自我介绍与研究经历分享

王伟琪 WEIQI WANG

2025/09/04

HTTPS://MIGHTY-WEAVER.GITHUB.IO/

教育背景

- 香港科技大学（HKUST）
  - **博士研究生（计算机科学与工程），2022九月 — 2026八月，导师：宋阳秋教授**
    - 香港政府博士奖学金获得者（Hong Kong PhD Fellowship Awardee）
    - 博士毕业论文：《大语言模型可泛化推理研究：从概念化到形而上学的框架与基准》
      From Conceptualization to Metaphysics: Frameworks and Benchmarks for Generalizable Reasoning with Large Language Models
  - **理学士（计算机科学和数学），2018九月 — 2022 七月，导师：宋阳秋教授**
    - 一等荣誉学位，香港科技大学本科奖学金获得者
    - 本科毕业论文：《常识推理中的事件转移》 Event Transition in Commonsense Reasoning

- 约翰斯·霍普金斯大学（Johns Hopkins University）
  - **访问博士学者，2024十一月 — 2025四月，导师：Daniel Khashabi、Benjamin Van Durme**
    - 研究课题：大语言模型驱动的科学知识探索 Scientific Knowledge Exploration via LLMs

实习与学术经历

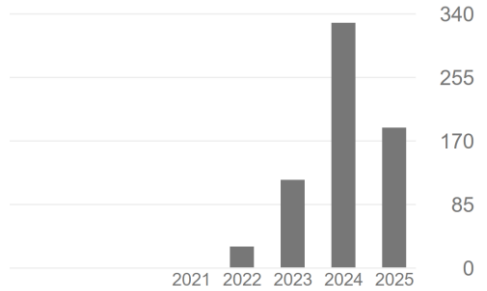
- Amazon Search Experience Science， 应用科学家实习生，2024/06-2024/09
  - 导师：崔荔萌博士、刘昕博士、罗琛博士
  - 研究主题：构建基准与方法：探索大语言模型驱动的电商脚本规划与客户智能体模拟

- Amazon Stores Foundational AI，应用科学家实习生，2025/06-2025/12
  - 导师：刘昕博士、杨靖锋博士、尹庆宇博士
  - 研究主题：突破强化学习Scaling Law瓶颈：通过动态数据生成提升大模型训练效率

- 领域主席 Area Chair:
  - ACL(2024,2025), EMNLP(2024,2025), COLING(2025), NAACL(2025), COLM(2025), ICML(2025), ICLR(2026)

论文发表情况

Publications	50
h-index	16
Citations	691
Highly Influential Citations	26



[HKUST] Weiqi Wang*, Tianqing Fang*, Baixuan Xu, Chun Yi Louis Bo, Yangqiu Song, and Lei Chen. CAT: A Contextualized Conceptualization and Instantiation Framework for Commonsense Reasoning. **ACL 2023 (Oral)**

[HKUST] Weiqi Wang*, Tianqing Fang*, Wenxuan Ding, Baixuan Xu, Xin Liu, Yangqiu Song, and Antoine Bosselut. CAR: Conceptualization-Augmented Reasoner for Zero-Shot Commonsense Question Answering. **Findings of EMNLP 2023**

[HKUST] Weiqi Wang, Tianqing Fang, Chunyang Li, Haochen Shi, Wenxuan Ding, Baixuan Xu, Zhaowei Wang, et al. CANDLE: Iterative Conceptualization and Instantiation Distillation from Large Language Models for Commonsense Reasoning. **ACL 2024.**

[HKUST] Weiqi Wang, Yangqiu Song. MARS: Benchmarking the Metaphysical Reasoning Abilities of Language Models with a Multi-task Evaluation Dataset. **ACL 2025.**

[HKUST] Weiqi Wang, Tianqing Fang, Haochen Shi, Baixuan Xu, Wenxuan Ding, Liyu Zhang, Wei Fan, Jiaxin Bai, Haoran Li, Xin Liu, Yangqiu Song. On the Role of Entity and Event Level Conceptualization in Generalizable Reasoning: A Survey of Tasks, Methods, Applications, and Future Directions. **Findings of EMNLP 2025.**

[JHU] Weiqi Wang, Jiefu Ou, Yangqiu Song, Benjamin Van Durme, Daniel Khashabi. Can LLMs Generate Tabular Summaries of Science Papers? Rethinking the Evaluation Protocol. **Under Review at ACL 2026.**

[Amazon] Weiqi Wang, Limeng Cui, Xin Liu, Sreyashi Nag, Wenju Xu, Sheikh Sarwar, Chen Luo, Yang Laurence Li, Hansu Gu, Hui Liu, Changlong Yu, Jiaxin Bai, Yifan Gao, Haiyang Zhang, Qi He, Shuiwang Ji, Yangqiu Song. EcomScriptBench: A Multi-task Benchmark for E-commerce Script Planning via Step-wise Intention-Driven Product Association. **ACL 2025.**

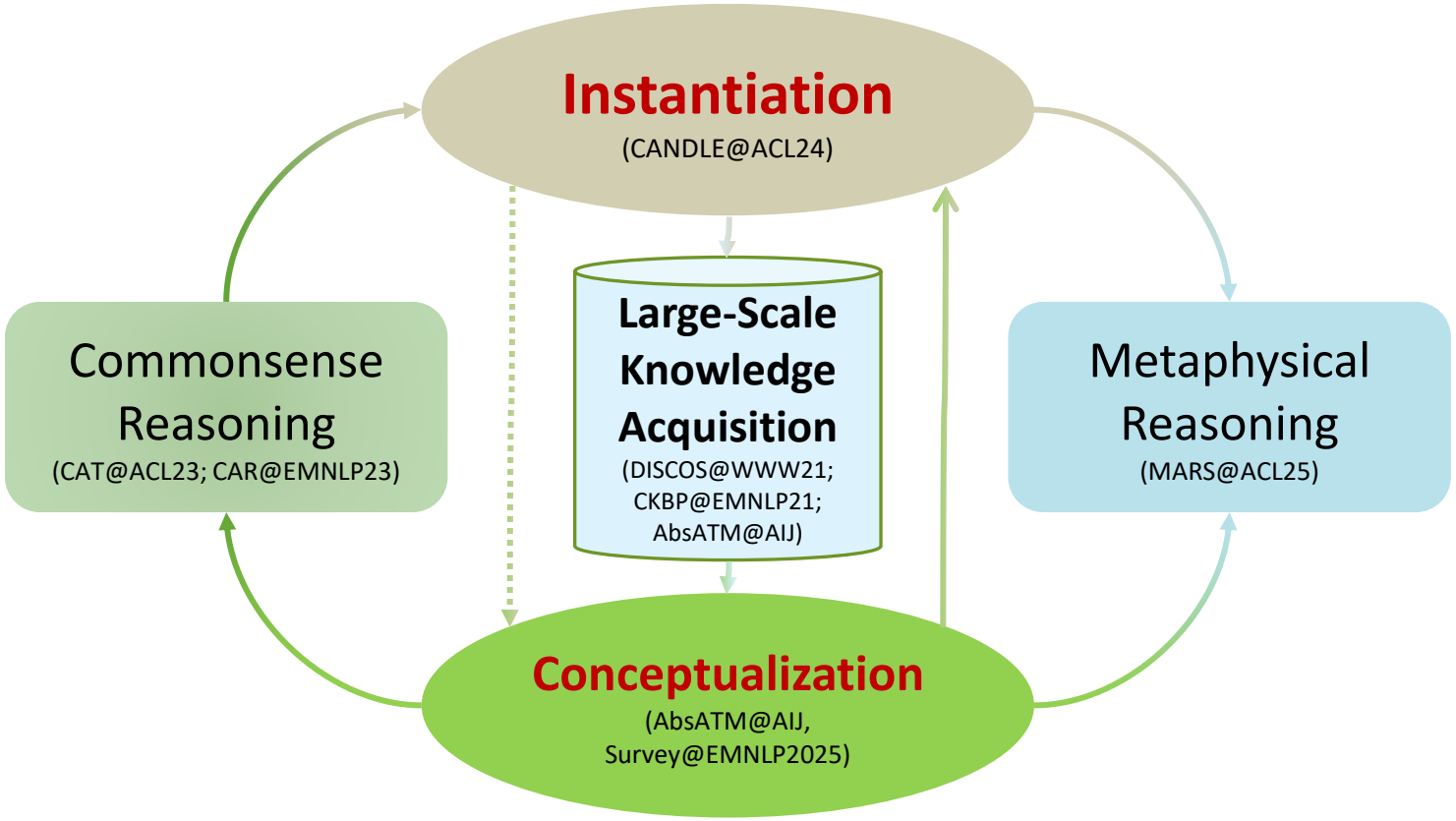
[Amazon] Weiqi Wang, Xin Liu, Jingfeng Yang, Hejie Cui, Binxuan Huang, Changlong Yu, Zheng Li, Yangqiu Song, Bing Yin. Breaking the RL Scaling Plateau: Dynamic Query Generation for Improved RL Training Data Efficiency. **Targeting ICLR 2026.**

