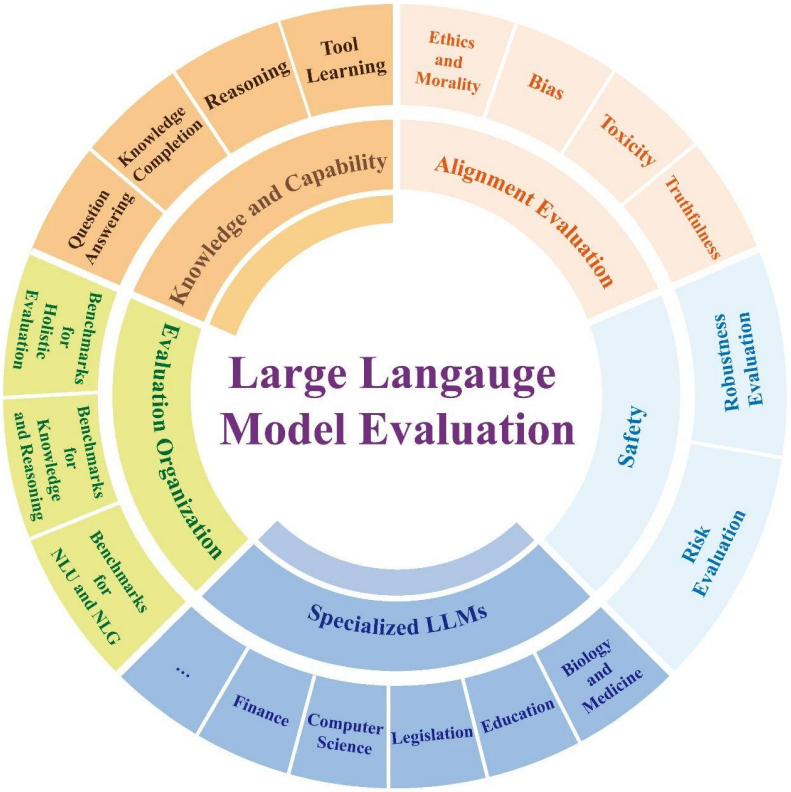
Evaluating Large Language Models on Knowledge and Capability

宁钰成

2024.09.27

Outline

- 大模型知识与能力评估-概述
  - （1）用什么评估？（任务、数据集）
  - （2）怎么评估？（评测指标）
- 大模型知识与能力评估-论文分享
  - （1）大模型综合评估
  - （2）大模型相关应用的评估（RAG & Agent）
  - （3）大模型特定任务 or 特定能力的评估
- 总结与展望
  - （1）未来可能的研究方向
  - （2）国内研究团队



大模型评估概述：通常需要回答的两个问题

- 用什么评估？
  - （1）需要评估模型的哪个方面（能力或知识等）？
  - （2）该方面包含哪些任务？
  - （3）与任务相对应的数据集？

- 怎么评估？
  - （1）自动评估：相应的评测指标（基于静态数据集）
  - （2）人工评估
  - （3）使用GPT-4等较强的模型进行评估（打分、排序等）

1. 用什么评估：从 NLP 任务到人类试题

- 传统 NLP 评估模型的方式：

  做什么任务，就在对应的数据集上进行评估

- 随着预训练模型的发展（BERT时代），单个的数据集已经逐渐不足以综合评估一个模型。那么，我们就拿多个NLP 任务数据集组成一个测试基准 benchmark 来测量，这些 NLP 任务往往代表着模型在解决某类特定问题上的表现。一些比较常见的测试基准有：

  GLUE、SuperGLUE、XTREME、 BIG-Bench 等

- 在ChatGPT/GPT-4公开后，评测大模型的方法逐渐从使用 NLP 任务过渡到使用人类试题。一方面，传统的 NLP 任务过于简单；另一方面，数据集在互联网上都可以轻易获取，从而容易混入大模型的预训练语料里，导致其上评估的结果没有那么有说服力（数据污染）。在GPT-4的官方介绍中，我们可以看到两个主要测试角度：

  人类考试+传统测试基准，例如AGIEval（下一页）、MMLU、C-Eval等，主要是选择题

AGIEval: 以人为中心的基础模型评估基准，评估大模型在人类认知水平任务上的表现

Exams	#Participants	Language	Tasks	Subject	# Instance	#Avg. Token
Gaokao 高考	12M	Chinese	GK-geography	Geography	199	144
			GK-biology	Biology	210	141
			GK-history	History	243	116
			GK-chemistry	Chemistry	207	113
			GK-physics	Physics	200	124
			GK-En	English	306	356
			GK-Ch	Chinese	246	935
			GK-Math-QA	Math	351	68
			GK-Math-Cloze	Math	118	60
SAT	1.7M	English	SAT-En.	English	206	656
			SAT-Math	Math	220	54
律师资格考试 Lawyer Qualification Test	820K	Chinese	JEC-QA-KD	Law	1000	146
			JEC-QA-CA	Law	1000	213
Law School Admission Test (LSAT)	170K	English	LSAT-AR	Law-Analytics	230	154
			LSAT-LR	Law-Logic	510	178
			LSAT-RC	Law-Reading	260	581
国家公务员考试 Civil Service Examination	2M	English	LogiQA-en	Logic	651	144
	2M	Chinese	LogiQA-ch	Logic	651	242
GRE	340K	English	AQuA-RAT	Math	254	77
GMAT	150K	English				
AMC	300K	English	MATH	Math	1000	40
AIME	3000	English				

部分常用的LLM评测基准

名称	发布机构	语言	题型	评估内容
MMLU	UC Berkeley等	英文	选择题	57个不同领域的通用知识
ARC	Allen AI	英文	选择题	小学3年级到9年级的科学考试问题
C-Eval	上海交大、清华	中文	选择题	52个不同领域
MT-bench	UC Berkeley等	英文	开放式问答	8个不同领域
HellaSwag	Stanford	英文	开放式问答	评估模型生成符合上下文的文本延续的能力
GSM8K	OpenAI	英文	开放式问答	8.5k个小学数学问题

更多评测基准介绍：汇总大语言模型LLM的评测基准数据集(BenchMarks)

2. 怎么评估：如何给模型输出打分

- 无论是传统的 NLP 任务，或者多选题形式的人类考题，都有相对明确的打分标准，例如准确率 accuracy，或者专门的衡量指标（基于静态数据集进行打分）：

  分类任务：Accuracy, Precision, Recall, F1-Score, PR曲线

  回归任务：MAE(平均绝对误差), MSE(均方误差), MSLE(均方误差对数).....

  语言模型：Cross-entropy, Perplexity.....

  文本生成：BLEU, ROUGE, METEOR, BERTScore.....

- 对于更丰富一些的任务，例如对一篇文章的质量进行评估，需要人类打分。此外，许多验证评价指标有效的方法也是计算该指标与人类评估的契合度。

- 随着 LLM 的发展，用模型来给模型打分也逐渐流行起来。对于不同模型的输出，用一个（相对来说比较强的）模型，例如 使用GPT-4 对不同模型的输出进行比较和打分。但是值得注意的是，最近的一些工作也发现模型做自动评估往往还是存在一些系统性的偏见。

分享论文目录

大模型综合评估

- 精确评估大语言模型的世界知识

  KoLA: Carefully Benchmarking World Knowledge of Large Language Models (Yu et al., ICLR 2024)

- 基于知识的交互式评估方法

  KIEval: A Knowledge-grounded Interactive Evaluation Framework for Large Language Models (Yu et al., ACL 2024)

大模型相关应用的评估（RAG & Agent）

- 检索增强生成（RAG）系统的自动评估

  RAGAs: Automated Evaluation of Retrieval Augmented Generation (Es et al., EACL 2024)

- 特定场景下检索增强生成（RAG）评估数据生成框架

  RAGEval: Scenario Specific RAG Evaluation Dataset Generation Framework, *ArXiv*, *abs*/2408.01262.

- 大模型工具使用能力评测（Agent 能力）

  T-Eval: Evaluating the Tool Utilization Capability of Large Language Models Step by Step (Chen et al., ACL 2024)

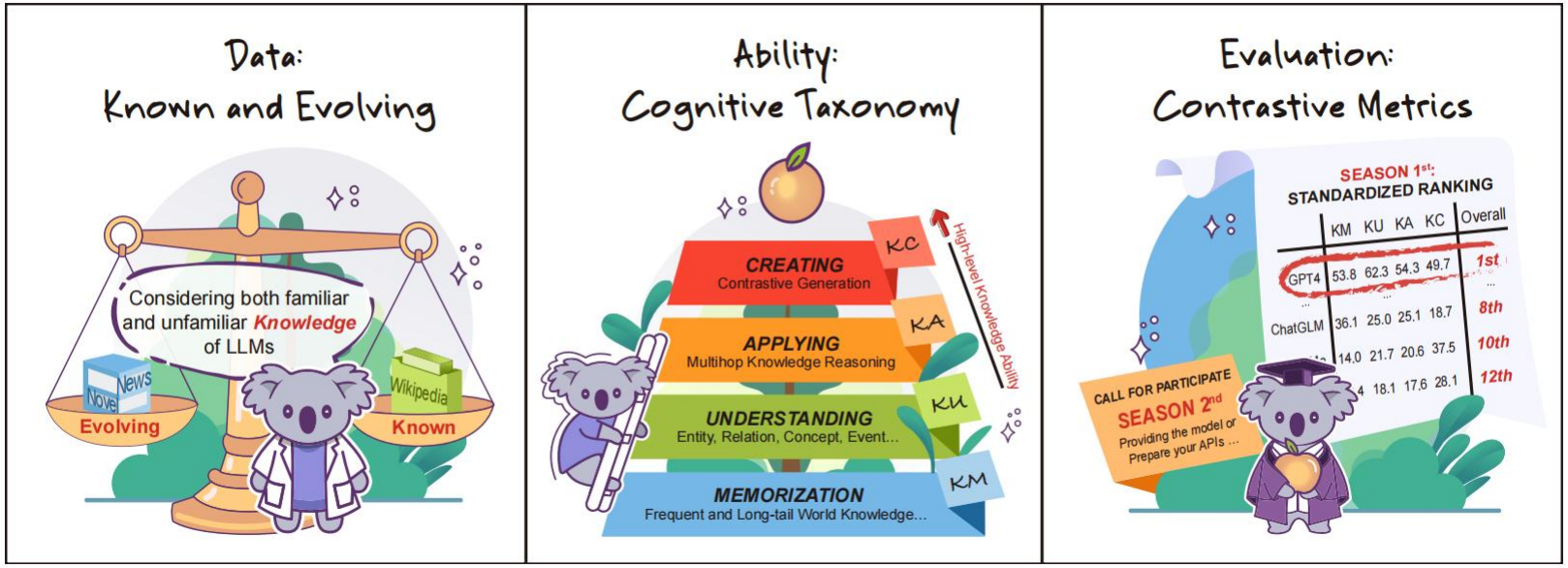
大模型特定任务or能力的评估

- 评估大语言模型在处理有争议知识的问题回答能力

  DEBATEQA: Evaluating Question Answering on Debatable Knowledge, *ArXiv*, *abs*/2408.01419.

KoLA: Carefully Benchmarking World Knowledge of Large Language Models

考拉：一种以世界知识为导向的LLM评估基准



Preliminary

世界知识：一种常见说法，指概念、实体和事件等知识，被广泛认为在LLM的出色性能中起着基础性的作用。偏向于Know-What

- 概念(concept)：“人类”、“动物”
- 实体(entity)：一棵树、一辆汽车、一座城市
- 事件(event)：“地震”、“金融危机”

还有哪些知识？

- 原理知识(Know-Why)：牛顿运动三大定律、行星运动规律
- 技能知识(Know-How)：如何做一道“鱼香肉丝”
- 人际知识(Know-Who)：哪位师兄师姐特别适合帮你debug

Motivation

- **Ability-更有价值的评估结果**

传统的评估集中在相对狭窄和表面的能力上，已经有些过时；最近的一些研究希望扩展评估范围，以涵盖更广泛的能力或更具挑战性的任务，追求评估的“广度”。

但作者认为，评估的“深度”也同样重要，主张考虑能力的分层和联系，而不是将这些评估数据集仅仅基于学科或难度离散地组织起来。

- **Data-评估的公平性**

LLM使用的预训练数据范围非常广泛。在评估LLM时，训练数据的变化和潜在的测试数据泄露风险会带来公平性问题。

- **Evaluation-为了具有高度的适用性**

传统的基准测试单独报告不同任务的绝对指标，不同任务的得分之间具有不可比性；此外不同指标的敏感性不同，可读性差。

KoLA: Carefully Benchmarking World Knowledge of Large Language Models (Yu et al., ICLR 2024)

Method

如何评估人→如何评估LLM

- **知识记忆(Knowledge Memorization, KM)**

  问题：汪国真的职业是什么？

  答案：诗人

- **知识理解(Knowledge Understanding, KU)**

  问题：“物种具有纤维素壁和其他多糖。”这句话是真的还是假的？请回答真假。

  答案：假

- **知识应用(Knowledge Applying, KA)**

  问题：《盲井》和《傅满洲的面具》哪部电影先上映？

  答案：《盲井》是 2003 年的电影；《傅满洲的面具》是1932 年的电影。所以答案是《傅满洲的面具》。

- **知识创造(Knowledge Creating, KC)**

  说明：完成下面的文字，确保包含所提供的所有事件。

  事件：### 标题： 弗雷迪-格雷之死;###已知事件;#### 事件触发器：指控;患者：六名警察;原因：法医报告裁定格雷之死为他杀;...给定背景： ### 2015年4月12日，25岁的黑人男子弗雷迪-卡洛斯-格雷...

