).

[285] Joseph L Fleiss, "Measuring nominal scale agreement among many raters.," *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971 (cit. on p. 109).

[286] Zhengxiang Shi et al., "Rethinking semi-supervised learning with language models," in *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, Anna Rogers et al., Eds., Association for Computational Linguistics, 2023, pp. 5614–5634. DOI: 10.18653/V1/2023.FINDINGS-ACL.347 (cit. on p. 110).

[287] Ebtesam Almazrouei et al., "The falcon series of open language models," *CoRR*, vol. abs/2311.16867, 2023. DOI: 10.48550/ARXIV.2311.16867 (cit. on p. 111).

[288] Yunfan Gao et al., "Retrieval-augmented generation for large language models: A survey," *CoRR*, vol. abs/2312.10997, 2023. DOI: 10.48550/ARXIV.2312.10997 (cit. on p. 111).

[289] Ruixin Yang et al., "Confidence calibration and rationalization for llms via multi-agent deliberation," *CoRR*, vol. abs/2404.09127, 2024. DOI: 10.48550/ARXIV.2404.09127 (cit. on p. 111).

[290] Liangming Pan et al., "Automatically correcting large language models: *Surveying the Landscape of Diverse Automated Correction Strategies*," *Trans. Assoc. Comput. Linguistics*, vol. 12, pp. 484–506, 2024. DOI: 10.1162/TACL_A_00660 (cit. on p. 111).

[291] Freda Shi et al., "Large language models can be easily distracted by irrelevant context," in *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Andreas Krause et al., Eds., ser. Proceedings of Machine Learning Research, vol. 202, PMLR, 2023, pp. 31 210–31 227 (cit. on p. 113).

[292] Junyi Li et al., "Halueval: A large-scale hallucination evaluation benchmark for large language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, Houda Bouamor et al., Eds., Association for Computational Linguistics, 2023, pp. 6449–6464. DOI: 10.18653/V1/2023.EMNLP-MAIN.397 (cit. on p. 113).

[293] Hongbin Ye et al., "Cognitive mirage: A review of hallucinations in large language models," in *Proceedings of the First International OpenKG Workshop: Large Knowledge-Enhanced Models co-locacted with The International Joint Conference on Artificial Intelligence (IJCAI 2024), Jeju Island, South Korea, August 3, 2024*, Ningyu Zhang et al., Eds., ser. CEUR Workshop Proceedings, vol. 3818, CEUR-WS.org, 2024, pp. 14–36 (cit. on p. 113).

[294] Anthropic, "Introducing the next generation of claude," *Anthropic Announcements*, 2024 (cit. on p. 117).

[295] Sondre Wold et al., "Compositional generalization with grounded language models," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Lun-Wei Ku et al., Eds., Association for Computational Linguistics, 2024, pp. 3447–3460. DOI: 10.18653/V1/2024.FINDINGS-ACL.205 (cit. on p. 124).

[296] Yuhang Zang et al., "Overcoming the pitfalls of vision-language model finetuning for OOD generalization," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, OpenReview.net, 2024 (cit. on p. 125).

[297] Fanyi Qu et al., "Unsupervised distractor generation via large language model distilling and counterfactual contrastive decoding," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Lun-Wei Ku et al., Eds., Association for Computational Linguistics, 2024, pp. 827–838. DOI: 10.18653/V1/2024.FINDINGS-ACL.47 (cit. on p. 125).

[298] Zhihong Shao et al., "Deepseekmath: Pushing the limits of mathematical reasoning in open language models," *CoRR*, vol. abs/2402.03300, 2024. DOI: 10.48550/ARXIV.2402.03300 (cit. on pp. 125, 126).

[299] Oleh Rybkin et al., "Value-based deep RL scales predictably," in *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, OpenReview.net, 2025 (cit. on p. 125).

[300] Siyuan Wang et al., "Benchmark self-evolving: A multi-agent framework for dynamic LLM evaluation," in *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, Owen Rambow et al., Eds., Association for Computational Linguistics, 2025, pp. 3310–3328 (cit. on p. 126).