本科/硕士/博士生科研指导

涵盖Topic：不同领域下的任务基准构建、知识编辑、电商意图理解、多模态推理、大模型不确定性Uncertainty、多智能体推理、多模型协作Routing，等等

Wenxuan Ding*, Weiqi Wang*, Sze Heng Douglas Kwok, Minghao Liu, Tianqing Fang, Jiaxin Bai, Xin Liu, Changlong Yu, Zheng Li, Chen Luo, Qingyu Yin, Bing Yin, Junxian He, Yangqiu Song. IntentionQA: A Benchmark for Evaluating Purchase Intention Comprehension Abilities of Language Models in E-commerce. Findings of EMNLP 2024.
Baixuan Xu*, Weiqi Wang*, Haochen Shi, Wenxuan Ding, Huihao Jing, Tianqing Fang, Jiaxin Bai, Xin Liu, Changlong Yu, Zheng Li, Chen Luo, Qingyu Yin, Bing Yin, Long Chen, Yangqiu Song. MIND: Multimodal Shopping Intention Distillation from Large Vision-language Models for E-commerce Purchase Understanding. EMNLP 2024.
Feihong Lu*, Weiqi Wang*, Yangyifei Luo, Ziqin Zhu, Qingyun Sun, Baixuan Xu, Haochen Shi, Shiqi Gao, Qian Li, Yangqiu Song, Jianxin Li. MIKO: Multimodal Intention Knowledge Distillation from Large Language Models for Social-Media Commonsense Discovery. ACM MM 2024.
Haochen Shi*, Weiqi Wang*, Tianqing Fang, Baixuan Xu, Wenxuan Ding, Xin Liu, Yangqiu Song. QaDynamics: Training Dynamics-Driven Synthetic QA Diagnostic for Zero-Shot Commonsense Question Answering. Findings of EMNLP 2023.
Liyu Zhang*, Weiqi Wang*, Tianqing Fang, Yangqiu Song. ConKE: Conceptualization-Augmented Knowledge Editing in Large Language Models for Commonsense Reasoning. Findings of ACL 2025.
Chunyang Li, Weiqi Wang, Tianshi Zheng, Yangqiu Song. Patterns Over Principles: The Fragility of Inductive Reasoning in LLMs under Noisy Observations. Findings of ACL 2025.
Zheye Deng, Weiqi Wang, Zhaowei Wang, Xin Liu, Yangqiu Song. Gold: A Global and Local-aware Denoising Framework for Commonsense Knowledge Graph Noise Detection. Findings of EMNLP 2023.
Jiayu Liu, Qing Zong, Weiqi Wang, Yangqiu Song. Revisiting Epistemic Markers in Confidence Estimation: Can Markers Accurately Reflect Large Language Models' Uncertainty?. ACL 2025.
Yuqi Yang*, Weiqi Wang*, Baixuan Xu, Wei Fan, Qing Zong, Chunkit Chan, Zheye Deng, Xin Liu, Yifan Gao, Changlong Yu, Chen Luo, Yang Li, Zheng Li, Qingyu Yin, Bing Yin, Yangqiu Song. SessionIntentBench: A Multi-task Inter-Session Intention-Shift Modeling Benchmark for E-commerce Customer Behavior Understanding. Under Review at EACL 2026.
Baixuan Xu*, Chunyang Li*, Weiqi Wang*, Wei Fan, Tianshi Zheng, Haochen Shi, Tao Fan, Yangqiu Song, Qiang Yang. Towards Multi-Agent Reasoning Systems for Collaborative Expertise Delegation: An Exploratory Design Study. Under Review at ACL 2026.
Haochen Shi*, Tianshi Zheng*, Weiqi Wang*, Baixuan Xu, Chunyang Li, Chunkit Chan, Tao Fan, Yangqiu Song, Qiang Yang. InferenceDynamics: Efficient Routing Across LLMs through Structured Capability and Knowledge Profiling. Under Review at ACL 2026.
Hongyu Luo, Weiqi Wang, Haochen Shi, Tianshi Zheng, Qing Zong, Baixuan Xu, Chunyang Li, Yangqiu Song. PAINT with Words: A Dataset and Cognitive Framework for Evaluating Visual Creativity. Under Review at EACL 2026.
Ching Ming Samuel Lau*, Weiqi Wang*, Haochen Shi, Baixuan Xu, Jiaxin Bai, Yangqiu Song. EcomEdit: An Automated E-commerce Knowledge Editing Framework for Enhanced Product and Purchase Intention Understanding. Under Review at EACL 2026.