  参考回答：2015年5月1日，巴尔的摩市州检察官玛丽莲-莫斯比宣布，在法医报告判定格雷死亡为他杀后，她的办公室已对六名警察提出指控。检察官表示，他们有合理的理由对六名警察提起指控，据信这六名警察与格雷的死有关。

Method 用什么评估：数据源和任务选择

Level	ID	Dataset	Metrics	Exclusive	Context Type	Test Set	Pool	Source
KM	1-1	High-Freq.	EM, **F1**	✓	Triple	100	20.6M	Known
	1-2	Low-Freq.	EM, **F1**	✓	Triple	100	20.6M	
	1-3	ETM	EM, **F1**	✓	Triple	100	2.7k	Evolving
KU	2-1	COPEN-CSJ	**Acc.**	✓	Entity, Concept	100	3.9k	Known
	2-2	COPEN-CPJ	**Acc.**	✓	Concept	100	4.7k	
	2-3	COPEN-CiC	**Acc.**	✓	Concept	100	2.3k	
	2-4	FewNERD	**F1**	×	Sentence	300	188.2k	
	2-5	DocRED	**F1**	✓	Document, Entity	100	12k	
	2-6	MAVEN	**F1**	✓	Document	100	20.4k	
	2-7	MAVEN-ERE	**F1**	✓	Document(s), Event	199	1.3M	
	2-8	ETU	**F1**	✓	Document, Entity	100	1.6k	Evolving
KA	3-1	HotpotQA	**F1**	×	Document(s)	100	7.4k	Known
	3-2	2WikiMulti.	**F1**	✓	Document(s)	100	12.6k	
	3-3	MuSiQue	**F1**	✓	Document(s)	100	2.5k	
	3-4	KQA Pro	**F1**	✓	KG	100	1.2k	
	3-5	KoRC	**F1**	✓	Document(s), KG	100	5.2k	
	3-6	ETA	**F1**	✓	Document(s), KG	49	1.6k	Evolving
KC	4-1	Encyclopedic	BLEU, **Rouge**	✓	Document, Event	95	4.5k	Known
	4-2	ETC	BLEU, **Rouge**	✓	Document, Event	95	100	Evolving

概念探测
命名实体识别
关系抽取
事件关系抽取
世界知识&多跳推理
读后续写

固定数据："经典例题"
(Known Data Source)

选用2021年前的维基百科数据，几乎所有大模型都会使用他们进行预训练。

不断演化的数据：
"大家都没见过的原创题"
(Evolving Data Source)

每季度都要更新，持续获取最近90天内发布的网络内容（事实新闻和虚构小说）作为数据源，并在其上构建新的数据集。

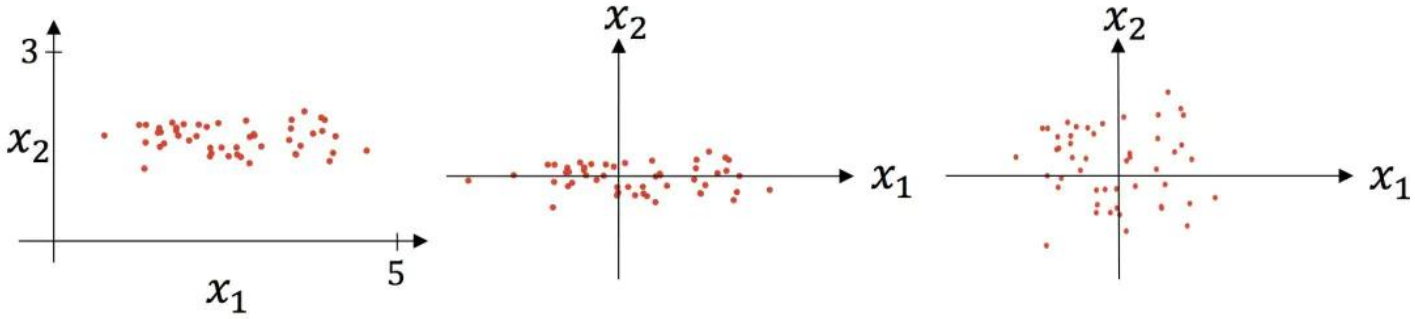
KoLA: Carefully Benchmarking World Knowledge of Large Language Models (Yu et al., ICLR 2024)

Method 怎么评估：在已有指标的基础上，设计对比评估系统

- 标准化整体得分

$$z_{ij} = \frac{x_{ij} - \mu\left(x_{i1}, \ldots, x_{i|M|}\right)}{\sigma\left(x_{i1}, \ldots, x_{i|M|}\right)}$$

score x_{ij} of model m_j on task d_i



将数据变换为均值为0，标准差为1的分布（保持原分布类型）

$$s_{ij} = 100\frac{z_{ij} - \min\left(z\right)}{\max\left(z\right) - \min\left(z\right)}$$

将所有结果调整到[0, 100]的范围，是一个相对得分。这使得同模型不同任务、甚至不同模型不同任务间的得分可比。

KoLA: Carefully Benchmarking World Knowledge of Large Language Models (Yu et al., ICLR 2024)

Experiment

Model	Level 1: KM				Level 2: KU									Level 3: KA							Level 4: KC		
	1-1	1-2	1-3	Rank	2-1	2-2	2-3	2-4	2-5	2-6	2-7	2-8	Rank	3-1	3-2	3-3	3-4	3-5	3-6	Rank	4-1	4-2	Rank
GPT-4	64.3	68.9	41.5	1st (—)	69.5	48.6	51.4	66.9	100.0	77.2	78.9	81.3	1st (—)	59.8	60.7	76.8	32.8	60.9	58.1	1st (—)	48.6	58.9	2nd (↑1)
GPT-3.5-turbo	53.4	60.0	38.3	2nd (↑2)	44.2	49.4	50.2	54.0	51.3	50.2	56.6	25.5	2nd (—)	58.5	42.6	53.9	45.6	30.9	19.6	5th (↓1)	52.2	46.4	3rd (↓1)
InstructGPT davinci v2 (175B*)	41.2	48.4	36.1	7th (↑1)	33.6	48.2	42.4	41.8	56.7	62.3	40.6	31.3	3rd (—)	30.6	39.7	44.2	23.6	50.3	23.4	7th (↓1)	54.4	56.8	1st (—)
Tulu (7B)	41.5	48.6	28.4	8th (↓2)	22.0	25.4	43.0	35.7	30.9	20.6	22.2	25.8	11th (↑1)	42.5	45.6	40.7	54.8	42.3	54.8	3rd (↑2)	32.8	42.5	9th (↑2)
Cohere-command (52.4B)	59.0	54.5	33.9	3rd (↓1)	40.0	47.1	46.3	26.6	38.6	20.6	46.9	25.0	4th (—)	36.2	41.7	45.3	49.3	54.7	44.0	4th (↓1)	17.1	15.1	25th (↓13)
FLAN-UL2 (20B)	53.0	42.6	30.7	6th (↓1)	59.0	47.1	53.0	16.0	25.0	20.6	22.2	25.0	6th (—)	49.6	47.5	39.4	51.1	43.5	53.3	2nd (—)	28.3	19.0	20th (↓3)
J2-Jumbo-Instruct (178B*)	32.6	33.7	19.2	12th (—)	27.3	23.5	31.2	37.1	32.0	32.7	51.3	25.0	7th (—)	45.2	31.6	31.6	38.3	28.3	25.5	8th (—)	43.7	53.3	4th (—)
ChatGLM (130B)	38.0	56.5	36.1	5th (↑2)	30.5	47.8	51.9	16.0	25.0	23.3	30.3	25.0	8th (—)	36.4	34.1	28.0	36.4	36.7	21.2	9th (↑1)	24.4	28.4	16th (↑2)
FLAN-T5 (11B)	56.1	51.5	32.9	4th (↓1)	63.2	47.8	49.1	18.7	—	—	—	25.0	5th (—)	45.0	49.0	32.9	51.1	39.7	16.1	6th (↑1)	20.2	0.0	28th (↓6)
InstructGPT curie v1 (6.7B*)	28.1	43.8	28.9	10th (↓1)	29.4	41.2	41.8	22.3	25.4	22.0	25.8	25.0	9th (—)	31.6	37.2	24.3	29.1	31.2	27.2	11th (—)	27.7	29.6	12th (↑3)
LLaMa (65B)	24.2	25.6	19.5	15th (↓1)	22.0	18.4	18.3	55.6	31.4	30.1	25.5	25.0	10th (↑1)	16.4	35.4	41.7	25.4	21.7	17.6	16th (↓2)	44.7	37.1	5th (—)
ChatGLM2-32k (6B)	25.3	22.8	20.1	16th (—)	22.0	40.8	20.0	19.0	26.5	20.6	22.7	25.4	17th (—)	35.7	31.1	24.0	40.1	20.7	17.3	13th (↓1)	34.5	41.5	8th (—)
Alpaca (7B)	21.5	25.3	18.2	17th (↓2)	22.0	18.4	18.8	25.3	26.4	31.4	22.2	25.0	20th (—)	15.1	19.3	20.9	18.1	45.5	50.7	12th (↑5)	35.7	41.0	7th (↑3)
Llama2-chat (7B)	21.6	19.7	17.9	22th (↓3)	25.2	18.4	24.5	37.4	32.5	20.6	27.2	25.6	14th (—)	17.5	16.8	23.4	14.4	40.5	51.7	14th (↑2)	31.7	38.1	10th (↓4)
ChatGLM (6B)	31.9	32.9	30.5	11th (—)	23.1	45.5	32.9	16.0	25.0	22.0	22.8	25.0	13th (—)	21.2	27.5	22.2	19.9	19.5	28.5	20th (—)	17.7	30.8	18th (↑6)
Vicuna (13B)	18.8	19.1	17.4	26th (—)	22.0	18.7	23.3	29.8	26.0	35.4	30.5	25.0	15th (—)	25.4	10.0	24.7	18.1	21.0	16.4	24th (↓1)	35.0	45.9	6th (↑1)
GLM (130B)	21.9	25.1	22.9	14th (↑3)	22.0	18.4	18.3	49.6	33.2	29.7	22.2	—	12th (↓2)	23.5	13.0	18.4	21.7	45.0	35.1	17th (↓4)	29.3	19.1	19th (↓3)
GPT-J (6B)	20.8	18.8	18.0	25th (↓1)	22.0	18.4	18.3	22.2	25.0	31.0	—	25.0	25th (—)	38.8	39.4	26.8	49.3	17.5	16.5	10th (↓1)	30.5	24.0	14th (↑9)
T0++ (11B)	41.9	38.4	23.9	9th (↑1)	30.5	39.2	27.8	16.0	—	—	—	25.0	16th (—)	22.4	23.0	23.8	14.4	39.7	16.1	19th (—)	18.1	3.7	27th (↓8)
Dolly-v2 (12B)	21.1	21.0	18.9	20th (↑2)	22.0	18.4	18.8	28.5	25.0	20.6	27.1	25.0	23th (—)	14.1	20.5	18.4	18.1	26.4	25.2	22th (—)	29.5	31.9	11th (↓2)
GPT-JT (6B)	19.8	18.9	19.0	24th (↑1)	22.0	18.4	18.3	19.2	25.0	36.7	—	25.0	22th (—)	31.4	38.2	23.0	32.8	18.5	17.5	15th (—)	21.2	19.5	21th (↑5)
Internlm-chat-8k (7B)	23.5	20.4	17.2	19th (↑1)	22.0	18.4	18.3	19.0	27.2	20.6	26.1	25.0	27th (—)	19.3	20.8	22.8	14.4	17.1	22.2	23th (↑1)	26.1	27.8	15th (↑5)
UL2 (20B)	26.5	28.1	18.9	13th (—)	22.0	18.4	18.3	18.0	—	—	—	25.0	28th (—)	25.2	28.0	25.9	38.3	16.1	16.1	18th (—)	26.9	9.0	23th (↓9)
GPT-3 davinci v1 (175B)	18.1	17.9	16.9	27th (—)	22.0	18.7	18.3	30.2	25.0	29.6	22.3	25.0	19th (—)	18.3	13.5	22.8	19.9	21.5	17.0	25th (—)	32.3	22.6	13th (—)
GPT-NeoX (20B)	19.9	20.7	18.2	23th (—)	22.0	18.4	18.3	25.7	25.0	32.1	—	25.0	21th (—)	14.0	13.9	17.1	18.1	22.7	16.7	28th (↓1)	32.2	19.2	17th (↑4)
BLOOM (7B)	21.0	21.9	18.3	18th (—)	22.0	18.4	18.3	30.1	28.0	29.2	22.2	25.0	18th (—)	18.6	22.6	17.1	19.9	27.4	23.3	21th (—)	17.9	21.4	22th (↑5)
GPT-3 curie v1 (6.7B)	17.2	17.7	16.8	28th (—)	22.0	18.4	18.3	21.5	25.0	25.6	23.9	25.0	26th (—)	22.6	15.2	20.5	18.1	19.1	16.4	26th (—)	23.7	10.1	24th (↑1)
RedPajama-Instruct (7B)	21.8	21.2	16.4	21th (—)	22.0	18.4	18.3	29.8	25.0	20.6	25.4	25.0	24th (—)	12.6	10.0	17.1	14.4	26.1	23.2	27th (↑1)	13.7	12.3	26th (↑2)

KoLA: Carefully Benchmarking World Knowledge of Large Language Models (Yu et al., ICLR 2024)