[301] Weiqi Wang et al., "Heapa: Difficulty-aware heap sampling and on-policy query augmentation for LLM reinforcement learning," *CoRR*, vol. abs/2601.22448, 2026. DOI: 10.48550/ARXIV.2601.22448 (cit. on p. 128).

[302] Ming Dai et al., "Improving generalized visual grounding with instance-aware joint learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 48, no. 1, pp. 448–465, 2026. DOI: 10.1109/TPAMI.2025.3607387 (cit. on p. 129).

[303] Qiying Yu et al., "DAPO: An open-source LLM reinforcement learning system at scale," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025 (cit. on p. 129).

[304] Zihan Zheng et al., "Planningarena: A modular benchmark for multidimensional evaluation of planning and tool learning," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Wanxiang Che et al., Eds., Association for Computational Linguistics, 2025, pp. 31 047–31 086 (cit. on p. 129).

[305] Changlong Yu et al., "Folkscope: Intention knowledge graph construction for e-commerce commonsense discovery," in *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, Anna Rogers et al., Eds., Association for Computational Linguistics, 2023, pp. 1173–1191. DOI: 10.18653/V1/2023.FINDINGS-ACL.76 (cit. on p. 163).

[306] Liyu Zhang et al., "Conke: Conceptualization-augmented knowledge editing in large language models for commonsense reasoning," in *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Wanxiang Che et al., Eds., Association for Computational Linguistics, 2025, pp. 627–635 (cit. on p. 163).

[307] Tianshi Zheng et al., "Knowshiftqa: How robust are RAG systems when textbook knowledge shifts in K-12 education?" In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Wanxiang Che et al., Eds., Association for Computational Linguistics, 2025, pp. 183–195. DOI: 10.18653/V1/2025.ACL-SHORT.16 (cit. on p. 163).

[308] Weiqi Wang et al., "Arxiv2table: Toward realistic benchmarking and evaluation for llm-based literature-review table generation," *CoRR*, vol. abs/2504.10284, 2025. DOI: 10.48550/ARXIV.2504.10284 (cit. on p. 163).

[309] Tianshi Zheng et al., "From automation to autonomy: A survey on large language models in scientific discovery," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Christos Christodoulopoulos et al., Eds., Suzhou, China: Association for Computational Linguistics, 2025, pp. 17 733–17 750. DOI: 10.18653/v1/2025.emnlp-main.895 (cit. on p. 163).

[310] Wei Fan et al., "Legal rule induction: Towards generalizable principle discovery from analogous judicial precedents," *CoRR*, vol. abs/2505.14104, 2025. DOI: 10.48550/ARXIV.2505.14104 (cit. on p. 163).

[311] Yuqi Yang et al., "Sessionintentbench: A multi-task inter-session intention-shift modeling benchmark for e-commerce customer behavior understanding," *CoRR*, vol. abs/2507.20185, 2025. DOI: 10.48550/ARXIV.2507.20185 (cit. on p. 163).

[312] Rui Wang et al., "Prospect theory fails for llms: Revealing instability of decision-making under epistemic uncertainty," *CoRR*, vol. abs/2508.08992, 2025. DOI: 10.48550/ARXIV.2508.08992 (cit. on p. 163).

[313] Zheye Deng et al., "Structuring the unstructured: A systematic review of text-to-structure generation for agentic AI with a universal evaluation framework," *CoRR*, vol. abs/2508.12257, 2025. DOI: 10.48550/ARXIV.2508.12257 (cit. on p. 163).

[314] Baixuan Xu et al., "The cognitive bandwidth bottleneck: Shifting long-horizon agent from planning with actions to planning with schemas," *CoRR*, vol. abs/2510.07091, 2025. DOI: 10.48550/ARXIV.2510.07091 (cit. on p. 163).