HKUST PhD Career Research Theme

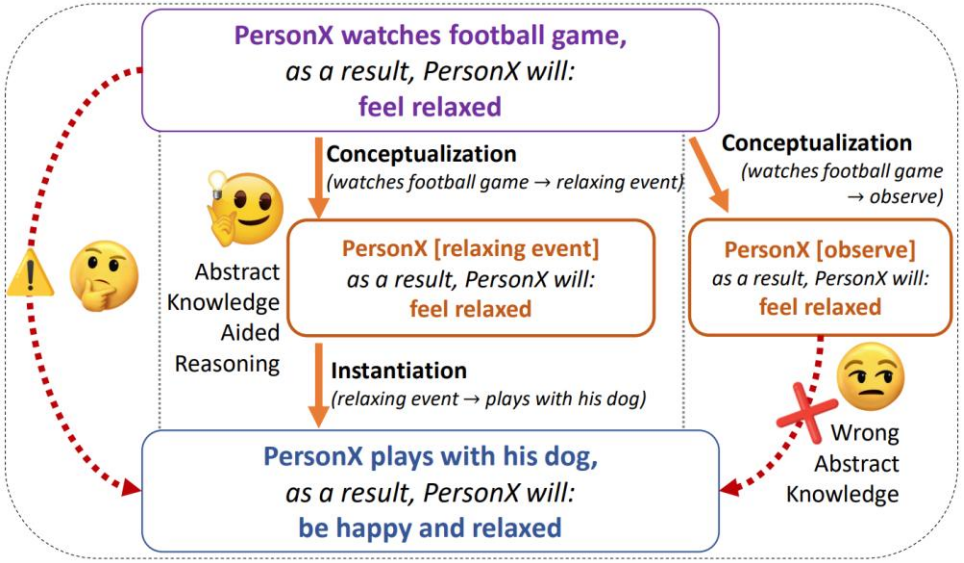


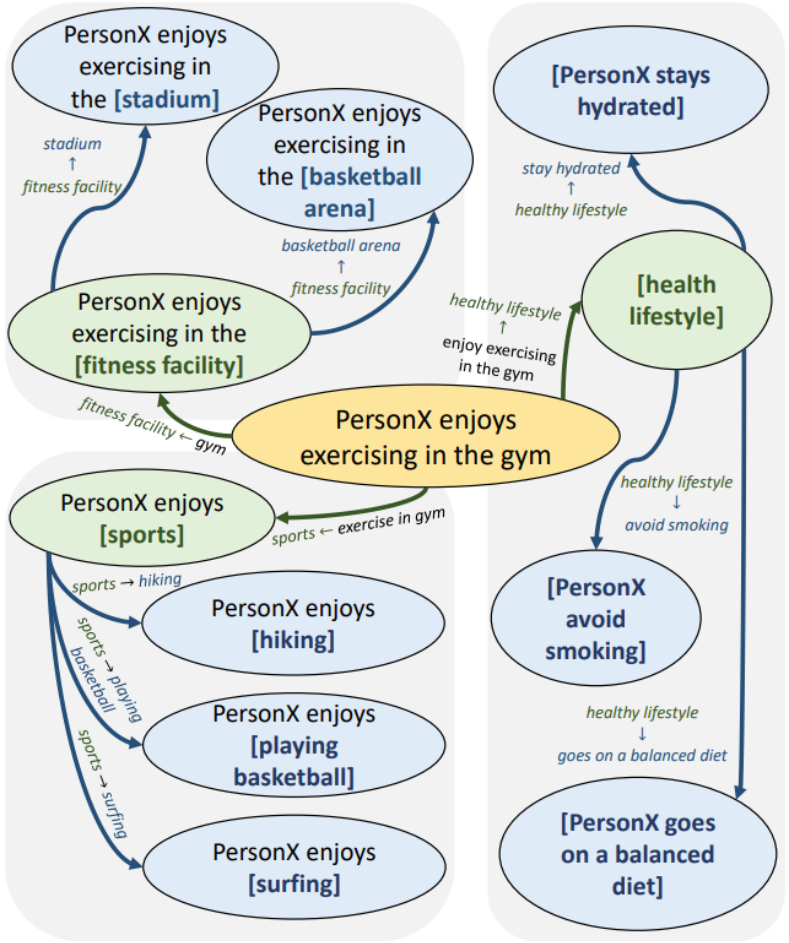
Tianqing Fang, Hongming Zhang, Weiqi Wang, Yangqiu Song, and Bin He. DISCOS: Bridging the Gap between Discourse Knowledge and Commonsense Knowledge. WWW 2021
Tianqing Fang*, Weiqi Wang*, Sehyun Choi, Shibo Hao, Hongming Zhang, Yangqiu Song, and Bin He. Benchmarking Commonsense Knowledge Base Population with an Effective Evaluation Dataset. EMNLP 2021
Mutian He, Tianqing Fang, Weiqi Wang, Yangqiu Song. Acquiring and Modelling Abstract Commonsense Knowledge via Conceptualization. Artificial Intelligence
Weiqi Wang*, Tianqing Fang*, Baixuan Xu, Chun Yi Louis Bo, Yangqiu Song, and Lei Chen. CAT: A Contextualized Conceptualization and Instantiation Framework for Commonsense Reasoning. ACL 2023 (Oral)
Weiqi Wang*, Tianqing Fang*, Wenxuan Ding, Baixuan Xu, Xin Liu, Yangqiu Song, and Antoine Bosselut. CAR: Conceptualization-Augmented Reasoner for Zero-Shot Commonsense Question Answering. Findings of EMNLP 2023
Weiqi Wang, Tianqing Fang, Chunyang Li, Haochen Shi, Wenxuan Ding, Baixuan Xu, Zhaowei Wang, et al. CANDLE: Iterative Conceptualization and Instantiation Distillation from Large Language Models for Commonsense Reasoning. ACL 2024.
Weiqi Wang, Yangqiu Song. Metaphysical Reasoning: Benchmark and Baselines. To be submitted to NeurIPS 2024 Dataset and Benchmark Track. ACL 2025

Conceptualization and Instantiation

概念化（Conceptualization）： 将一组实体或事件抽象为一个通用概念，从而在其原有语境中形成抽象知识。

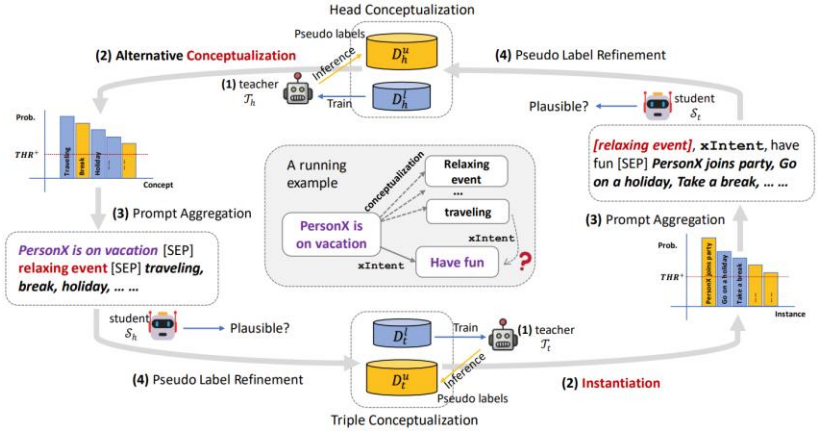
实例化（Instantiation）： 将所抽象出的概念落实到其他实例和事件中，以引入新的具体知识。