Method 怎么评估：在已有指标的基础上，设计对比评估系统

- **自对比度量（Self-contrast Metric）**

  评估知识创造更重要的是评估生成的知识是否忠实和合理，即知识幻觉

这样的计算方式消除了LLM和人类作者之间的风格差异的影响

防止模型忽略知识K而导致的评估崩溃

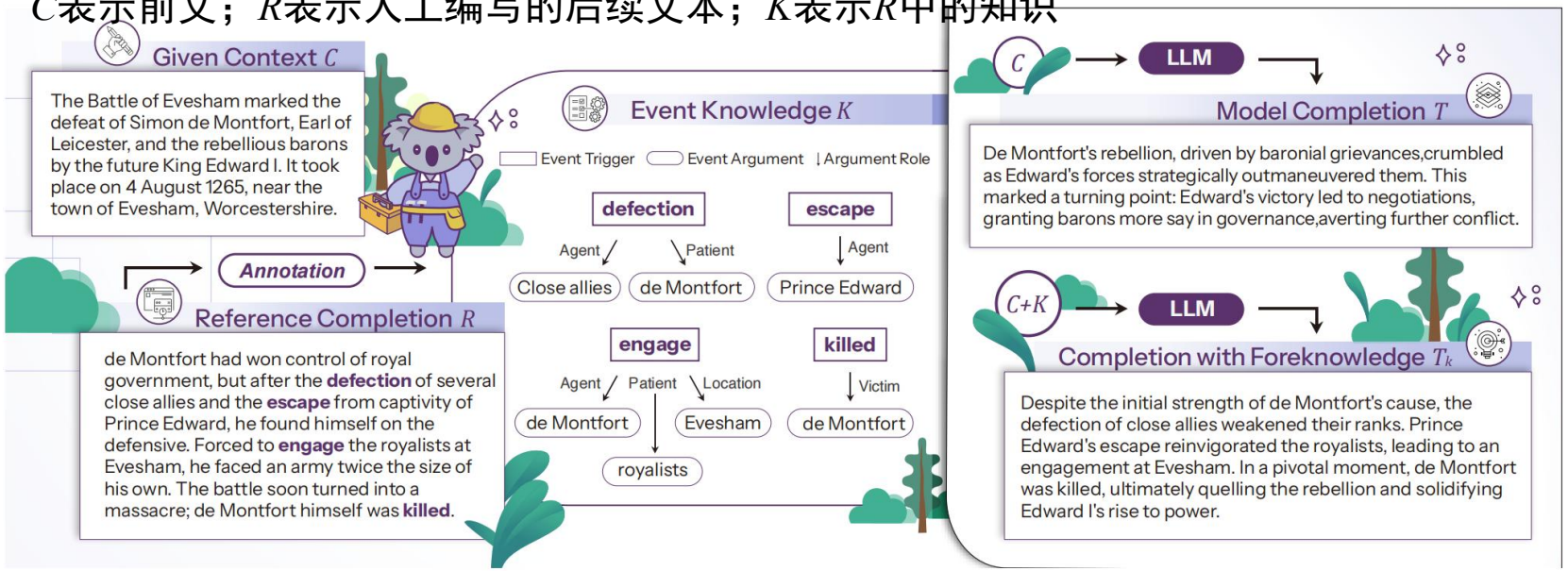
$$x = \text{avg}(\quad\quad \partial(T, T_k) \quad\quad)$$

$$x = \text{avg}(\partial(T, R), \partial(T, T_k), \partial(T_k, R))$$

avg(·)表示平均值；

∂(·)用于计算两个文本的相似度，使用Rouge-L (F1)

C表示前文；R表示人工编写的后续文本；K表示R中的知识



KoLA: Carefully Benchmarking World Knowledge of Large Language Models (Yu et al., ICLR 2024)

Experiment

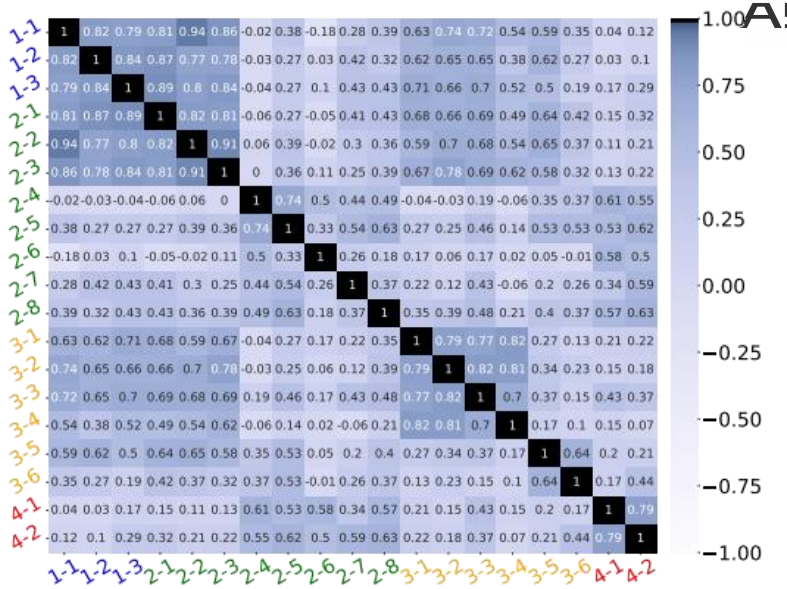
能力层级间的相关性（上图）

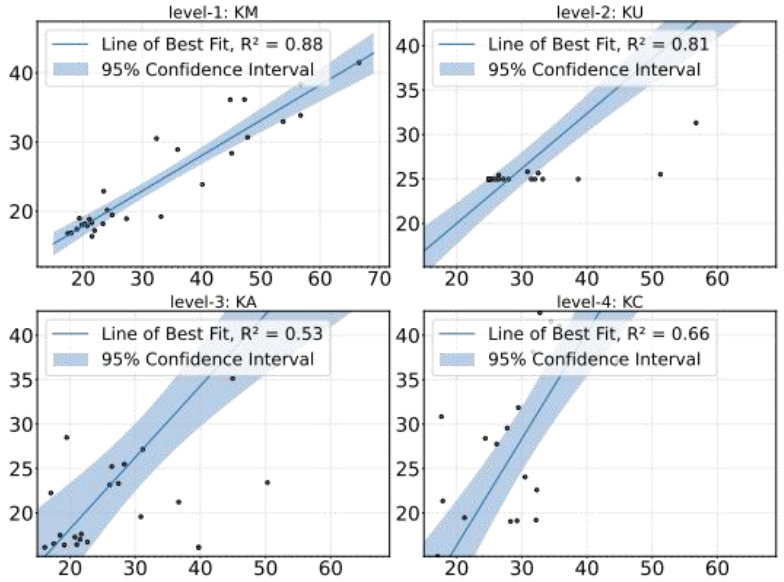
① 每个层级内的任务之间存在很高的相关性，表明LLM的能力确实具有一定的层次结构;

② 知识记忆(KM)层级与其他层级之间存在显著的相关性，表明高级任务在很大程度上依赖于低级任务;

③ 没有进行对齐或微调的模型，在知识记忆(KM)层级的排名和模型大小之间存在很强的相关性;

④ 经过对齐或微调的模型，高级能力与模型大小之间的相关性显著增加；然而，低级能力(KM)与模型大小之间的相关性呈下降趋势——“对齐税”。





证明数据集是可靠的（下图）

① 模型在不断演化和非演化任务上的结果显示明显的线性相关性，表明构建的演化数据集可靠。

KoLA: Carefully Benchmarking World Knowledge of Large Language Models (Yu et al., ICLR 2024)

分享论文目录

大模型综合评估

- 精确评估大语言模型的世界知识

  KoLA: Carefully Benchmarking World Knowledge of Large Language Models (Yu et al., ICLR 2024)

- 基于知识的交互式评估方法

  KIEval: A Knowledge-grounded Interactive Evaluation Framework for Large Language Models (Yu et al., ACL 2024)

大模型相关应用的评估（RAG & Agent）

- 检索增强生成（RAG）系统的自动评估

  RAGAs: Automated Evaluation of Retrieval Augmented Generation (Es et al., EACL 2024)

- 特定场景下检索增强生成（RAG）评估数据生成框架

  RAGEval: Scenario Specific RAG Evaluation Dataset Generation Framework, *ArXiv*, *abs*/2408.01262.

- 大模型工具使用能力评测（Agent 能力）

  T-Eval: Evaluating the Tool Utilization Capability of Large Language Models Step by Step (Chen et al., ACL 2024)

大模型特定任务or能力的评估

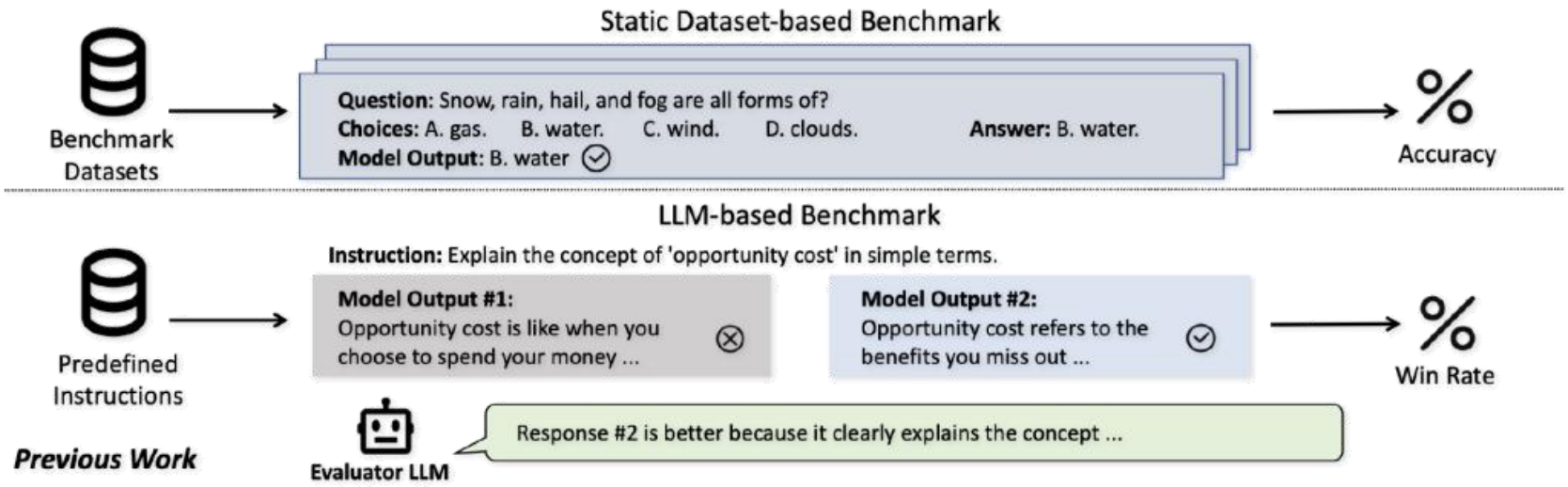
- 评估大语言模型在处理有争议知识的问题回答能力

  DEBATEQA: Evaluating Question Answering on Debatable Knowledge, *ArXiv*, *abs*/2408.01419.

KIEval: A Knowledge-grounded Interactive Evaluation Framework for Large Language Models

KIEval：一个基于知识的动态交互式评估框架

Motivation



- **当前的“静态数据集+做选择题”方法不能全面评估模型的性能**

主流的基于静态的数据集的评估方法，往往只需大模型生成对于QA任务或选择题的选项答案，这一点与目前大语言模型的开放式文本生成的主要用途并不匹配。

- **数据污染问题**

数据污染，即模型在训练过程中接触到评测基准的测试集数据，导致其在自动评测基准的表现被高估。

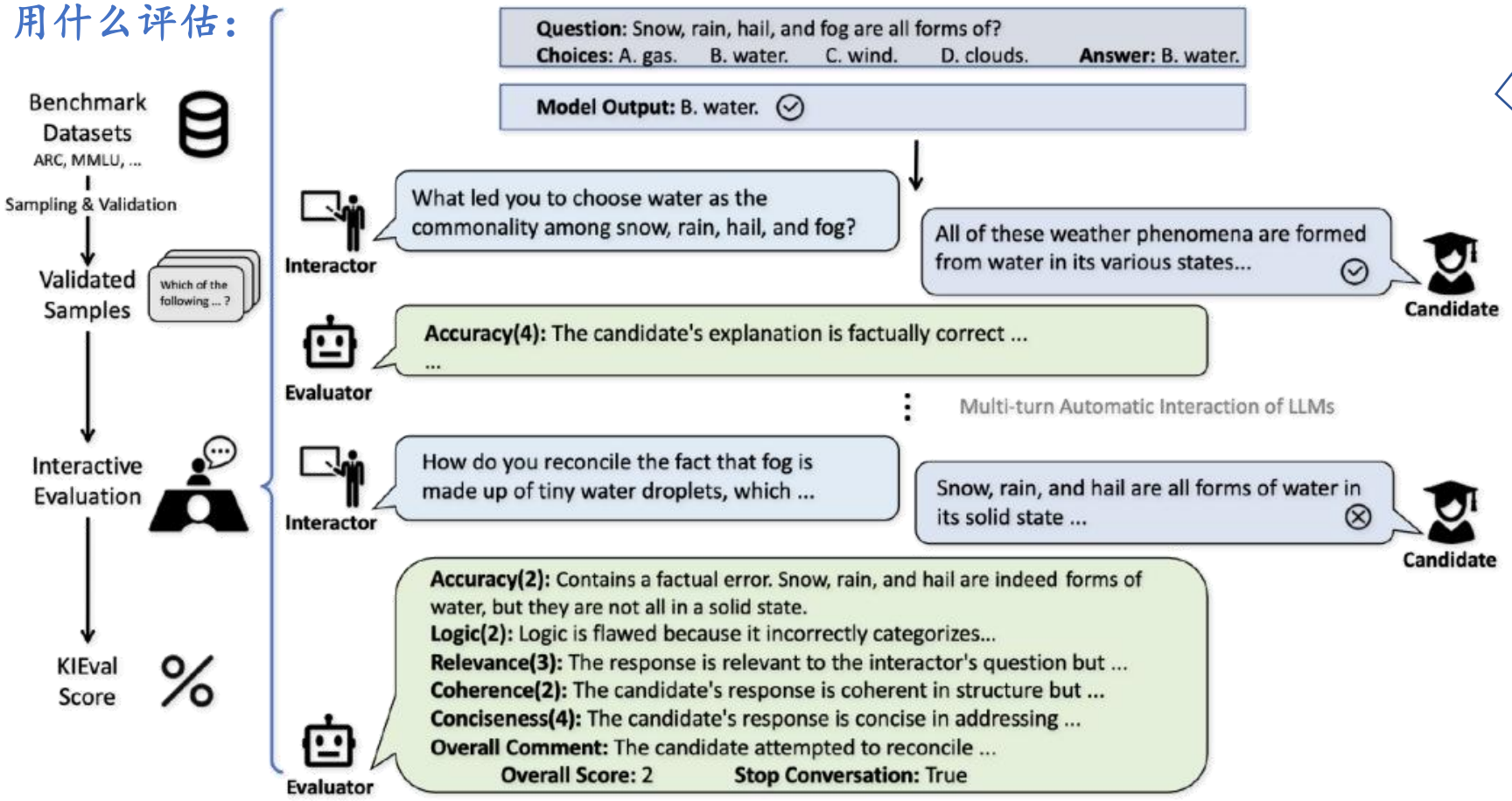
KIEval: A Knowledge-grounded Interactive Evaluation Framework for Large Language Models (Yu et al., ACL 2024)