[315] Tianshi Zheng et al., "Newtonbench: Benchmarking generalizable scientific law discovery in LLM agents," *CoRR*, vol. abs/2510.07172, 2025. DOI: 10.48550/ARXIV.2510.07172 (cit. on p. 163).

[316] Qing Zong et al., "Critical: Can critique help LLM uncertainty or confidence calibration?" *CoRR*, vol. abs/2510.24505, 2025. DOI: 10.48550/ARXIV.2510.24505 (cit. on p. 163).

[317] Yuetong Wu et al., "Knowcomp at dialam-2024: Fine-tuning pre-trained language models for dialogical argument mining with inference anchoring theory," in *Proceedings of the 11th Workshop on Argument Mining, ArgMining 2024, Bangkok, Thailand, August 15, 2024*, Yamen Ajjour et al., Eds., Association for Computational Linguistics, 2024, pp. 103–109 (cit. on p. 163).

[318] Wei Fan et al., "Chain-of-choice hierarchical policy learning for conversational recommendation," in *Database Systems for Advanced Applications - 29th International Conference, DASFAA 2024, Gifu, Japan, July 2-5, 2024, Proceedings, Part V*, Makoto Onizuka et al., Eds., ser. Lecture Notes in Computer Science, vol. 14854, Springer, 2024, pp. 120–137. DOI: 10.1007/978-981-97-5569-1_8 (cit. on p. 163).

[319] Wei Fan et al., "Goldcoin: Grounding large language models in privacy laws via contextual integrity theory," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, Yaser Al-Onaizan et al., Eds., Association for Computational Linguistics, 2024, pp. 3321–3343. DOI: 10.18653/V1/2024.EMNLP-MAIN.195 (cit. on p. 163).

[320] Feihong Lu et al., "Miko: Multimodal intention knowledge distillation from large language models for social-media commonsense discovery," in *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, Jianfei Cai et al., Eds., ACM, 2024, pp. 3303–3312. DOI: 10.1145/3664647.3681339 (cit. on p. 163).

[321] Weiqi Wang et al., "Knowcomp at semeval-2024 task 9: Conceptualization-augmented prompting with large language models for lateral reasoning," in *Proceedings of the 18th International Workshop on Semantic Evaluation, SemEval@NAACL 2024, Mexico City, Mexico, June 20-21, 2024*, Atul Kr. Ojha et al., Eds., Association for Computational Linguistics, 2024, pp. 1639–1645. DOI: 10.18653/V1/2024.SEMEVAL-1.233 (cit. on p. 163).

[322] Ying Su et al., "Entaile: Introducing textual entailment in commonsense knowledge graph completion," *CoRR*, vol. abs/2402.09666, 2024. DOI: 10.48550/ARXIV.2402.09666 (cit. on p. 163).

[323] Xin Liu et al., "Towards subgraph isomorphism counting with graph kernels," *CoRR*, vol. abs/2405.07497, 2024. DOI: 10.48550/ARXIV.2405.07497 (cit. on p. 163).

[324] Liyu Zhang et al., "Conceptedit: Conceptualization-augmented knowledge editing in large language models for commonsense reasoning," *CoRR*, vol. abs/2412.11418, 2024. DOI: 10.48550/ARXIV.2412.11418 (cit. on p. 163).

[325] Jiaxin Bai et al., "Intention knowledge graph construction for user intention relation modeling," in *19th Conference of the European Chapter of the Association for Computational Linguistics*, 2026 (cit. on p. 163).

[326] Zhaowei Wang et al., "COLA: contextualized commonsense causal reasoning from the causal inference perspective," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, Anna Rogers et al., Eds., Association for Computational Linguistics, 2023, pp. 5253–5271. DOI: 10.18653/V1/2023.ACL-LONG.288 (cit. on p. 163).

[327] Qing Zong et al., "TILFA: A unified framework for text, image, and layout fusion in argument mining," in *Proceedings of the 10th Workshop on Argument Mining, ArgMining 2023, Singapore, December 7, 2023*, Milad Alshomary et al., Eds., Association for Computational Linguistics, 2023, pp. 139–147. DOI: 10.18653/V1/2023.ARGMINING-1.14 (cit. on p. 163).