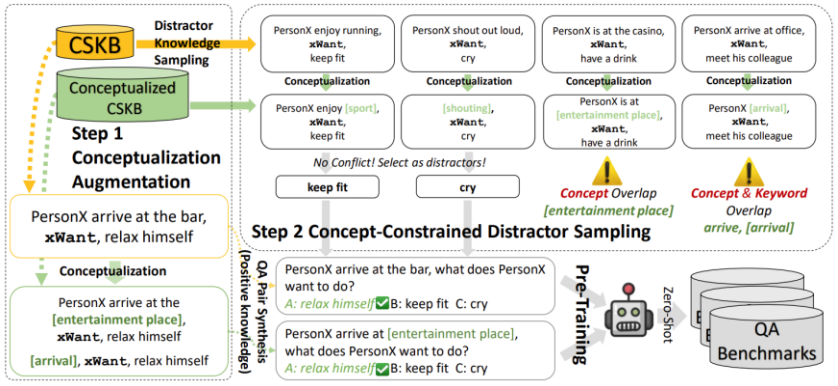


Conceptualization and Instantiation

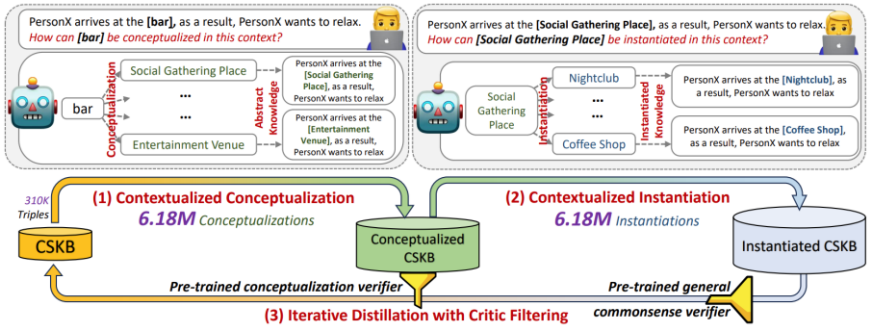
我们首先提出了一种**弱监督训练方法**，通过匹配 Probase、WordNet 等概念语料库中的知识，实现任意知识的**概念化与实例化**，效果显著优于传统的监督学习方法



我们进一步将**概念化**视作一种知识增广手段，设计了一个**数据合成框架**，将概念化知识融入**负采样流程**并**扩充训练数据**，从而训练出更强大的问答模型，性能超越以往所有方法及 GPT-3.5



最后，我们将**概念化与实例化**整合为一套完整的知识增强流程，提出了一个**迭代式知识蒸馏框架**：从 GPT-3.5中循环获取概念化与实例化结果，并通过严格的批判性过滤机制筛选高质量数据。借助这一增广方式，我们训练出的模型在多个基准任务上均超越最佳基准与GPT-3.5



Conceptualization and Instantiation

在已有数据的基础上，我们将所提出的方法应用于多类常识推理任务，并在以下方向上均取得了显著进展：

- 常识知识库的概念化与实例化
- 常识知识生成
- 常识问答

值得强调的是，我们的方法不依赖特定的知识表示形式，能够在任意类型的数据上有效运作，展现了较强的通用性与扩展性。

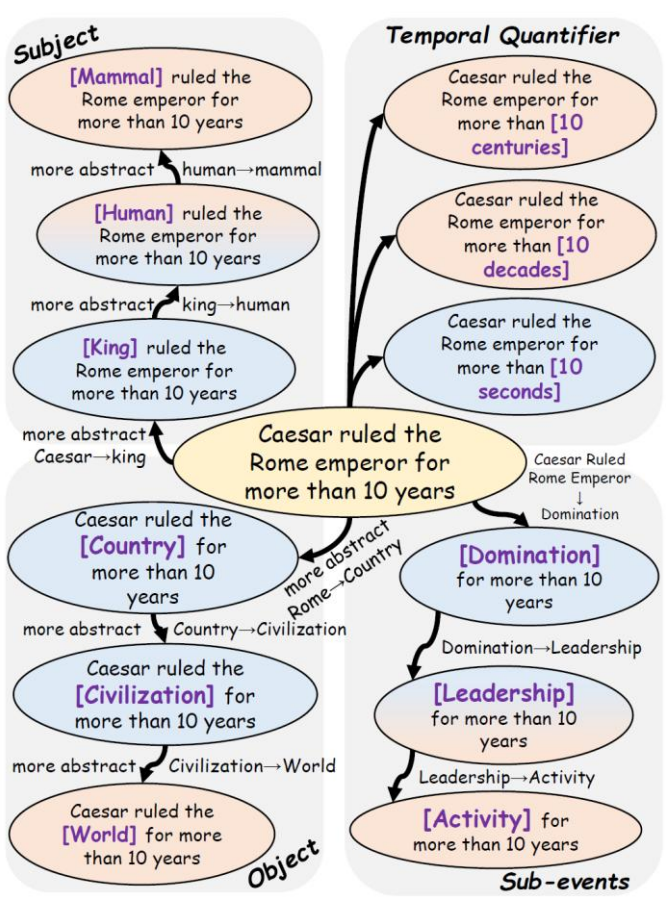
Model/Method	CSKB	a-NLI	CSQA	PIQA	SIQA	WG	Avg.
Pre-trained Language Models							
RoBERTa-L (Liu et al., 2019)	-	65.5	45.0	67.6	47.3	57.5	56.6
DeBERTa-v3-L (He et al., 2023)	-	59.9	25.4	44.8	47.8	50.3	45.6
Self-talk (Shwartz et al., 2020)	-	-	32.4	70.2	46.2	54.7	-
SMLM (Banerjee and Baral, 2020)	*	65.3	38.8	-	48.5	-	-
COMET-DynGen (Bosselut et al., 2021)	ATOMIC	-	-	-	50.1	-	-
MICO (Su et al., 2022)	ATOMIC	-	44.2	-	56.0	-	-
STL-Adapter (Kim et al., 2022)	ATOMIC	71.3	66.5	71.1	64.4	60.3	66.7
DeBERTa-v3-L (MR) (Ma et al., 2021)	ATM10X	75.1	71.6	79.0	59.7	71.7	71.4
DeBERTa-v3-L (MR) (Ma et al., 2021)	ATOMIC	76.0	67.0	78.0	62.1	76.0	71.8
CAR-DeBERTa-v3-L (Wang et al., 2023a)	ATOMIC	78.9	67.2	78.6	63.8	78.1	73.3
CAR-DeBERTa-v3-L (Wang et al., 2023a)	AbsATM	79.6	69.3	78.6	64.0	78.2	73.9
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	75.5	79.2	**70.7**	63.6	71.9
+ Self-consistent chain-of-thought	-	73.2	**75.7**	81.7	69.7	64.1	72.9
GPT-4 (`gpt-4`)	-	75.0	43.0	73.0	57.0	77.0	65.0
LLAMA2 (7B; Touvron et al., 2023)	-	57.5	57.8	78.8	48.3	69.2	62.3
LLAMA2 (13B; Touvron et al., 2023)	-	55.9	67.3	80.2	50.3	72.8	65.3
Mistral-v0.1 (7B; Jiang et al., 2023)	-	51.0	59.6	**83.0**	42.9	75.3	62.4
VERA-T5-xxl (Liu et al., 2023)	ATOMIC	71.2	61.7	76.4	57.7	67.5	66.9
VERA-T5-xxl (Liu et al., 2023)	ATM10X	70.3	59.5	75.1	58.2	67.2	66.1
VERA-T5-xxl (Liu et al., 2023)	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$_{\uparrow 1.4}$