Method

KIEval引入了一个「交互者」大模型，与被评估模型进行多轮对话。在每一轮交互中，「交互者」根据先前的对话历史，动态生成新的、更为深入的问题，引导被评估模型灵活运用其知识，生成连贯、相关的回复。

在对话过程中，引入「评估者」大模型，重点关注模型回复的准确性、逻辑性、相关性、连贯性、简洁性等指标，而非仅仅考察其回复是否与参考答案匹配。

用什么评估：



怎么评估：

$$\text{KIEvalScore} = \frac{\sum_{i=1}^{n} s_i w_i}{\sum_{i=1}^{n} w_i}$$

$$w_i = \exp(-\frac{i}{n})$$

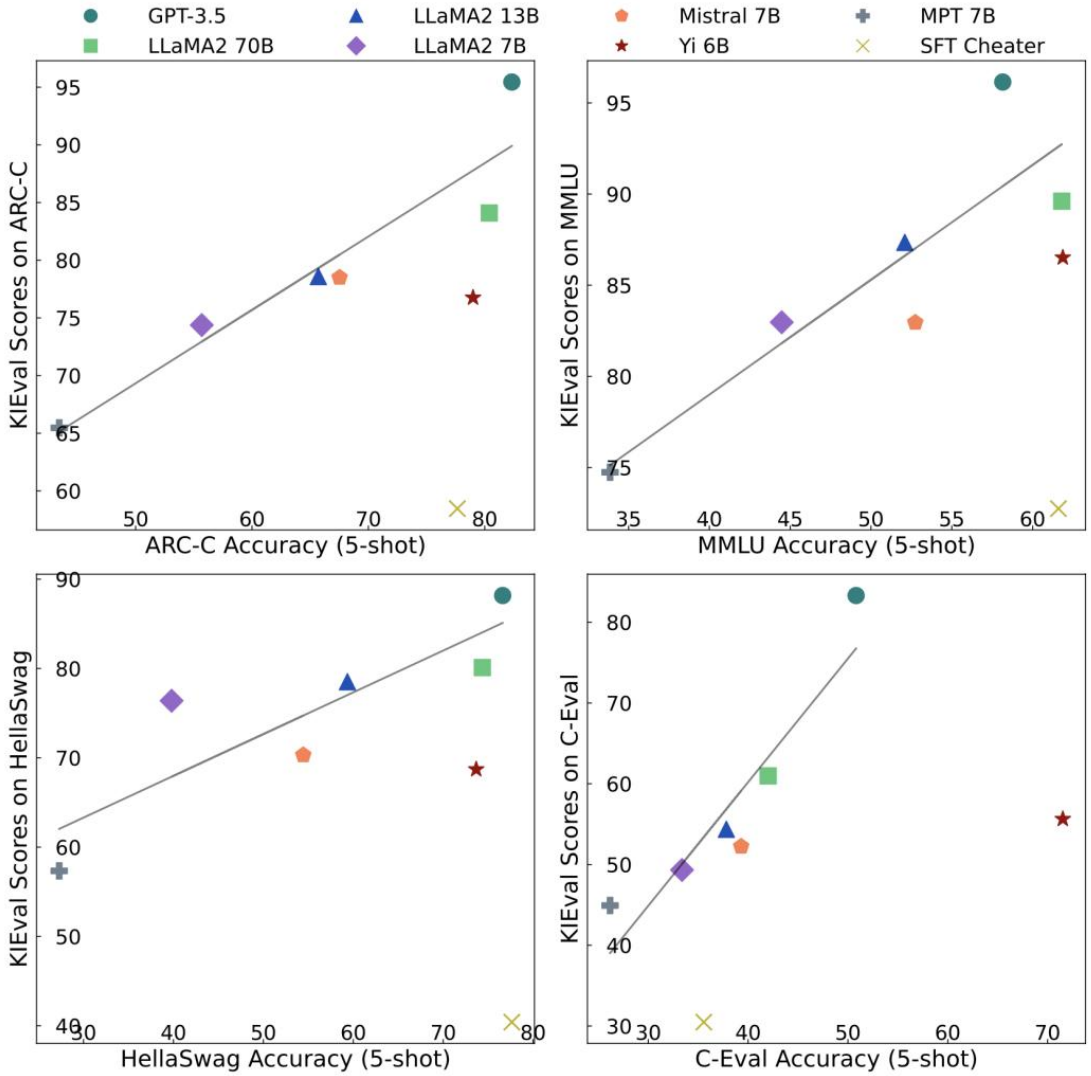
更强调早期的对话内容。
实验中$n=5$，但「评估者」有权利提前终止对话。

KIEval: A Knowledge-grounded Interactive Evaluation Framework for Large Language Models (Yu et al., ACL 2024)

Experiment

①传统的基准测试低估了模型之间的真实性能差距

在静态数据集上，不同模型的得分差异可能并不明显。但将这些模型置于 KIEval的动态对话场景中时，它们在知识运用、逻辑推理等方面的差距被显著放大。



KIEval: A Knowledge-grounded Interactive Evaluation Framework for Large Language Models (Yu et al., ACL 2024)

Experiment

②数据污染只是提高了模型对特定答案的记忆，而非真正增强其知识理解和运用的能力

作者构造了若干「作弊」模型，将评测数据集的一部分测试样本加入到「作弊」模型的训练集中作者发现，这些在训练时接触过测试集的模型，虽然在对应的测试集上取得了很高的分数，但在KIEval的动态对话中却表现平平，并未在「作弊」训练中得到正向提升。

应用：可以通过观察KIEval分数与静态评估数据集准确率的分数关系，推测数据泄露的存在（新方法）。

Table 3: Comparison on different data contamination scenarios on ARC-Challenge (Clark et al., 2018) dataset. 'SFT-Cheater' and 'PT-Cheater' denote leaking test-set labels during supervised fine-tuning phase and pre-training phase. We report 5-shot accuracy on ARC-Challenge dataset and KIEval scores. We detect data contamination with differences in average language modelling loss (Wei et al., 2023) and Min-K% Prob (Shi et al., 2023).

	ARC-C	Avg. LM Loss			Min-K%	KIEval					
	Acc.(5-shot)	$\mathcal{L}_{train}$	$\mathcal{L}_{test}$	Δ	AUC	Acc.	Log.	Rel.	Coh.	Con.	Overall
Normal (LLaMA 2 7B + SFT)	52.8	3.12	3.10	-0.02	0.53	61.7	62.1	84.4	69.2	70.6	66.3
SFT-Cheater	69.8	4.05	3.95	-0.09	0.54	52.8	52.3	72.8	60.2	57.7	56.1
PT-Cheater	76.8	3.88	2.02	-1.86	0.89	50.8	49.9	65.6	54.5	49.0	51.2
LLaMA 2 7B Chat	57.8	3.05	3.01	-0.04	0.55	75.3	75.9	90.1	80.2	74.0	77.9

Experiment

③大模型本身偏向性对于评估结果的影响不大

大模型本身可能具有一定偏向性（例如GPT系列模型可能更倾向于自身的输出）。

对一组相同的被评估模型使用相同的「交互者」，即可确保交互的双方输出保持不变；针对相同的交互输出，仅需使用不同的「评估者」对被评估模型的输出进行重复评价，即可得到不同模型针对同一被试模型的评估结果。

实验表明，尽管在样本级别上，这一偏向性确实存在，但在总体评估分数上，不同评估者模型给出的分数具有较强的正相关性，因此大模型的偏向性不易影响总体。

Table 13: KIEval scores using Claude 3 Opus (`claude-3-opus-20240229`) and GPT-4 Turbo (`gpt-4-turbo-preview-1106`) as evaluators on ARC-Challenge dataset.

Model	Evaluator	Accuracy	Logic	Relevance	Coherence	Conciseness	Overall
GPT-3.5	GPT-4	94.6	94.7	98.5	96.1	97.3	95.5
	Claude-3	98.6	98.8	99.8	99.4	99.0	98.7
LLaMA-2 70B	GPT-4	81.9	82.8	92.2	85.3	75.6	84.1
	Claude-3	98.3	98.7	98.2	96.9	84.6	96.4
LLaMA-2 7B	GPT-4	70.6	71.6	90.4	77.9	71.7	74.4
	Claude-3	90.9	91.8	98.0	95.0	85.2	91.0

Table 14: Pearson (r), Spearman (ρ), Kendall-Tau (τ) correlation efficients of KIEval scores evaluated by `claude-3-opus-20240229` and `gpt-4-turbo-preview-1106`.

Metric	Corr. Coeff.	P-Value
Pearson r	0.822	2.87e-05
Spearman ρ	0.898	4.17e-07
Kendall τ	0.761	1.10e-05

KIEval: A Knowledge-grounded Interactive Evaluation Framework for Large Language Models (Yu et al., ACL 2024)

分享论文目录

大模型综合评估

- 精确评估大语言模型的世界知识

  KoLA: Carefully Benchmarking World Knowledge of Large Language Models (Yu et al., ICLR 2024)

- 基于知识的交互式评估方法

  KIEval: A Knowledge-grounded Interactive Evaluation Framework for Large Language Models (Yu et al., ACL 2024)

大模型相关应用的评估（RAG & Agent）

- 检索增强生成（RAG）系统的自动评估

  RAGAs: Automated Evaluation of Retrieval Augmented Generation (Es et al., EACL 2024)

- 特定场景下检索增强生成（RAG）评估数据生成框架

  RAGEval: Scenario Specific RAG Evaluation Dataset Generation Framework, *ArXiv*, *abs*/2408.01262.

- 大模型工具使用能力评测（Agent 能力）

  T-Eval: Evaluating the Tool Utilization Capability of Large Language Models Step by Step (Chen et al., ACL 2024)

大模型特定任务or能力的评估

- 评估大语言模型在处理有争议知识的问题回答能力

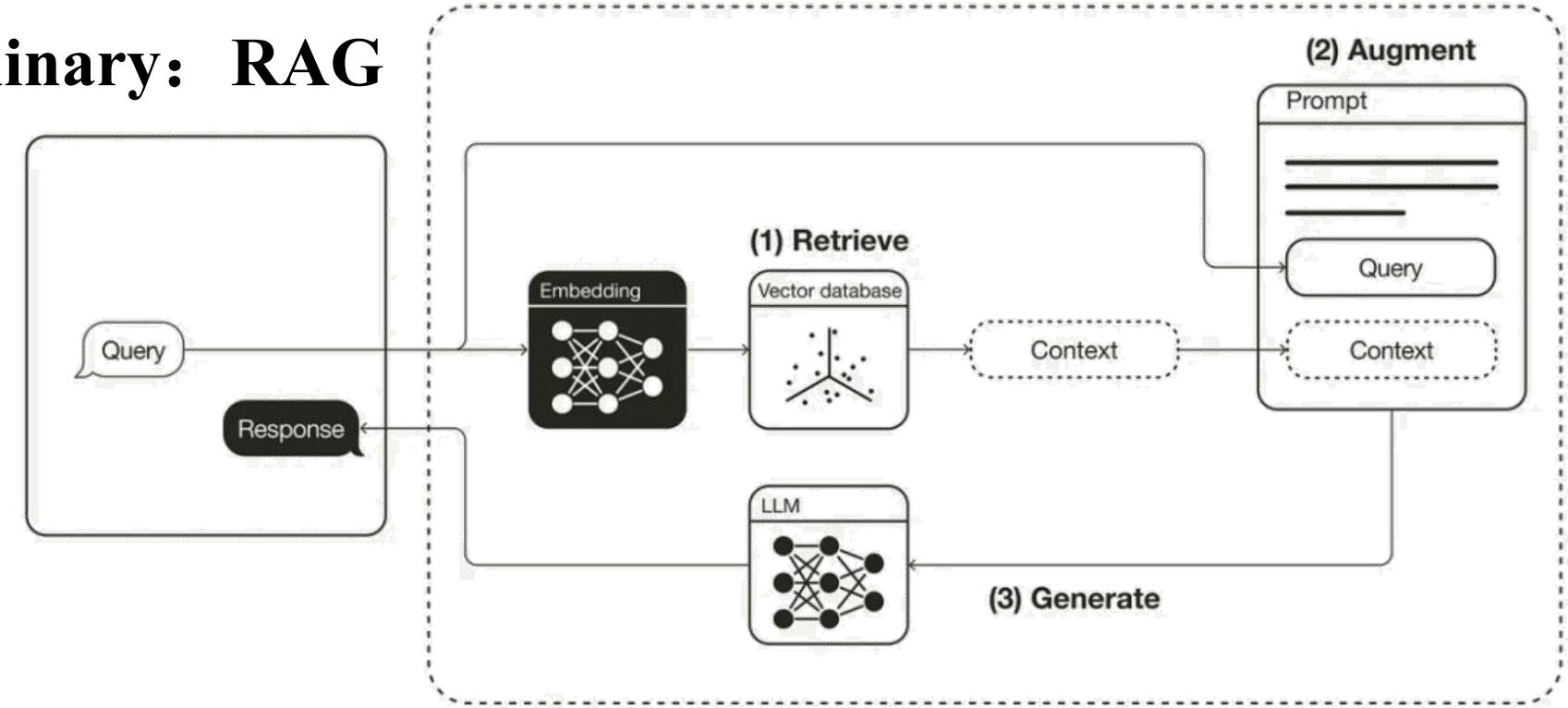
  DEBATEQA: Evaluating Question Answering on Debatable Knowledge, *ArXiv*, *abs*/2408.01419.