[328] Cheng Jiayang et al., "Storyanalogy: Deriving story-level analogies from large language models to unlock analogical understanding," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, Houda Bouamor et al., Eds., Association for Computational Linguistics, 2023, pp. 11 518–11 537. DOI: 10.18653/V1/2023.EMNLP-MAIN.706 (cit. on p. 163).

[329] Weiqi Wang et al., "Knowcomp at semeval-2023 task 7: Fine-tuning pre-trained language models for clinical trial entailment identification," in *Proceedings of the The 17th International Workshop on Semantic Evaluation, SemEval@ACL 2023, Toronto, Canada, 13-14 July 2023*, Atul Kr. Ojha et al., Eds., Association for Computational Linguistics, 2023, pp. 1–9. DOI: 10.18653/V1/2023.SEMEVAL-1.1 (cit. on p. 163).

[330] Yi Wu et al., "Knowcomp submission for WMT23 word-level autocompletion task," in *Proceedings of the Eighth Conference on Machine Translation, WMT 2023, Singapore, December 6-7, 2023*, Philipp Koehn et al., Eds., Association for Computational Linguistics, 2023, pp. 882–889. DOI: 10.18653/V1/2023.WMT-1.79 (cit. on p. 163).

[331] Baixuan Xu et al., "Knowcomp submission for WMT23 sign language translation task," in *Proceedings of the Eighth Conference on Machine Translation, WMT 2023, Singapore, December 6-7, 2023*, Philipp Koehn et al., Eds., Association for Computational Linguistics, 2023, pp. 351–358. DOI: 10.18653/V1/2023.WMT-1.36 (cit. on p. 163).

APPENDIX A

PROMPTS USED IN MARS

A.1 MARS Benchmark Curation

An overview of our benchmark construction pipeline is shown in Figure 8.3. We first present our prompts used in each step for sequentially instructing ChatGPT to generate candidate data for MARS [242].

A.1.1 Text Decomposition and Event Component Extraction

To decompose a lengthy text from the source corpora into several action events, we use the following prompt to instruct ChatGPT.

> **You are required to decompose the given long sentence into several short yet semantically complete events, each describing an action. An action event refers to those describing an action or a state change that occurs at a specific time and place. The key components of each event should be preserved: including the subject, verb, object, temporal and spatial quantifiers, numerical properties of the subject and objects, and sub-events. Generate one event as a whole sentence per line. You can generate as many events as you need. Below are some examples:**
>
> ...
>
> **Sentence <i>: In November 2010, after years of planning and development, SpaceX successfully launched their Falcon 9 rocket into orbit for the first time. The launch took place at Cape Canaveral Air Force Station in Florida. The Falcon 9 carried a Dragon spacecraft mock-up, representing a major milestone in SpaceX's efforts to develop a reliable and cost-effective means of transporting cargo and eventu-**

ally astronauts to the International Space Station.
Event 1: SpaceX successfully launched their Falcon 9 rocket into orbit for the first time in November 2010.
Event 2: The Falcon 9 carried a Dragon spacecraft mock-up.
Event 3: The launch of the Falcon 9 took place at Cape Canaveral Air Force Station in Florida.
...
Sentence <N>: In May 1934, following reports of a Japanese spy operating out of Dutch Harbor, the United States Navy dispatched Edwin T. Layton to the Aleutians to investigate the allegations.

We then use the following prompt to extract seven types of components from the decomposed events.