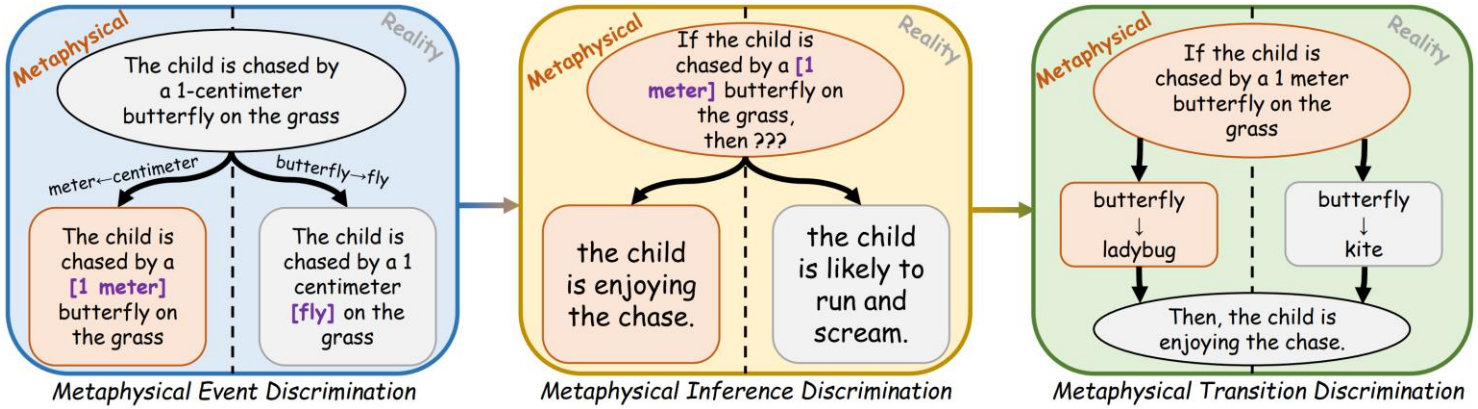
Training Data	Bleu-1	Bleu-2	Bleu-3	Bleu-4	METEOR	ROUGE-L	CIDEr	BERTScore	Human
Backbone: GPT2-XL (Radford et al., 2019) *1.5B*									
Zero-shot	4.350	1.598	0.732	0.293	5.702	5.030	0.792	37.11	14.50
ATOMIC	45.72	29.18	21.12	16.15	29.97	49.69	64.61	76.09	70.50
ATOMIC$^{20}_{20}$	42.15	25.77	17.82	13.14	29.82	47.61	63.70	70.39	76.50
ATOMIC-10X	45.38	29.20	21.09	16.15	30.09	49.86	65.02	75.89	77.50
AbstractATOMIC	45.30	29.08	21.00	16.06	29.98	48.61	63.98	75.56	71.50
CANDLE Distilled	**50.71**	**33.85**	**25.55**	**20.43**	**32.45**	**51.91**	**69.68**	**76.86**	78.50
Backbone: ChatGPT (OpenAI, 2022) (`openai/gpt-3.5-turbo`)									
Zero-shot	11.82	4.258	1.891	0.926	13.87	13.73	4.350	49.28	78.50
Five-shot	26.32	12.50	7.160	4.415	18.60	24.65	8.313	58.69	**81.00**
Chain-of-thought	9.906	3.568	1.556	0.736	11.85	11.02	2.905	46.17	64.00
Backbone: LLAMA2 (Touvron et al., 2023) *7B*									
Zero-shot	18.26	7.453	3.594	1.945	15.90	20.28	8.872	48.23	48.50
Five-shot	31.22	16.87	9.767	5.989	19.74	27.67	17.83	58.41	65.50
ATOMIC	42.04	23.01	14.10	9.125	27.80	42.90	53.17	71.52	68.50
ATOMIC$^{20}_{20}$	41.07	22.46	13.62	8.619	27.74	42.42	53.28	71.77	74.00
ATOMIC-10X	42.07	23.08	14.14	9.198	28.14	42.75	53.69	71.93	76.50
AbstractATOMIC	42.78	23.64	14.58	9.471	27.74	42.55	53.12	71.51	71.00
CANDLE Distilled	43.86	24.40	15.12	10.00	28.36	43.86	54.25	72.94	79.50

Table 3: 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**.

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*	**80.97**$_{\uparrow 1.42}$	**81.14**$_{\uparrow 1.75}$	**83.64**$_{\uparrow 0.76}$	**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$_{\uparrow 2.36}$

Metaphysical Reasoning with LLMs





针对大模型在面对未知情境和创造性上的局限性，我基于System-2 Reasoning的思路进一步提出了**形而上学推理**（Metaphysical Reasoning）的全新问题定义，用于刻画大语言模型**在面对环境或他人行动引发的情境变化时的推理能力**。

我构建了首个专门的评测基准 MARS，包含三个子任务，全面检验了大语言模型在**可行性判断、因果推理和情境修复**关键环节的推理能力，为评估和提升模型的泛化推理提供了系统化框架。

Metaphysical Reasoning with LLMs

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

通过对 20 多个语言模型的系统评测，我们发现当前模型在该任务上普遍表现不佳，即便经过微调也难以达到理想效果。

同时，实验表明结合概念化知识库（如 CANDLE）进行迁移学习能够显著提升模型在 MARS 上的表现，揭示了抽象概念知识对增强大模型在位置情景下推理能力泛化的重要性。

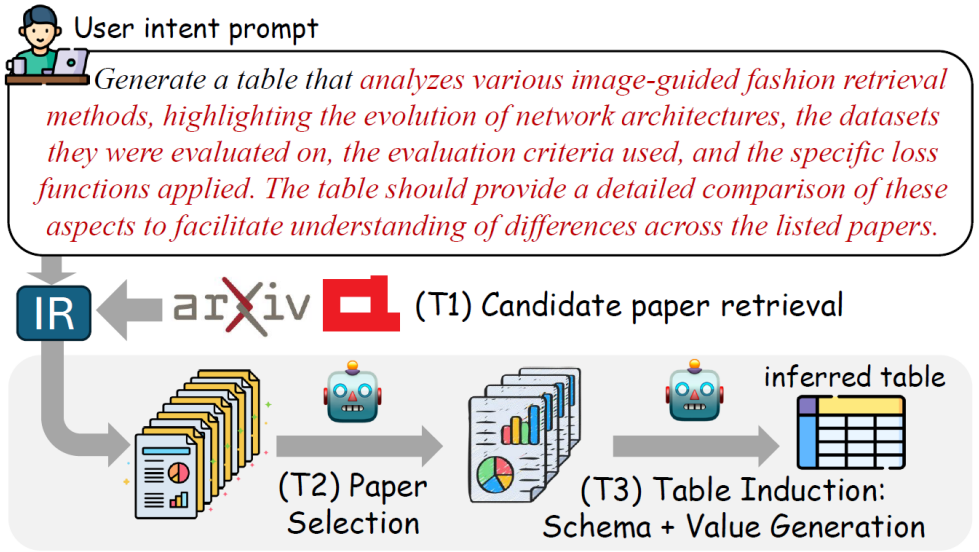
LLM for Agentic Scientific Table Generation

在JHU访问期间，我针对科学论文综述表格生成任务，提出了更贴近真实场景的新定义和解决方案：

1. **任务与基准改进**：构建了arXiv2Table基准，引入干扰文献和抽象化用户需求，让大模型像agent一样，**需要在嘈杂环境中自主筛选与决策，而不是依赖预设的干净输入**。