RAGAs: Automated Evaluation of Retrieval Augmented Generation

RAGAs: 检索增强生成系统的自动评估工具

https://github.com/explodinggradients/ragas

Preliminary：RAG



「**检索(Retrive)**」 根据用户请求从外部知识源检索相关上下文。为此，使用嵌入模型将用户查询嵌入到与向量数据库中的附加上下文相同的向量空间中。这允许执行相似性搜索，并返回矢量数据库中最接近的前 k 个数据对象。

「**增强(Augment)**」 用户查询和检索到的附加上下文被填充到提示模板中。

「**生成(Generate)**」 最后，检索增强提示被馈送到 LLM。

RAG可以为LLM提供外部知识源，使它们能够生成准确且符合上下文的答案，同时能够减少模型幻觉。

Motivation

- 早期评估直接套用语言模型、问答系统的评估方法，不完全适配RAG系统
- 实际应用中并不总是有参考答案(*Ground Truth*)

Contribution

- 提出了一个针对RAG系统的自动评估框架，不需要在特定的数据集上进行评估，只需单次问答过程的 Question + Answer + Context 即可评估

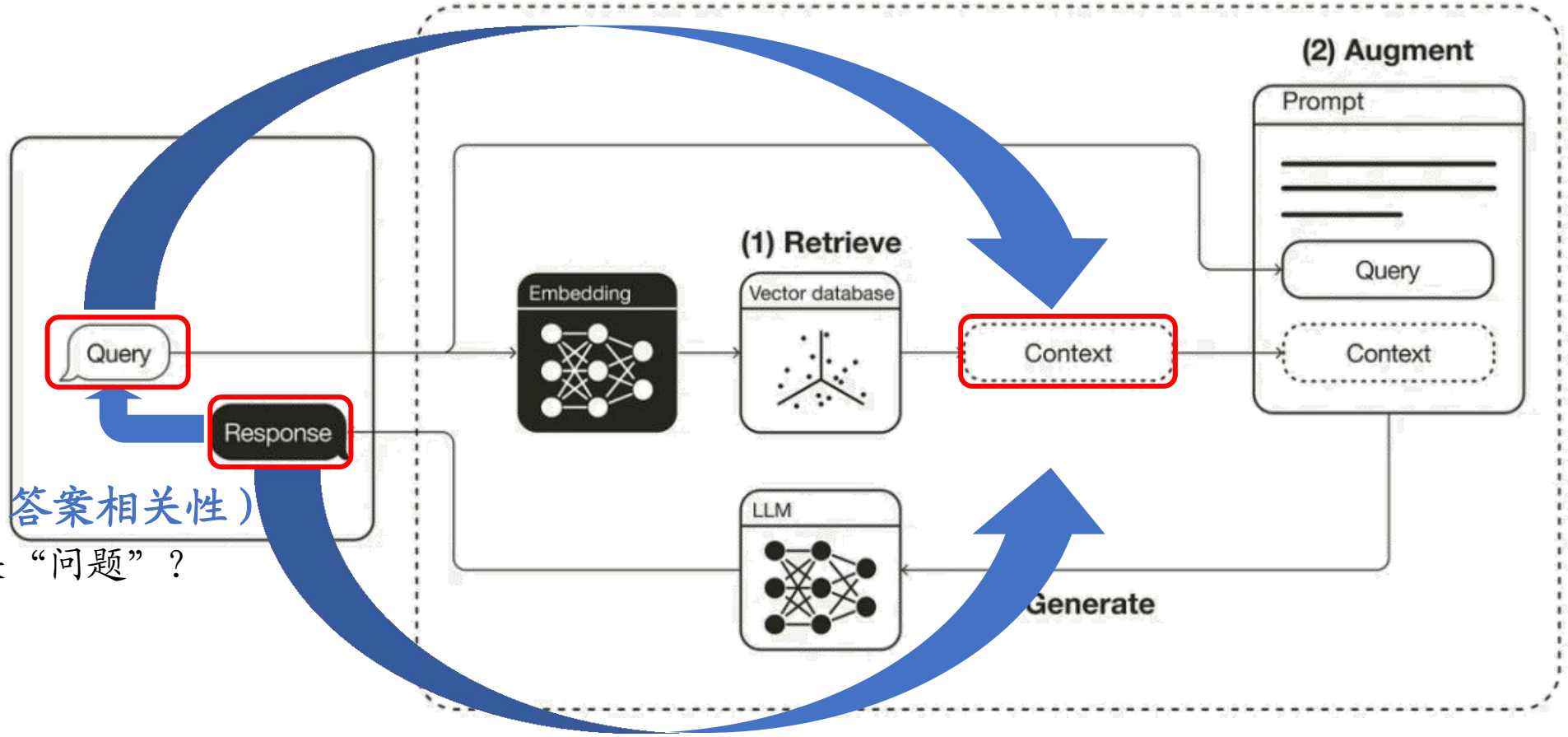
RAGAs: Automated Evaluation of Retrieval Augmented Generation (Es et al., EACL 2024)

Method

怎么评估？

Context_Relevancy（上下文相关性）

“上下文”中是否只包含了回答问题所需信息？



Answer_Relevancy（答案相关性）

“答案”是否可以直接解决“问题”？

Faithfulness（忠实度）

“答案”是否可以从“上下文”中推断出来？

RAGAs: Automated Evaluation of Retrieval Augmented Generation (Es et al., EACL 2024)

Method

怎么评估？

首先让LLM将“答案”提炼为一个或多个statement
↓
再让LLM判断每个statement是否由context支持

Faithfulness（忠实度）

“答案”是否可以从“上下文”中推断出来？

$$F = \frac{|V|}{|S|}$$

——LLM支持的statement数
——statement总数

Answer_Relevancy（答案相关性）

“答案”是否可以直接解决“问题”？

$$\text{AR} = \frac{1}{n}\sum_{i=1}^{n} \text{sim}(q, q_i)$$

让LLM根据“答案”生成多个可能的问题q_i，计算真实“问题”q和每个q_i嵌入向量之间的余弦相似度

Context_Relevancy（上下文相关性）

“上下文”中是否只包含了回答问题所需信息？

$$\text{CR} = \frac{\text{number of extracted sentences}}{\text{total number of sentences in } c(q)}$$

让LLM从“上下文”中提取出所有对回答“问题”有帮助的句子

*本页的LLM均为gpt-3.5-turbo-16k

RAGAs: Automated Evaluation of Retrieval Augmented Generation (Es et al., EACL 2024)

Experiment

- WikiEval**数据集的构建方法**

首先选择50个维基百科页面，涵盖了自2022年初以来发生的事件。在选择这些页面时，我们优先考虑了那些最近进行过编辑的页面。

对于每个页面，要求ChatGPT给出“问题”和“答案”。

- **在WikiEval数据集上进行实验，结果表明，相比于另外两种方法，RAGAs方法与人类评估的重合度最高。**

GPT Score：让ChatGPT对每个回答进行0-10评分

GPT Ranking：让ChatGPT对每个回答进行排序

	Faith.	Ans. Rel.	Cont. Rel.
RAGAs	**0.95**	**0.78**	**0.70**
GPT Score	0.72	0.52	0.63
GPT Ranking	0.54	0.40	0.52

Table 1: Agreement with human annotators in pairwise comparisons of faithfulness, answer relevance and context relevance, using the WikEval dataset (accuracy).

RAGAs: Automated Evaluation of Retrieval Augmented Generation (Es et al., EACL 2024)

RAGAs工具提供的其他指标：如果你拥有*Ground Truth*

- Faithfulness
- Answer relevancy
- Context recall
- Context precision
- Context utilization
- Context entity recall
- Noise Sensitivity
- Summarization Score

$$\text{Context Precision@K} = \frac{\sum_{k=1}^{K} (\text{Precision@k} \times v_k)}{\text{Total number of relevant items in the top } K \text{ results}}$$

$$\text{Precision@k} = \frac{\text{true positives@k}}{(\text{true positives@k} + \text{false positives@k})}$$

Where K is the total number of chunks in `contexts` and $v_k \in \{0, 1\}$ is the relevance indicator at rank k.

$$\text{context recall} = \frac{|\text{GT claims that can be attributed to context}|}{|\text{Number of claims in GT}|}$$

不足：仍然是根据“答案”的*Ground Truth*间接计算，没有“上下文”的*Ground Truth*
——可不可以做一个数据集，让“上下文”也有*Ground Truth*？

https://github.com/explodinggradients/ragas

分享论文目录

大模型综合评估

- 精确评估大语言模型的世界知识

  KoLA: Carefully Benchmarking World Knowledge of Large Language Models (Yu et al., ICLR 2024)

- 基于知识的交互式评估方法

  KIEval: A Knowledge-grounded Interactive Evaluation Framework for Large Language Models (Yu et al., ACL 2024)

大模型相关应用的评估（RAG & Agent）

- 检索增强生成（RAG）系统的自动评估

  RAGAs: Automated Evaluation of Retrieval Augmented Generation (Es et al., EACL 2024)

- **特定场景下检索增强生成（RAG）评估数据生成框架**

  RAGEval: Scenario Specific RAG Evaluation Dataset Generation Framework, *ArXiv*, *abs*/2408.01262.

- 大模型工具使用能力评测（Agent 能力）

  T-Eval: Evaluating the Tool Utilization Capability of Large Language Models Step by Step (Chen et al., ACL 2024)

大模型特定任务or能力的评估

- 评估大语言模型在处理有争议知识的问题回答能力

  DEBATEQA: Evaluating Question Answering on Debatable Knowledge, *ArXiv*, *abs*/2408.01419.

RAGEval: Scenario Specific RAG Evaluation Dataset Generation Framework

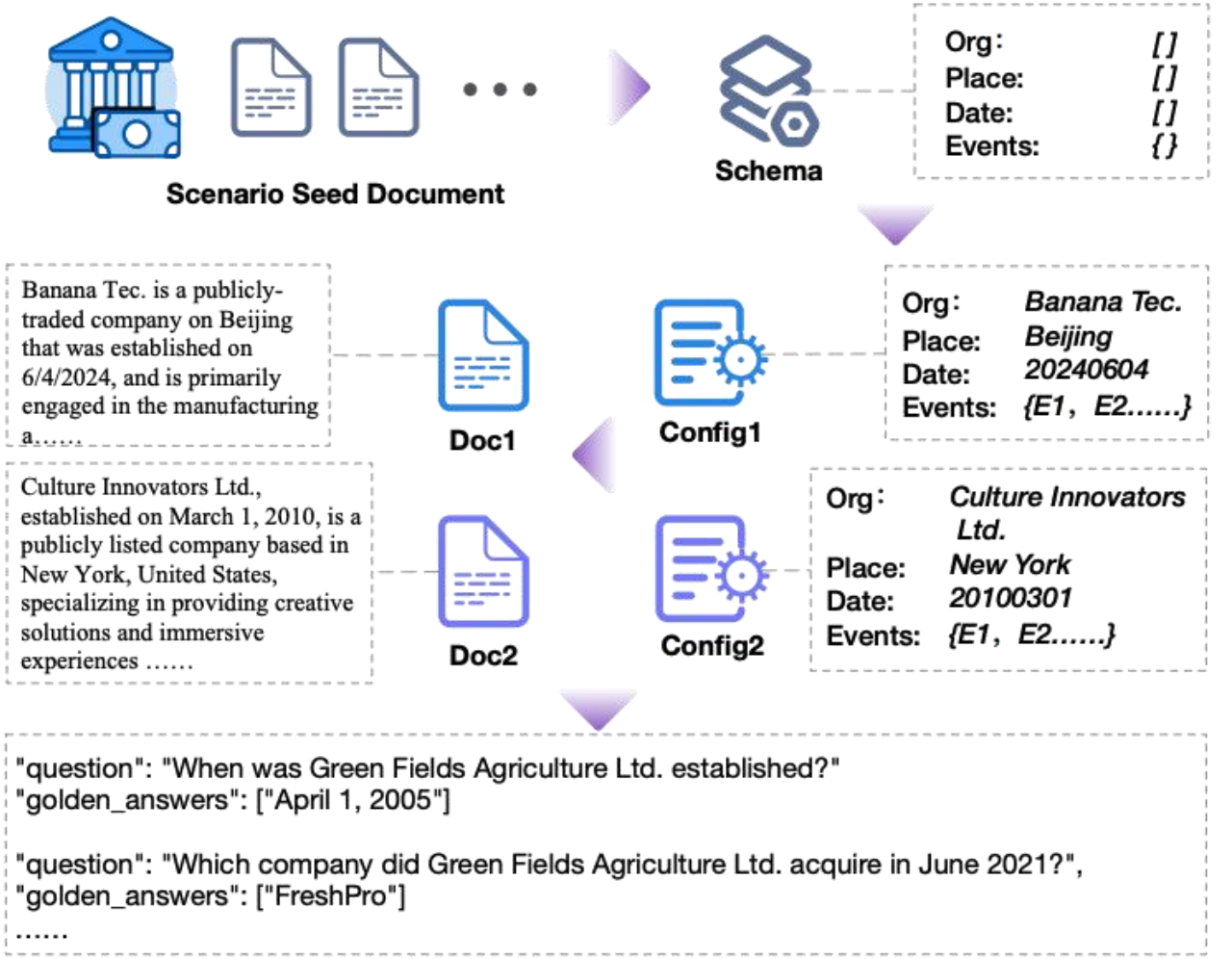
特定场景下 RAG 评估数据集生成框架

经济/法律/医学

Motivation

- 现有的RAG benchmark主要在一般领域，而缺乏在垂直领域（如金融、医疗保健和法律）评估RAG的数据集；
- 在垂直领域获取高质量和多样化的真实数据成本高，难度大。