Given a short event, extract these components:
1. Subject: The noun that performs the action in the sentence.
2. Verb: The action word in the sentence.
3. Object: The noun that receives the action of the verb.
4. Temporal Quantifier: The time or time period of the event in the sentence.
5. Spatial Quantifier: The location or spatial extent of the event in the sentence.
6. Numerical Quantities and Properties of Objects: Numerical values describing the number or properties of the subject, object, or sub-events.
7. Sub-events: Complete events that are part of the main event in the sentence.
For each component, if there are more than one, separate them with |. If you cannot find one for a component, generate ``None'' only. Below are some examples:
...
Event <i>: After the First Battle of Naktong Bulge, the US Army's 2nd Infantry Division was moved to defend the Naktong

River line.
Subject: US Army's 2nd Infantry Division
Verb: moved | defend
Object: None
Temporal Quantifier: After the First Battle of Naktong Bulge
Spatial Quantifier: Naktong River line
Quantities and Properties of Objects: None
Sub-events: The US Army's 2nd Infantry Division was moved | The US Army's 2nd Infantry Division was moved to defend the Naktong River line.
...
Event <N>: The University of Colorado created the Department of Medicine in September 1883 in the Old Main building on the Boulder campus.

A.1.2 Component Abstraction and Variation

For each type of component, we customize the prompt according to the nature of the component and whether the changes are implemented via abstraction or numerical variation. Here, we take the subject category with its abstraction as an example.

Given an event and a subject within the event, abstract the given subject in the given sentence into three different concepts. Each concept should be more abstract than the previous one. You are encouraged to be creative, but please ensure the three concepts gradually cover more instances. Below are some examples:
...
Event <i>: World's leading scientists announce breakthrough in clean energy technology, revolutionizing global sustainability efforts.
Subject: World's leading scientists
Concepts: expert, human, organism
...
Event <N>: A driver is speeding down the highway.

Subject: A driver

Note that leveraging LLM to perform contextualized abstraction [305–331] has been shown to result in better quality, larger coverage, and stronger downstream benefits compared to previous conceptualization methods [78, 85], such as retrieving from a pre-defined concept taxonomy or human annotation. Our knowledge distillation-based method is justifiable and enables large-scale benchmark construction.

A.1.3 Inference Generation

We use different prompts to collect plausible inferential states and metaphysical inferential states for each changed action event. Here, we provide the prompt for generating a metaphysical inference as an example.

Given an action event, generate a short metaphysical if-then inferential statement that describes an inferential state that only occurs in metaphysical space. A state is a condition or situation in which someone or something exists in the past or present that will last for a certain time if no changes occur. An action is a thing that can be done in a time interval that is usually not long. Metaphysical inference is a type of inference that is not based on empirical evidence but rather on the nature of things. It can be a counterfactual inference that is contrary to the facts or reality, meaning that it is usually not true in reality world. Below are some examples:

...

Event <i>: In 2003, he played a recurring role on two episodes of The Bill.

Metaphysical Inference: Everyone criticizes his performance in the show.

...

Event <N>: Sam drives down the road with fast speed.

Model	Task 1 Plaus.	Expert.	Task 2 Plaus.	Expert.	Task 3 Plaus.	Expert.
ChatGPT	60.98	92.0	58.56	96.5	50.25	93.5
Meta-LLaMa-3.1-405B	62.2	93.2	57.0	95.8	51.0	94.6
GPT-4o	64.6	94.8	59.2	98.4	53.4	96.0

Table A.1: Annotation results of evaluation data curated with different LLMs as backbones. Plaus. refers to plausible event/inference/transition rate and Expert. refers to ratio of data accepted by expert annotators.

A.1.4 Metaphysical Transition Generation

Finally, we use the prompt below to collect the change needed to transition a metaphysical inference into a plausible one.

You will be given an event and its metaphysical inference, meaning that such an inference is impossible or rarely occurring in reality. Please generate a transition that would make the inference plausible or possible in real life. Specifically, you are required to only change a component of the event. The component must be one of the Subject, Verb, Object, Temporal Quantifier, Spatial Quantifier, Numerical Properties of Subject or Objects, and Sub-events of the event. Below are some examples:

...

Event <i>: The boss of the company is monitoring the employees.

Metaphysical Inference: The boss feels nervous and is expecting a rise.

Transition: employees -> stocks (Object)

...

Event <N>: The man is being chased by a 100 meters butterfly in the forest.