2. **新的评价框架**：提出了一个无需人工标注的 LLM 生成问答评估方法，从表格的schema、单元格信息和单元格间关系三个维度全面衡量表格质量。

3. **生成方法创新**：提出了迭代式批处理生成框架，**模拟agent在多轮交互中不断检索、筛选和优化schema的过程**，使表格生成更稳健、更符合用户需求。

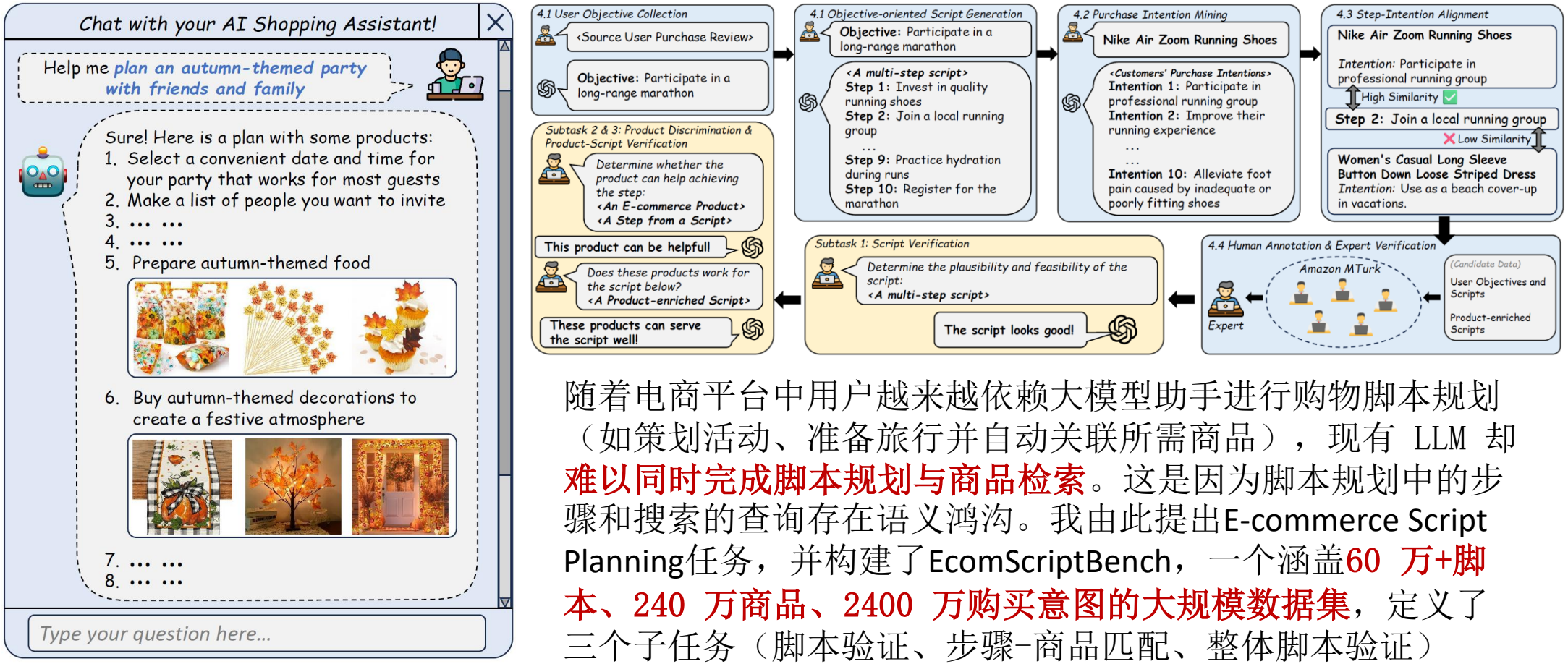


LLM for Agentic Scientific Table Generation

Backbone Model	Method	Paper	Schema			Unary Value			Pairwise Value			Avg
		Recall	P	R	F1	P	R	F1	P	R	F1	
LLAMA-3.3 (70B)	Baseline 1	52.8	31.3	37.7	34.2	29.6	40.4	34.2	28.4	31.8	30.0	32.8
	Baseline 2	65.4	26.7	**69.3**	38.5	17.0	56.8	26.2	11.2	22.5	15.0	26.6
	Newman et al.	61.9	36.4	40.5	38.3	32.8	44.5	37.8	29.5	30.2	29.8	35.3
	Ours	69.3	41.9	55.4	47.7	43.1	62.6	51.1	36.4	46.9	41.0	46.6
Mistral-Large (123B)	Baseline 1	54.7	33.1	34.5	33.8	31.6	30.4	31.0	15.5	24.7	19.0	27.9
	Baseline 2	66.8	27.4	65.0	38.5	22.7	47.4	30.7	17.8	30.7	22.6	30.6
	Newman et al.	67.9	39.9	41.6	40.7	34.7	46.3	39.7	29.9	35.1	32.3	37.6
	Ours	71.3	45.4	56.7	50.4	43.3	61.5	50.8	42.0	49.2	45.3	48.8
DeepSeek-V3 (685B)	Baseline 1	57.5	38.7	41.7	40.1	32.5	43.8	37.3	28.7	31.8	30.1	35.8
	Baseline 2	69.8	34.9	69.0	46.4	27.1	55.5	36.4	25.7	32.7	28.8	37.2
	Newman et al.	70.9	39.4	44.2	41.7	36.6	49.2	42.0	33.3	36.5	34.8	39.5
	Ours	74.3	39.6	56.9	46.7	47.7	65.2	55.1	40.4	49.8	44.6	48.8
GPT-4o-mini	Baseline 1	55.9	32.0	35.7	33.7	28.9	39.3	33.3	25.0	31.0	27.7	31.6
	Baseline 2	68.2	31.5	67.7	43.0	27.7	50.8	35.9	21.6	28.3	24.5	34.5
	Newman et al.	69.3	40.3	45.9	42.9	38.3	47.5	42.4	35.0	37.8	36.3	40.5
	Ours	72.6	46.5	59.7	52.3	**49.0**	**66.7**	**56.5**	43.5	51.9	47.3	52.0
GPT-4o	Baseline 1	58.5	35.8	43.2	39.2	36.9	41.8	39.2	29.0	34.7	31.6	36.7
	Baseline 2	70.2	34.2	68.0	45.5	27.9	56.0	37.2	19.4	33.6	24.6	35.8
	Newman et al.	71.3	45.0	47.9	46.4	38.7	49.8	43.6	36.9	40.0	38.4	42.8
	Ours	**74.6**	**51.5**	59.4	**55.2**	46.1	**66.7**	54.5	**45.9**	**55.7**	**50.3**	**53.3**

实验结果表明，现有大模型在该任务上仍存在显著挑战，而我们提出的方法在各个子任务的多个指标上均实现了显著提升。这不仅体现了任务定义与方法设计在真实科研场景中的实用价值，也表明我们的方法能够有效增强大模型作为科研信息Agent在复杂环境下的筛选、归纳与知识组织能力。

E-commerce Script Planning with LLMs



随着电商平台中用户越来越依赖大模型助手进行购物脚本规划（如策划活动、准备旅行并自动关联所需商品），现有 LLM 却**难以同时完成脚本规划与商品检索**。这是因为脚本规划中的步骤和搜索的查询存在语义鸿沟。我由此提出E-commerce Script Planning任务，并构建了EcomScriptBench，一个涵盖**60 万+脚本、240 万商品、2400 万购买意图的大规模数据集**，定义了三个子任务（脚本验证、步骤-商品匹配、整体脚本验证）