→不如自己来“虚构”数据



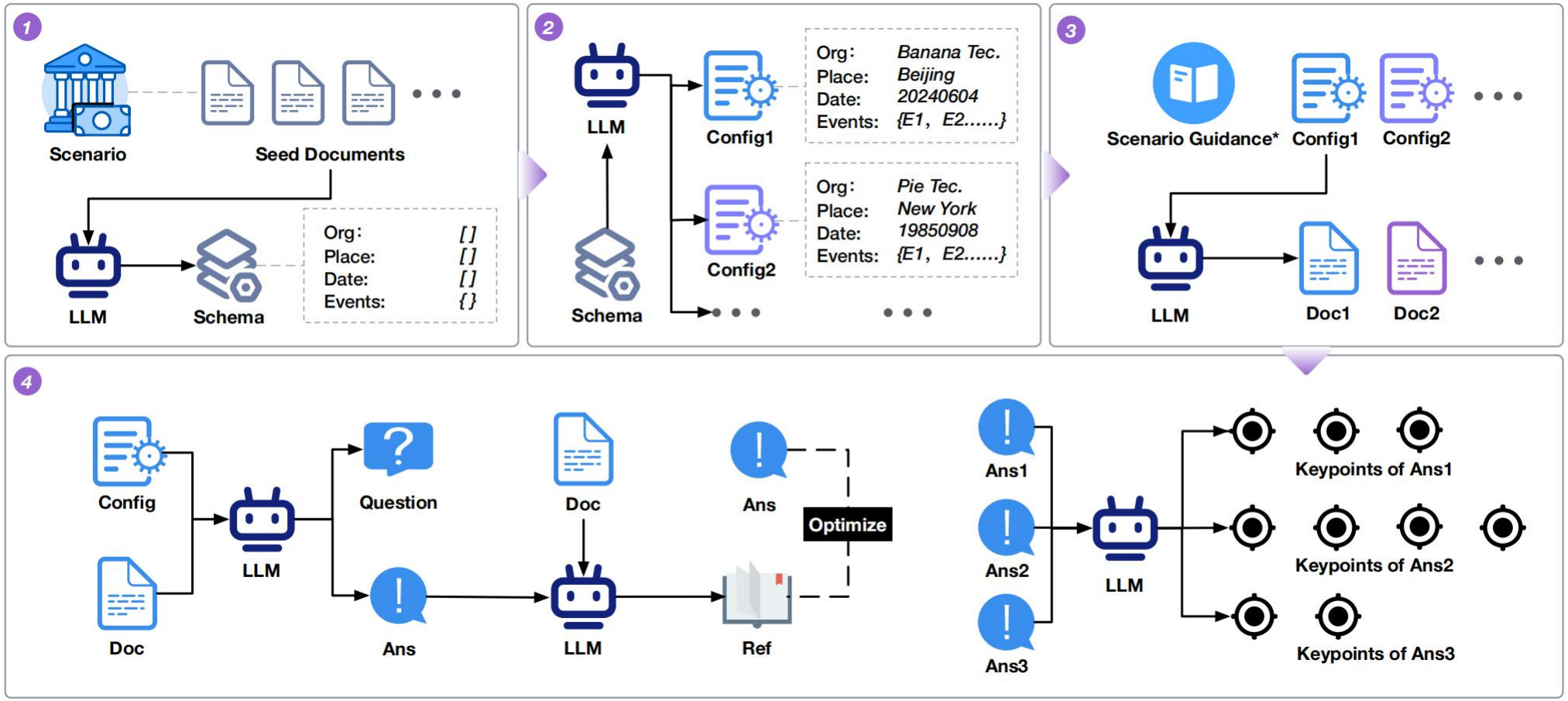
RAGEval （schema-configuration-document-QAR-keypoint）

RAGEval: Scenario Specific RAG Evaluation Dataset Generation Framework, *ArXiv*, *abs*/2408.01262.

Method

用什么评估

① 首先收集一小组特定于领域的种子文档来总结一个「大纲(schema)」，其中封装了基本的特定于领域的知识；
② 根据该「大纲」生成不同的「要点(config)」；
③ 进一步利用这些「要点」来生成不同的「文档(doc)」；
④ 使用给定的文档*D*和要点*C*生成，问题-参考-答案(QRA)三元组



a) 利用config + LLM→问题Q和初始答案A
b) 使用问题Q+初始答案A，从doc中提取参考文献R
c) 保证答案A和参考文献R对齐：R有A没有则补A，A有R遗漏则找R
d) 从每个标准答案A中生成key points

RAGEval: Scenario Specific RAG Evaluation Dataset Generation Framework, *ArXiv*, *abs*/2408.01262.

Method：怎么评估

检索指标

- Recall： 评估检索结果的ref是否全部找回了ground truth中的ref

$$\text{Recall} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(M(G_i, \mathcal{R})),$$

- Effective Information Rate (EIR)：检索结果的ref中与ground truth中的ref重叠字数的百分比（信噪比）

$$\text{EIR} = \frac{\sum_{i=1}^{m}|G_i \cap R_t|}{\sum_{j=1}^{k}|R_j|}$$

RAGEval: Scenario Specific RAG Evaluation Dataset Generation Framework, *ArXiv*, *abs*/2408.01262.

Method：Metrics（怎么评估）

生成指标

- Completeness：生成的答案与ground truth中key points重叠的比例

$$\text{Comp}(A, K) = \frac{1}{|K|} \sum_{i=1}^{n} \mathbb{1}[A \text{ covers } k_i]$$

- Hallucination：生成的答案中与key points相矛盾的比例

$$\text{Hallu}(A, K) = \frac{1}{|K|} \sum_{i=1}^{n} \mathbb{1}[A \text{ contradicts } k_i]$$

- Irrelevancy：不相关性评估的是来自基本事实的关键点的比例，这些关键点既没有被生成的答案覆盖，也没有与之相矛盾。

$$\text{Irr}(A, K) = 1 - \text{Comp}(A, K) - \text{Hallu}(A, K)$$

RAGEval: Scenario Specific RAG Evaluation Dataset Generation Framework, *ArXiv*, *abs*/2408.01262.

Experiment

较高的Recall 和EIR得分（检索得好）通常会导致较好的completeness和hallucination得分（生成得好）

Table 3: Overall Model Performance Results. (Without irrelevant result)

Model	Completeness (↑)		Hallucination (↓)		Irrelevance (↓)		Rouge-L (↑)	
	CN	EN	CN	EN	CN	EN	CN	EN
MiniCPM-2B-sft	0.4114	0.5484	0.4080	0.2115	0.1803	0.2401	0.2773	0.2505
Baichuan-2-7B-chat	0.4009	0.5498	0.4181	0.2212	0.1809	0.2290	**0.3262**	**0.3039**
Qwen1.5-7B-chat	0.3983	0.5704	0.4058	0.1953	0.1957	0.2340	0.2040	0.1862
Qwen2-7B-Instruct	0.4564	0.6052	0.3829	0.1955	**0.1596**	0.1988	0.2035	0.2182
Llama3-8B-Instruct	0.4427	0.6524	0.3888	**0.1582**	0.1679	0.1894	0.1982	0.2406
Qwen1.5-14B-chat	0.4926	0.6053	0.3440	0.1795	0.1630	0.2152	0.2611	0.2330
GPT3.5-Turbo	0.4774	0.6540	0.3601	0.1901	0.1626	0.1556	0.2309	0.2563
GPT-4o	**0.5187**	**0.6845**	**0.2797**	0.1636	0.1972	**0.1520**	0.1527	0.2190

Table 4: Retrieval Model Performance Results.

Model	Retrieve				Generation					
	Recall (↑)		EIR (↑)		Completeness (↑)		Hallucination (↓)		Irrelevance (↓)	
	CN	EN	CN	EN	CN	EN	CN	EN	CN	EN
BM25	0.7662	0.6717	0.0470	0.1162	0.6316	0.6649	0.2441	0.1264	0.1242	0.2087
GTE-Large	0.5760	**0.7542**	0.0362	**0.1372**	0.5337	0.6921	0.2851	**0.1042**	0.1813	0.2037
BGE-Large	0.6881	0.7321	0.0465	0.1362	0.5780	**0.7077**	0.2794	0.1129	0.1426	**0.1795**
BGE-M3	**0.8387**	0.6928	**0.0541**	0.1243	**0.6980**	0.6556	**0.2004**	0.1254	**0.1010**	0.2190

RAGEval: Scenario Specific RAG Evaluation Dataset Generation Framework, *ArXiv*, *abs*/2408.01262.

Experiment

Table 5: TopK & Chunk-TopK Performance Results.

Settings	Retrieve				Generation					
	Recall (↑)		EIR (↑)		Completeness (↑)		Hallucination (↓)		Irrelevance (↓)	
	CN	EN	CN	EN	CN	EN	CN	EN	CN	EN
TopK										
2	0.4667	0.5685	**0.0764**	**0.2491**	0.5004	0.5682	0.3226	0.1693	0.1770	0.2625
4	0.6362	0.6976	0.0553	0.1591	0.5517	0.6503	0.3127	0.1303	0.1352	0.2194
6	**0.7259**	**0.7542**	0.0408	0.1182	**0.5835**	**0.7087**	**0.2974**	**0.1227**	**0.1191**	**0.1686**
Chunk-TopK										
128-8	**0.5031**	0.5472	**0.0884**	0.2222	0.4549	**0.6683**	**0.2861**	**0.1168**	0.2591	**0.2148**
256-4	0.4393	**0.6161**	0.0824	**0.2628**	0.4855	0.6509	0.3196	0.1241	0.1944	0.2250
512-2	0.4667	0.5685	0.0764	0.2491	**0.4932**	0.5609	0.3195	0.1635	**0.1873**	0.2756

① TopK的实验结果比较符合直觉：topK越大，recall越大，导致EIR更低，回答的完整性、幻觉和不相关性都更高（也许这就是多答多错）。

② retrieval和generation的表现并不一定一直正相关，需要trade-off，根据不同场景和任务选择适合的模型和超参数配置。

RAGEval: Scenario Specific RAG Evaluation Dataset Generation Framework, *ArXiv*, *abs*/2408.01262.

分享论文目录

大模型综合评估

- 精确评估大语言模型的世界知识

  KoLA: Carefully Benchmarking World Knowledge of Large Language Models (Yu et al., ICLR 2024)

- 基于知识的交互式评估方法

  KIEval: A Knowledge-grounded Interactive Evaluation Framework for Large Language Models (Yu et al., ACL 2024)

大模型相关应用的评估（RAG & Agent）

- 检索增强生成（RAG）系统的自动评估

  RAGAs: Automated Evaluation of Retrieval Augmented Generation (Es et al., EACL 2024)

- 特定场景下检索增强生成（RAG）评估数据生成框架

  RAGEval: Scenario Specific RAG Evaluation Dataset Generation Framework, *ArXiv*, *abs*/2408.01262.

- 大模型工具使用能力评测（Agent 能力）

  T-Eval: Evaluating the Tool Utilization Capability of Large Language Models Step by Step (Chen et al., ACL 2024)

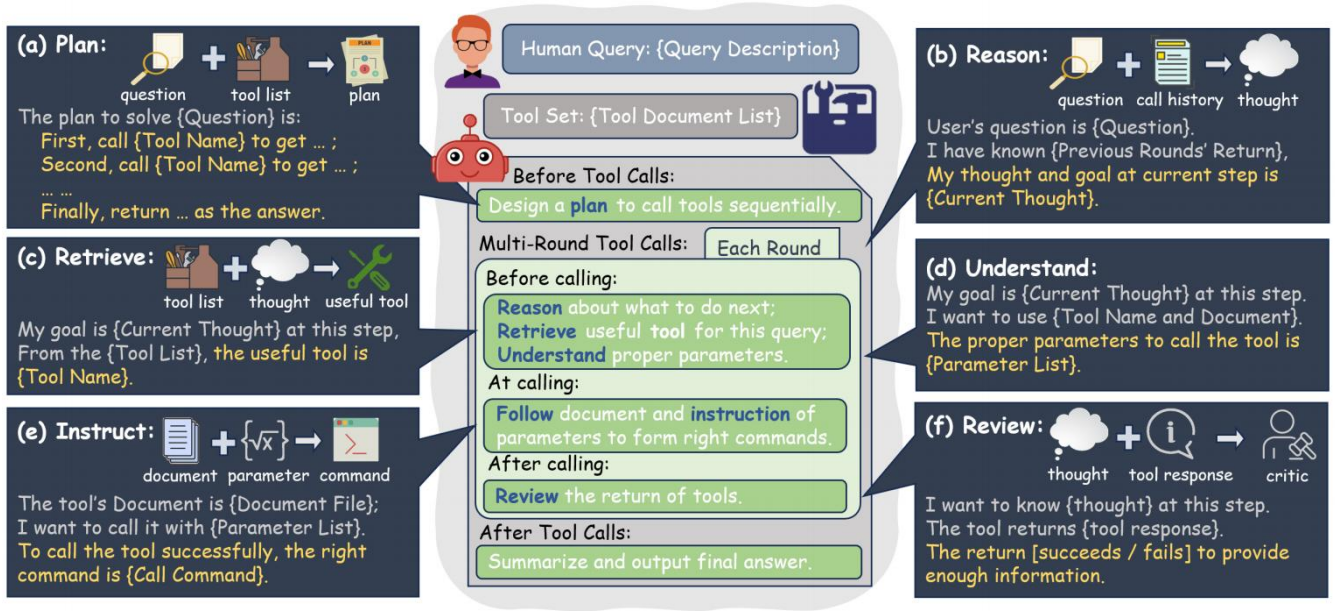
大模型特定任务or能力的评估

- 评估大语言模型在处理有争议知识的问题回答能力

  DEBATEQA: Evaluating Question Answering on Debatable Knowledge, *ArXiv*, *abs*/2408.01419.