Metaphysical Inference: The man is not scared and is laughing.

A.2 Main Evaluations on MARS

To evaluate LLMs on three tasks in MARS, we show our evaluating prompts in zero-shot scenario in Table A.2. Note that we are aware that LLMs may not be familiar with the word "metaphysical." Therefore, we also experimented with replacing the word with "implausible," and the best performances from both types of prompts are reported. These models are consistent across all models' evaluations for fair comparison.

For few-shot evaluations, few shot examples are added after task descriptions and before the prompted test entry. The exemplars are randomly sampled for each different test entry. For COT prompting, we specifically ask LLMs to "think step by step and generate a short rationale to support your reasoning." Then, we ask it to give an answer based on its generated rationale. The sampling temperature τ is set to 0.1 by default, and 5 COT responses are sampled with τ set to 0.7 in the SC-COT setting.

A.3 Leveraging Open-sourced LLM for Benchmark Curation

In Chapter 8, we use proprietary LLMs and human annotation for data construction, which can be expensive and labor-intensive. However, this approach serves the best pursuit of data quality, which is crucial for an evaluation benchmark. Prior to our data collection, we tested a wide variety of LLMs, and ChatGPT outperformed almost all of them. Therefore, we opted to use it for data construction. Nevertheless, with the recent advancements in state-of-the-art LLMs, we have found that `meta-llama/Llama-3.1-405B-Instruct` and `GPT-4o` also achieve satisfactory performance within our data collection framework. We sampled 500 original data entries and employed similar prompts and data collection processes to gather metaphysical reasoning evaluation data entries. We then asked expert annotators to rate the plausibility of the obtained data. The results are shown in Table A.1. We observe that LLAMA3.1-405B can achieve comparable performance to ChatGPT in terms of plausible data (evaluation data that reflects reality rather than metaphysics, similar to the majority vote results in Table 2) and expert acceptance rates. Additionally, we find that GPT-4o can even improve the data collection process, resulting in higher quality data. Thus, we believe this represents a compromise between data quality, reproducibility, and cost. It would also be feasible for data collectors to use LLAMA3.1 in the future for collecting metaphysical data, although leveraging proprietary LLMs can be more reliable to some extent.

Task	Prompt
ME.	Given an event, determine whether it is a metaphysical event or not. A metaphysical event refers to event that is implausible or rarely occurring in reality. If it is plausible and commonly accepted in the real world, answer yes. On the contrary, if the event is metaphysical, answer No. The event you need to discriminate is: **<TEST-ENTRY-EVENT>**. Answer Yes or No only with one word:
MI.	Given an assertion that describes a if-then inference, determine whether the inference is plausible or metaphysical. A plausible inference is an inference that is likely to be true or reasonable based on the information provided in the assertion. A metaphysical inference is an inference that is not based on empirical evidence but rather on the nature of things, it rarely occurs in the real world and can be counterfactual or implausible. The assertion is: **<TEST-ENTRY-INFERENCE>**. Answer Yes or No only with one word.
MT.	You are given an event, an inference based on the event that rarely occurs in the real world (a metaphysical inference), and a transition in the event that would make the inference plausible or possible in the real world, please determine whether the transition is correct or not in terms of making the inference plausible or possible. The event is: **<TEST-ENTRY-EVENT>**. The inference is: **<TEST-ENTRY-INFERENCE>**. The transition is: **<TEST-ENTRY-TRANSITION>**. Answer Yes or No only with one word.

Table A.2: Prompts used for evaluating LLMs across three tasks in MARS in zero-shot scenario. ME. MI., and MT. stand for three tasks, respectively.

A.4 Additional Statistics on MARS

Table 8.3 presents detailed statistics on the number of unique identified and modified components by type in the annotated splits of each task. The majority (approximately 80%) of the components focus on the subject, verb, and object, while the remainder (around 20%) concentrate on temporal quantifiers, spatial quantifiers, numerical properties, and sub-events. On average, each annotated event in MARS features 8.15 identified components for changes and 7.81 transitions.