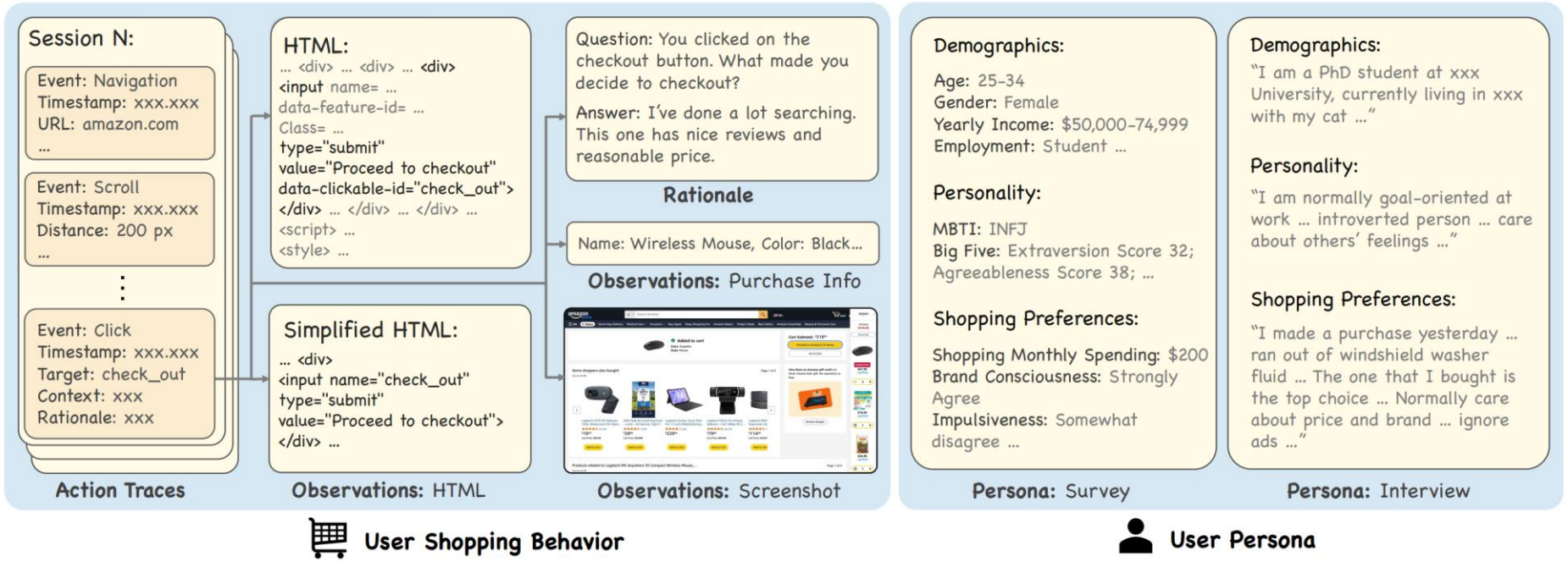
E-commerce Script Planning with LLMs

Methods	Backbone	Script Verification			Product Discrimination			Product-Script Veri.		
		Acc	AUC	Ma-F1	Acc	AUC	Ma-F1	Acc	AUC	Ma-F1
Random	N/A	50.00	-	50.00	50.00	-	50.00	50.00	-	50.00
Majority	N/A	60.98	-	60.05	57.67	-	57.10	56.46	-	56.24
PTLM *(Zero-shot)*	RoBERTa-Large *340M*	52.04	51.79	51.21	50.80	50.74	50.68	51.39	51.37	51.32
	DeBERTa-Large *435M*	51.98	52.06	51.82	52.00	51.96	51.23	52.34	52.59	51.81
	CAR *435M*	52.77	52.75	51.95	51.98	52.10	51.88	53.06	53.25	52.90
	CANDLE *435M*	53.76	53.61	53.20	52.89	53.10	52.28	52.40	52.37	51.91
	VERA-xl *3B*	53.63	53.50	53.18	52.94	52.87	52.21	52.18	52.09	51.94
	VERA-xxl *11B*	55.77	55.66	54.79	54.49	54.61	53.92	54.90	54.94	54.34
LLM *(Zero-shot)*	Meta-Llama-3-8B	70.05	-	69.98	64.83	-	64.36	61.16	-	60.22
	Meta-Llama-3-70B	71.74	-	71.52	66.02	-	65.05	62.00	-	61.33
	Meta-Llama-3.1-8B	71.45	-	71.30	65.74	-	65.69	61.63	-	60.96
	Meta-Llama-3.1-70B	72.65	-	72.42	66.15	-	65.54	62.50	-	62.22
	Meta-Llama-3.1-405B	75.26	-	74.97	68.16	-	67.33	65.66	-	65.65
	Gemma-2-2B	66.82	-	66.80	60.56	-	60.22	58.95	-	58.10
	Gemma-2-9B	71.27	-	70.98	65.14	-	64.15	61.07	-	60.40
	Gemma-2-27B	71.77	-	71.27	66.86	-	66.20	63.15	-	62.70
	Phi-3.5-mini *4B*	68.18	-	68.05	61.92	-	61.15	60.36	-	59.79
	Falcon2 *11B*	71.73	-	71.68	65.70	-	65.12	61.89	-	61.65
	Mistral-7B-v0.3	72.38	-	71.49	66.42	-	65.77	62.18	-	61.47
	Mistral-Nemo *12B*	73.18	-	72.51	66.98	-	66.78	62.95	-	62.71
	Mixtral-8x7B-v0.1	75.06	-	74.25	66.39	-	65.59	63.64	-	62.84
PTLM & LLM *(Fine-tuned)*	RoBERTa-Large *340M*	79.18	79.27	78.86	72.26	72.32	71.74	70.26	70.38	69.83
	DeBERTa-v3-Large *435M*	81.10	80.76	81.03	74.26	74.56	73.78	72.00	71.93	71.99
	Meta-LLaMa-3-8B	83.48	83.38	82.64	75.75	75.52	**75.73**	73.06	73.33	72.84
	Meta-LLaMa-3.1-8B	85.24	85.07	84.64	**76.44**	**76.51**	75.53	**74.48**	**74.44**	**74.38**
	Gemma-2-2B	81.06	80.95	80.82	73.43	73.51	73.09	69.61	69.79	68.78
	Gemma-2-9B	82.04	82.20	81.35	73.58	73.94	73.15	71.65	71.41	71.44
	Mistral-7B-v0.3	**85.72**	**85.61**	**85.51**	75.63	75.61	75.33	73.18	73.09	72.62
LLM *(API)*	GPT4o-mini	74.30	-	73.54	69.03	-	68.47	69.68	-	69.16
	GPT4o-mini (5-shots)	74.56	-	73.61	71.56	-	71.09	71.39	-	71.04
	GPT4o-mini (COT)	71.66	-	71.59	69.31	-	68.63	70.62	-	70.23
	GPT4o-mini (SC-COT)	72.74	-	72.38	71.13	-	70.79	70.93	-	70.26
	GPT4o-mini (SR)	73.32	-	72.35	72.46	-	71.89	71.08	-	70.43
	GPT4o	77.50	-	77.23	73.04	-	72.06	71.50	-	71.33
	GPT4o (5-shots)	77.92	-	76.93	73.90	-	73.68	72.85	-	72.83
	GPT4o (COT)	74.89	-	74.12	71.05	-	70.58	70.32	-	69.68
	GPT4o (SC-COT)	73.84	-	73.16	71.08	-	70.67	69.26	-	68.67
	GPT4o (SR)	76.22	-	76.13	71.97	-	71.28	71.90	-	70.96