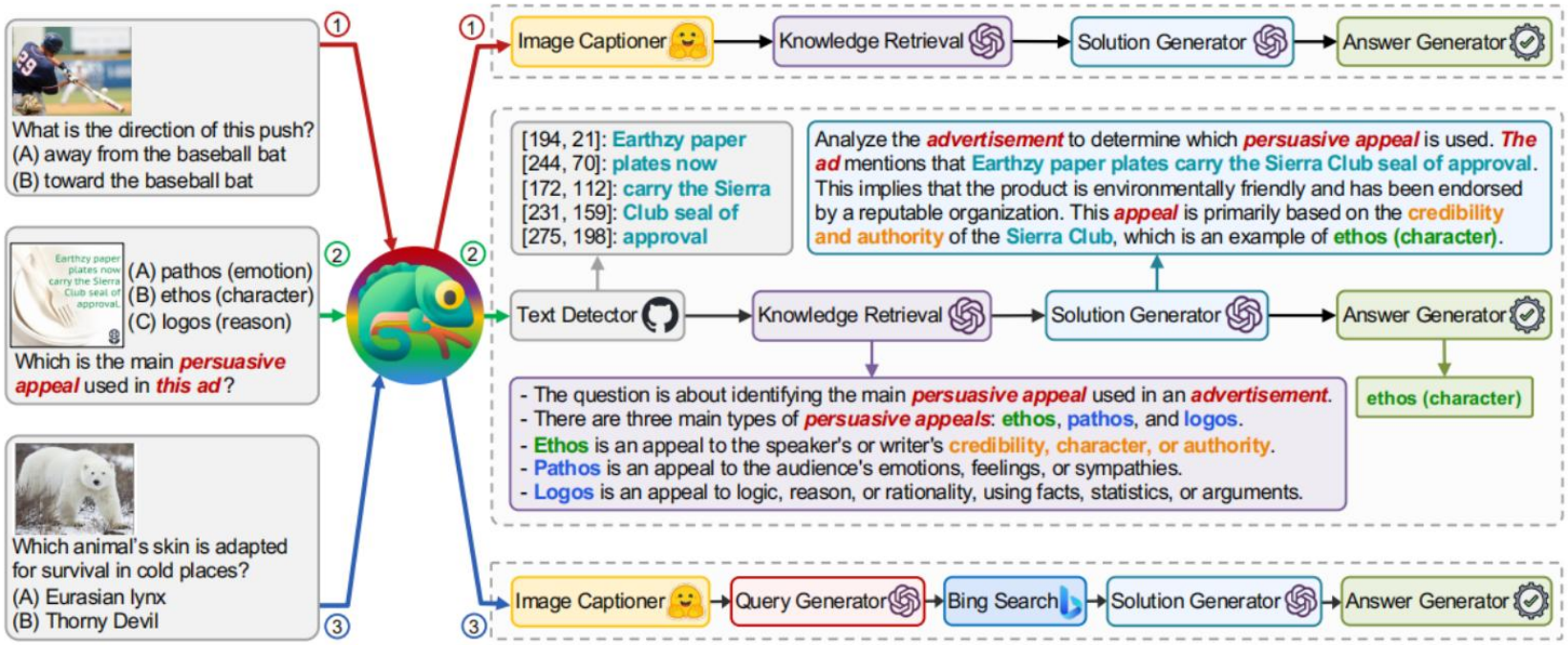
T-Eval: Evaluating the Tool Utilization Capability of Large Language Models Step by Step

T-Eval: 逐步评估大语言模型的工具使用能力（Agent 能力）



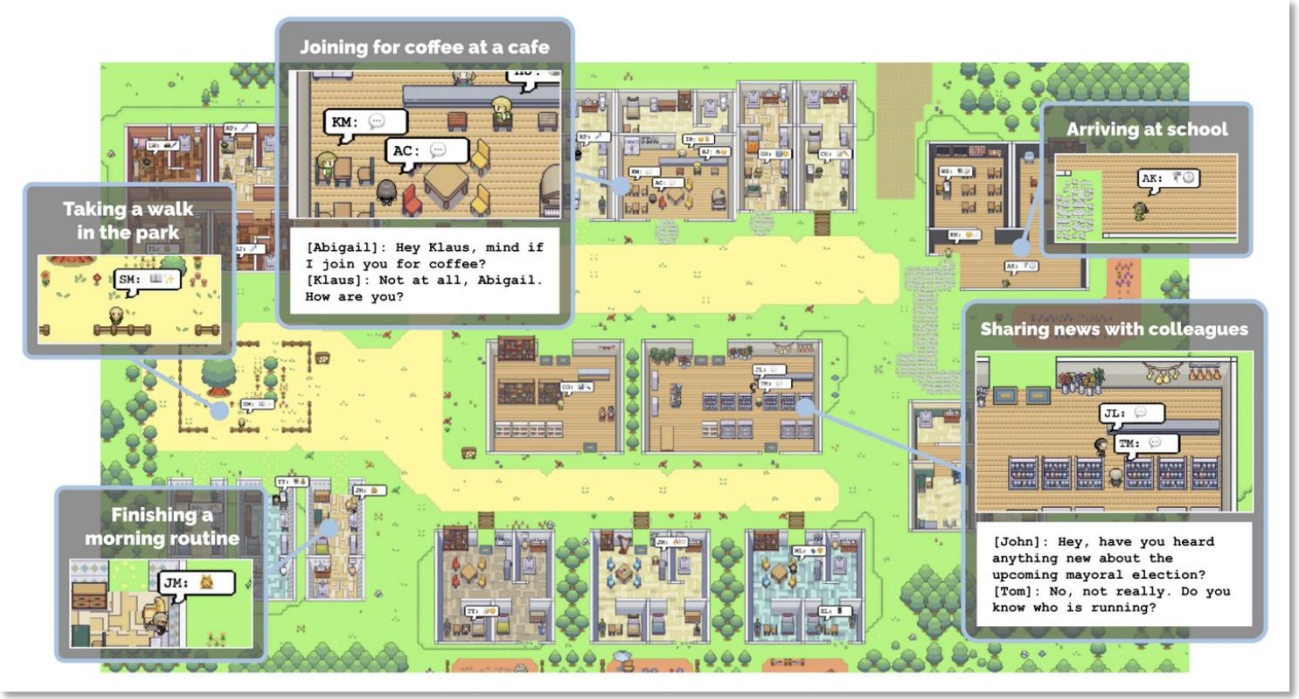
Preliminary：LLM Agent 与工具调用

将LLM作为“中枢大脑(Manager)”，根据解决问题的需要调用不同工具，即为一个Agent



以图中的问题2为例，当 Chameleon 接收到相关的输入内容时，首先调用「文字检测」工具检测输入图像中的文本信息，接着调用「知识搜索」工具检索与输入选项 Ethos, Pathos and Logos 相关的知识，然后调用「方案解决」工具生成推理步骤，最后使用「答案生成」工具总结出最终答案。

Preliminary：LLM Multi-Agent 协作



斯坦福AI小镇：25个AI智能体组成的虚拟世界

多智能体系统安全性研究

PsySafe: A Comprehensive Framework for Psychological-based Attack, Defense, and Evaluation of Multi-agent System Safety

Zaibin Zhang[1,2*], Yongting Zhang[1,3*], Lijun Li[1], Hongzhi Gao[1,3], Lijun Wang[2], Huchuan Lu[2], Feng Zhao[3], Yu Qiao[1], Jing Shao[1†]
[1] Shanghai Artificial Intelligence Laboratory
[2] Dalian University of Technology
[3] University of Science and Technology of China
{zhangzaibin, zhangyongting, shaojing}@pjlab.org.cn

ACL 2024 Outstanding paper

人类难以同时胜任所有任务 → 团队合作、层级合作、师徒传承

如果强行对单一智能体赋予过多的能力，急剧上升的复杂度会让其难以驾驭，从而降低实用价值 → 借鉴人类社会发展出的协作模式，对能力单一的智能体进行协作编排

Motivation

现有评估方法通常只关注模型处理单步骤任务时的工具调用表现，缺少在多步骤复杂任务场景下模型使用工具能力的评估。

Contribution

相较于之前整体评估模型的方式，论文中将大模型的工具使用分解为多个子过程，包括：

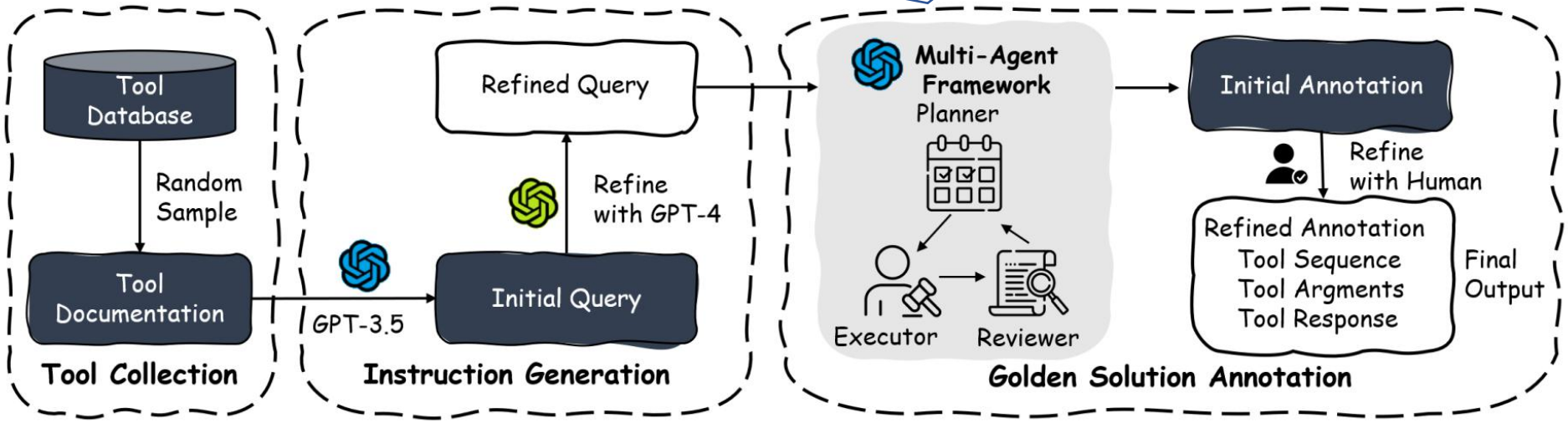
规划、推理、检索、理解、指令跟随和审查。

规划（plan）： 将用户问题分解为多个子问题，制定行动计划。
推理（reason）： 对上个状态的完成情况的判断，下一步行动的思考。
检索（retrieve）： 从给定的工具列表中选择合适的工具。
理解（understand）： 正确理解工具使用的参考文档和所需参数。
指令跟随（instruct）： 生成指定格式的工具调用请求。
审查（review）： 评估每个工具调用执行的结果，确保回答满足预期目标。

T-Eval: Evaluating the Tool Utilization Capability of Large Language Models Step by Step (Chen et al., ACL 2024)

Method：用什么评估

多智能体框架：不是只实例化一个LLM来处理整个解决方案注释路径，而是明确地将注释任务分解为三个不同的功能，包括计划者、执行者和评审者：「计划者」决定下一步应该做什么；「执行者」负责生成确切的工具名及其参数，并执行工具以获得响应；「评审员」被指派修改工具的响应，并根据外部反馈判断任务是否完成。



T-Eval 数据集的构建主要包括 3 个阶段：**工具收集、指令生成和参考答案标注**，具体流程如下：

① 根据可用性和使用率，挑选15种基本工具，涵盖了研究、旅行、娱乐、网络、生活和金融等多个领域。此外，还为每个工具生成了详细的API文档，以减少因工具描述不充分而导致的工具调用失败案例。

② 利用 GPT-3.5 生成了初始问题，并通过 GPT-4 进一步完善问题。

③ 开发了一个多智能体框架，利用所提供的工具解决问题，同时收集解决方案路径和工具响应。最后，使用人类专家来挑选高质量样本。

T-Eval: Evaluating the Tool Utilization Capability of Large Language Models Step by Step (Chen et al., ACL 2024)

Method：怎么评估

规划（plan）：将用户问题分解为多个子问题，制定行动计划

$$P^{pred} = [a_1^{pred}, a_2^{pred}, ..., a_{n^{pred}}^{pred}] \quad \text{for pairs } (a_i = (tool_i, args_i), a_j = (tool_j, args_j))$$

$$P^{gt} = [a_1^{gt}, a_2^{gt}, ..., a_{n^{gt}}^{gt}] \qquad S_{i,j} = \beta\sigma(tool_i, tool_j) + (1-\beta)\sigma(args_i, args_j)$$

$$p = l/n^{pred} \text{ and } r = l/n^{gt}, \qquad \text{plan score} = \frac{2pr}{p+r}$$

推理（reason）：对上个状态的完成情况的判断，下一步行动的思考

给定一个工具列表T、查询q和解决方案的前缀 t_i ，要求LLM生成 t_{i+1}^{pred} ，计算与 t_{i+1}^{gt} 的余弦相似度

检索（retrieve）：从给定的工具列表中选择合适的工具

给定一个工具列表T、查询q和解决方案的前缀 t_i ，要求LLM生成需要调用的工具$tool_{i+1}^{pred}$，与 $tool_{i+1}^{gt}$ 相同得1分，否则0分

理解（understand）：正确理解工具使用的参考文档和所需参数（重内容）

给定一个工具列表T、查询q和解决方案的前缀 t_i ，要求LLM生成工具调用的参数$args_{i+1}^{pred}$，计算与$args_{i+1}^{gt}$ 的余弦相似度

指令跟随（instruct）：生成指定格式的工具调用请求（重格式）

满足正确格式 $(tool, args)$ 的得0.5分；剩余0.5分看正确的参数所占的百分比，最后求和

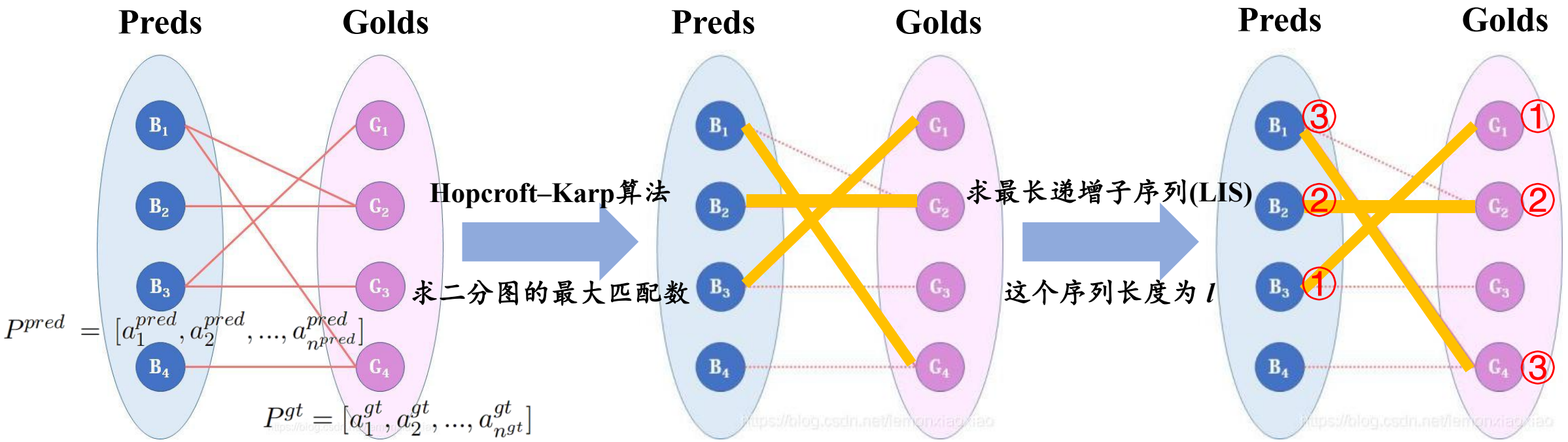
审查（review）：评估每个工具调用执行的结果，确保回答满足预期目标

给定解决方案的前缀 t_i 和调用工具的响应 o_i ，LLM须对响应进行分类：成功/错误/不相关...，分类正确得1分，否则0分

T-Eval: Evaluating the Tool Utilization Capability of Large Language Models Step by Step (Chen et al., ACL 2024)

Method：怎么评估

规划（plan）：将用户问题分解为多个子问题，制定行动计划



$$S_{i,j} = \beta\sigma(tool_i, tool_j) + (1-\beta)\sigma(args_i, args_j)$$

画一个二分图，$S \geqslant 0.7$则连线

精度和召回率：$p = l/n^{pred}$ and $r = l/n^{gt}$.

$$\text{plan score} = \frac{2pr}{p+r}$$

分享论文目录

大模型综合评估

- 精确评估大语言模型的世界知识

  KoLA: Carefully Benchmarking World Knowledge of Large Language Models (Yu et al., ICLR 2024)

- 基于知识的交互式评估方法

  KIEval: A Knowledge-grounded Interactive Evaluation Framework for Large Language Models (Yu et al., ACL 2024)

大模型相关应用的评估（RAG & Agent）

- 检索增强生成（RAG）系统的自动评估

  RAGAs: Automated Evaluation of Retrieval Augmented Generation (Es et al., EACL 2024)

- 特定场景下检索增强生成（RAG）评估数据生成框架

  RAGEval: Scenario Specific RAG Evaluation Dataset Generation Framework, *ArXiv*, *abs*/2408.01262.