Backbone	Training Data	Script Verification			Product Discrimination			Product-Script Veri.		
		Acc	AUC	Ma-F1	Acc	AUC	Ma-F1	Acc	AUC	Ma-F1
Llama-3.1 *8B*	Zero-shot	71.45	-	71.30	65.74	-	65.69	61.63	-	60.96
	ECOMSCRIPTBENCH	83.86	83.94	83.05	77.70	77.87	77.59	75.88	75.58	75.58
	FolkScope + MIND	67.74	67.63	67.38	66.79	66.43	66.11	64.91	64.87	64.42
	+ ECOMSCRIPTBENCH	84.65	84.84	84.13	78.60	78.83	78.27	76.35	76.50	76.08
Mistral-v0.3 *7B*	Zero-shot	72.38	-	71.49	66.42	-	65.77	62.18	-	61.47
	ECOMSCRIPTBENCH	85.72	85.61	85.51	75.63	75.61	75.33	73.18	73.09	72.62
	FolkScope + MIND	69.77	70.00	69.56	67.78	67.75	67.39	63.70	63.41	63.66
	+ ECOMSCRIPTBENCH	**85.87**	**85.80**	**86.37**	**81.18**	**80.96**	**80.54**	**78.94**	**78.94**	**78.66**

实验发现：现有模型即便经过微调仍面临显著挑战，但引入购买意图知识能显著提升性能，说明了意图建模在电商智能体中的核心价值。数据被引入为Amazon Rufus LLM内部评测指标之一。

[Ongoing Production Work @Amazon]
Customer Agentic Simulation with LLMs



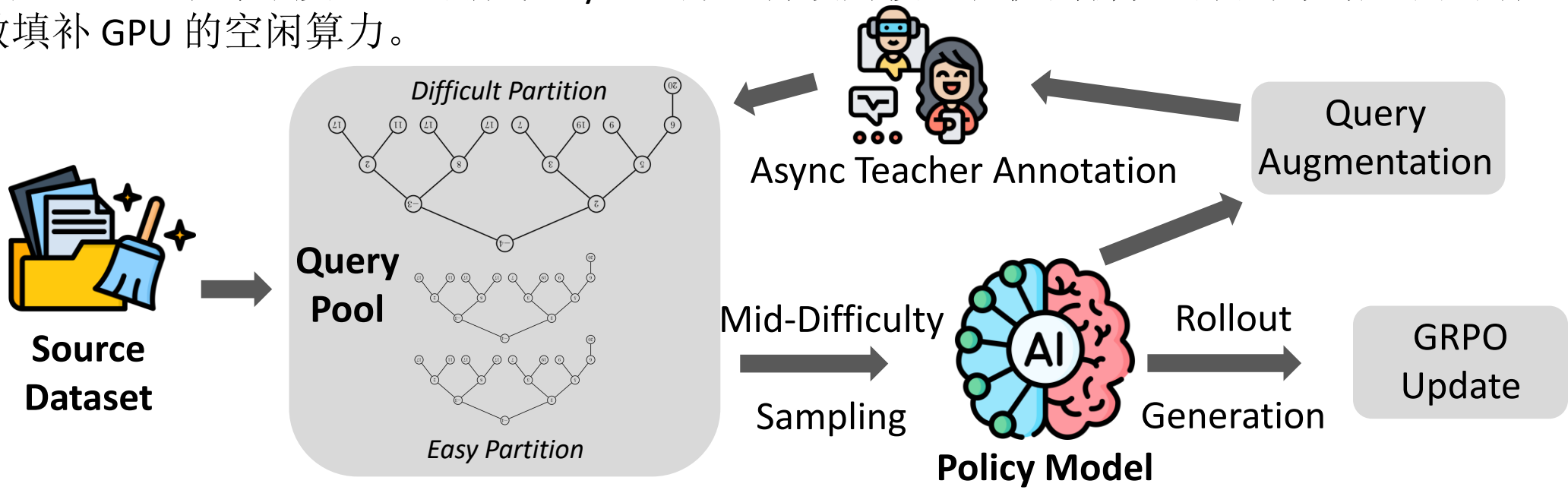
在我们的研究中，我们希望让**大语言模型作为顾客智能体**，自动参与类似 A/B 测试的实验。例如，测试一个推荐系统新上线的feature，可以将一部分Agent分配到原系统，另一部分分配到新系统，并让它们在模拟购物或交互中真实体验两种设计。随后，我们测量和比较两组智能体的行为与反馈，以此评估哪种方案更能提升客户体验和带来更大的购买率。这种方式不仅借鉴了亚马逊 **Weblab** 的实验范式，还通过 LLM 智能体大规模、自动化地探索产品与交互改进的潜力。

Wang, Z., Lu, Y., Li, W., Amini, A., Sun, B., Bart, Y., ... & Wang, D. (2025). OPeRA: A Dataset of Observation, Persona, Rationale, and Action for Evaluating LLMs on Human Online Shopping Behavior Simulation. *arXiv preprint arXiv:2506.05606*.

[Ongoing Research Work @Amazon]
Breaking the RL Scaling Plateau

在大模型的强化学习训练中，随着数据量增加，性能提升往往停滞（Scaling Plateau），原因是采样的轨迹多数过于简单或过于困难，无法提供有效学习信号基于以上动机。

我们提出了一个**动态数据生成框架**，能够在训练过程中**实时评估任务难度**，并动态调度或生成处于“中等难度区”的新Query。该方法持续为模型提供高价值的训练数据，同时有效填补 GPU 的空闲算力。



[Ongoing Research Work @Amazon] Breaking the RL Scaling Plateau

1. 动态数据池:

- 我们设计了由三个Min Heap构成的动态数据池，按照 reward 将数据分为平衡的**难 / 易**两部分。训练过程中的每个step从“中间难度区”采样，即难的一半和易的一半交界区域，确保训练样本始终处于模型的 **最近发展区 (Zone of Proximal Development)**。
- 使用一个堆在难的部分和两个堆在简单的部分保证了 **插入 / 取出操作均为 O(logN)**，在大规模分布式训练中保持高效。

2. On-policy Query Augmentation

由 **Policy Model 自身生成**与已有高价值数据分布相似的新 Query，并根据 reward 打分确定难度。将这些 Query 异步交给更强的大模型（当前是 **GPT-5**）做 **Ground Truth 标注**，提供额外的强监督信号。(与传统数据蒸馏不同，该过程并不是让大模型生成数据，而是提供 **更可靠答案**，增强训练数据质量)

3. 效果与对比

相比 **DAPO** 基线方法，我们的方法能显著：

- **加快训练收敛速度**：更快达到同等性能。
- **提升数据利用率**：更少数据即可带来更多增益。
- **保持高资源利用率**：不额外增加过重的GPU计算负担，达到同等表现的Flops没有显著增加。