- 大模型工具使用能力评测（Agent 能力）

  T-Eval: Evaluating the Tool Utilization Capability of Large Language Models Step by Step (Chen et al., ACL 2024)

大模型特定任务or能力的评估

- 评估大语言模型在处理有争议知识的问题回答能力

  DEBATEQA: Evaluating Question Answering on Debatable Knowledge, *ArXiv*, *abs*/2408.01419.

DEBATEQA: Evaluating Question Answering on Debatable Knowledge

评估大模型在处理有争议知识的问题回答能力

Field	Content
Question	Does birth order influence personality traits?
Partial Answer 1	**POV** Birth order does not have a meaningful and lasting effect on broad Big Five personality traits. **Explan** The influence of birth order on personality traits has been a topic of interest for over a century. However, based on extensive research combining large datasets from three national panels in the United States, Great Britain, and Germany, it is evident that birth order does not have a meaningful effect on broad Big Five personality traits . . .
Partial Answer 2	**POV** Firstborns score higher on intelligence and intellect. **Explan** Yes, birth order does influence personality traits, particularly in the domain of intelligence and intellect. Research has consistently shown that firstborns tend to score higher on objectively measured intelligence tests . . .
Partial Answer 3	**POV** No birth-order effects on extraversion, emotional stability, agreeableness, or conscientiousness. **Explan** The influence of birth order on personality traits such as extraversion, emotional stability, agreeableness, and conscientiousness has been a topic of interest for over a century. However, recent comprehensive studies have provided substantial evidence that birth order does not significantly impact these personality traits . . .

Motivation

传统的问答基准一般假定答案是固定的，但对于有争议知识的问题，这是不够的。

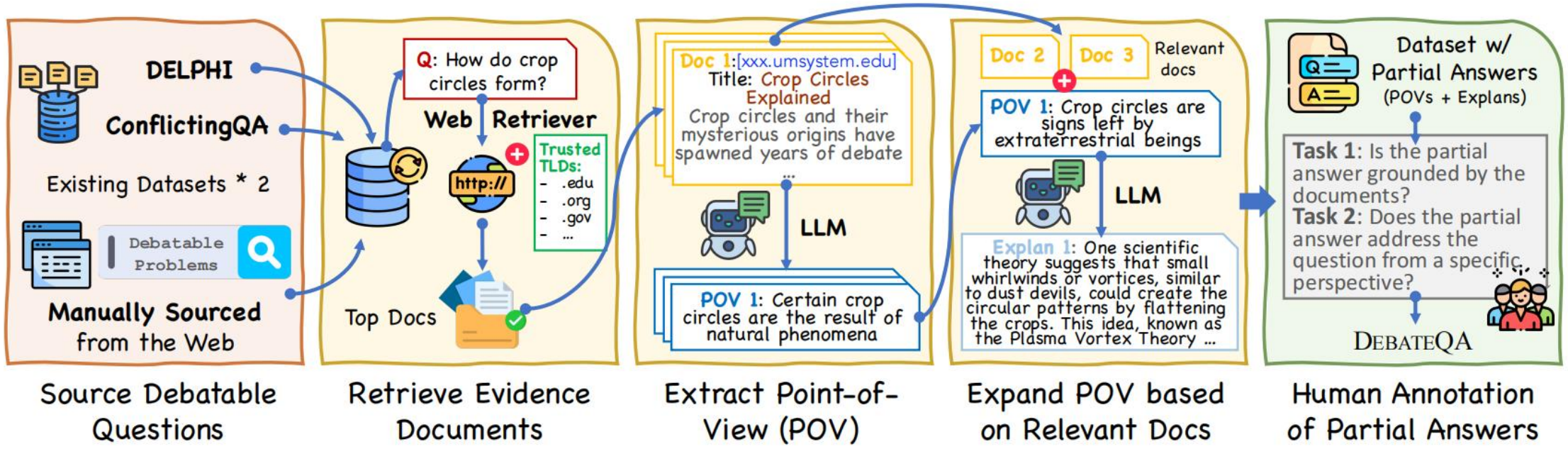
Contribution

- 介绍了 DebateQA 数据集，这是一个包含 2,941 个有争议的问题的数据集，每个问题都附带有多个人工注释的部分答案，涵盖了多种角度。
- 开发了两个指标：①Perspective Diversity-视角多样性，评估视角的全面性；②Dispute Awareness-争议意识，评估 LLM 是否承认这个问题的有争议的本质。
- 评估了12种流行的 LLM 和检索增强生成方法。研究结果显示，虽然 LLM 通常擅长识别有争议的问题，但它们提供包含不同视角的全面答案的能力差异很大。

DEBATEQA: Evaluating Question Answering on Debatable Knowledge, ArXiv, abs/2408.01419.

Method：用什么评估

- **问题收集：**重新利用了 DELPHI 和 ConflictingQA 两个数据集，并从网络上获取了其他有争议的问题，去重后得到3216个问题。
- **答案收集：**核心在于通过将响应与多个部分答案进行比较来评估模型，因此采用三阶段管道从网站中收集证据并拓展为长篇解释。
- **人工注释：**设计了两个注释任务来评估数据集的质量，确保解释基于文档、答案能从某个角度回答问题。



DEBATEQA: Evaluating Question Answering on Debatable Knowledge, ArXiv, abs/2408.01419.

Method：怎么评估

①Perspective Diversity-视角多样性

答案如何很好地从不同的角度提供信息和可信的信息？

$$\text{P.D.} = \sum_{i=1}^{n} \text{PPL}(\text{PA}^i | A)$$

$$\text{PPL}(Y|X) = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log P(y_i|X, y_{<i})\right)$$

$A = \texttt{chatTemplate}(\texttt{concat}(A, \text{“Please restate.”}))$
表示模型对有争议问题的答案

$\text{PA}^i = \texttt{concat}(\text{POV}^i, \text{Explan}^i)$
表示第 i 个部分答案

②Dispute Awareness-争议意识

答案是否承认争议性的存在？

$$\text{D.A.} := \begin{cases} 1 & \text{if } \mathcal{M}_{\text{eval}}(p_{\text{D.A.}}(q, Ans)) \text{ returns “1”,} \\ 0 & \text{otherwise.} \end{cases}$$

很简单，让LLM判断

Experiment

① 语言模型在识别辩论存在方面表现良好，但在提供包含多样观点的综合答案方面表现不一。

② RAG方法虽然不总是有益，但能提高闭源模型的性能，可能是由于更好地利用了上下文信息。

Model	Avg. Len. (#tokens)	Perspective Diversity (P.D.)					Dispute Awareness (D.A.)				
		$\mathcal{M}_{eval}$=Qwen2 0.5B		$\mathcal{M}_{eval}$=GPT-2		Norm. Rank	$\mathcal{M}_{eval}$=Phi-3 M.		$\mathcal{M}_{eval}$=Qwen2 1.5B		Norm. Rank
		Score ↓	Rank	Score ↓	Rank		Score ↑	Rank	Score ↑	Rank	
Closed-Source LLMs											
GPT-4o	434	3.07	1	4.03	1	1	0.952	1	0.979	1	1
GPT-4o mini	252	4.09	6	5.88	6	6	0.937	4	0.964	4	4
GPT-3.5 Turbo	141	5.28	10	8.25	10	10	0.904	6	0.947	6	6
Claude 3.5 Sonnet	199	4.63	8	6.96	8	8	0.856	10	0.920	9	10
Open-Source LLMs (Medium to Large)											
Llama3 70B	432	3.09	2	4.07	3	2=	0.945	3	0.977	2	2=
Llama3 8B	381	3.51	5	5.02	5	5	0.928	5	0.964	4	5
Qwen2 7B	255	4.18	7	6.10	7	7	0.895	8	0.923	8	8
Phi-3 small 128k	412	3.50	4	4.31	4	4	0.899	7	0.924	7	7
Gemma 2 9B	395	3.12	3	4.04	2	2=	0.947	2	0.967	3	2=
Open-Source LLMs (Tiny to Small)											
Qwen2 1.5B	169	5.60	11	8.67	11	11	0.864	9	0.875	10	9
Qwen2 0.5B	72	6.56	12	10.87	12	12	0.792	11	0.836	11	11
Phi-3 mini 128k	218	4.82	9	7.33	9	9	0.716	12	0.794	12	12

Table 4: Main results of P.D. and D.A. for LLMs on DEBATEQA-test. Avg. Len.: average length of the answers, GPT-2: GPT-2 (117M), Phi-3 M.: Phi-3 medium 128k, Norm. Rank: normalized average rank of different $\mathcal{M}_{eval}$. The best and worst results of each metric (*w.r.t.* a specific $\mathcal{M}_{eval}$) are highlighted.

Model	P.D. (↓ $\mathcal{M}_{eval}$=Qwen2 0.5B)		
	No RAG	Vanilla RAG	ReAct
GPT-4o mini	4.02	3.94	**3.70**
Claude 3.5 Sonnet	4.63	4.12	**3.65**
Llama3 8B	**3.55**	4.01	3.99
Qwen2 7B	**3.79**	5.96	5.29
Phi-3 mini 128k	**4.82**	7.01	6.86

Table 6: Effect of two RAG strategies on P.D. scores.

Model	P.D. (↓ $\mathcal{M}_{eval}$=Qwen2 0.5B)	
	Vanilla RAG	RAG w. T. Docs
GPT-4o mini	3.77	**3.63**
Claude 3.5 Sonnet	3.92	**3.54**
Llama3 8B	3.78	**3.62**
Qwen2 7B	5.91	**5.57**
Phi-3 mini 128k	6.77	**6.50**

Table 7: Effect of RAG sources on P.D. scores. RAG w. T. Docs: RAG using trustworthy documents.

DEBATEQA: Evaluating Question Answering on Debatable Knowledge, ArXiv, abs/2408.01419.

总结与展望-未来可能的研究方向

★ **本次报告的不足之处：**缺少对大模型推理能力评估的探讨

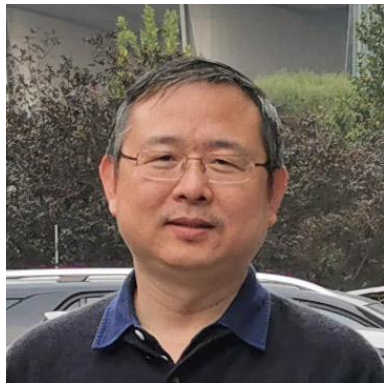
★ **大模型综合评估：**数据集、评测指标已经足够多，期望借鉴人类评估的思想(KoLA)，寻求评估的新视角、新方法；

★ **大模型特定任务、能力的评估：**期望新任务，如辩证性(DebateQA)、价值一致性等新方向；

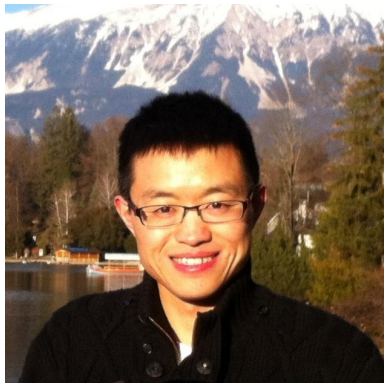
★ **动态评估：**现有的评测方法通常是静态评测，其测试样本总是长时间保持不变。未来可以采用动态评测方法，持续更新测试样本(KoLA)，引入开放式问题；并探索评测新方法，如使用多个大模型通过辩论的方式(KIEval)进行评测；

★ **Agent评估(多智能体评估)：**现有的智能体评测方法大多需要一个特定的环境，并且总是聚焦于智能体的能力。然而，这些方法往往缺乏专门用于评测智能体潜在风险的环境，因此可以进一步增加智能体所处环境的多样性，以便更全面地评估其能力和风险；

LLM评估-国内研究团队

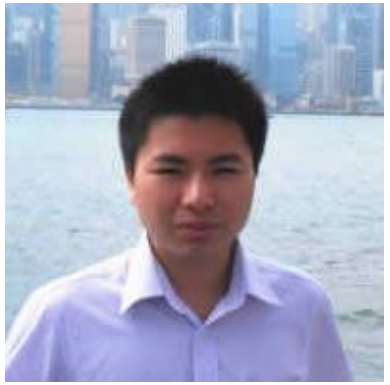



北京大学软件工程国家工程研究中心
知识计算实验室 张世琨/叶蔚
https://se.pku.edu.cn/kcl/

清华大学计算机科学与技术系
知识工程研究室 李娟子/唐杰等
https://keg.cs.tsinghua.edu.cn/


清华大学人工智能研究院
THUNLP 孙茂松/刘知远等
https://nlp.csai.tsinghua.edu.cn/

中国科学院软件研究所
中文信息处理实验室 孙乐/韩先培
https://www.icip.org.cn/zh/homepage/


中国科学院自动化研究所
赵军/刘康

Thank you