

内外知识增强的大模型可信增强

宗成庆

中国科学院自动化研究所

汇报提纲

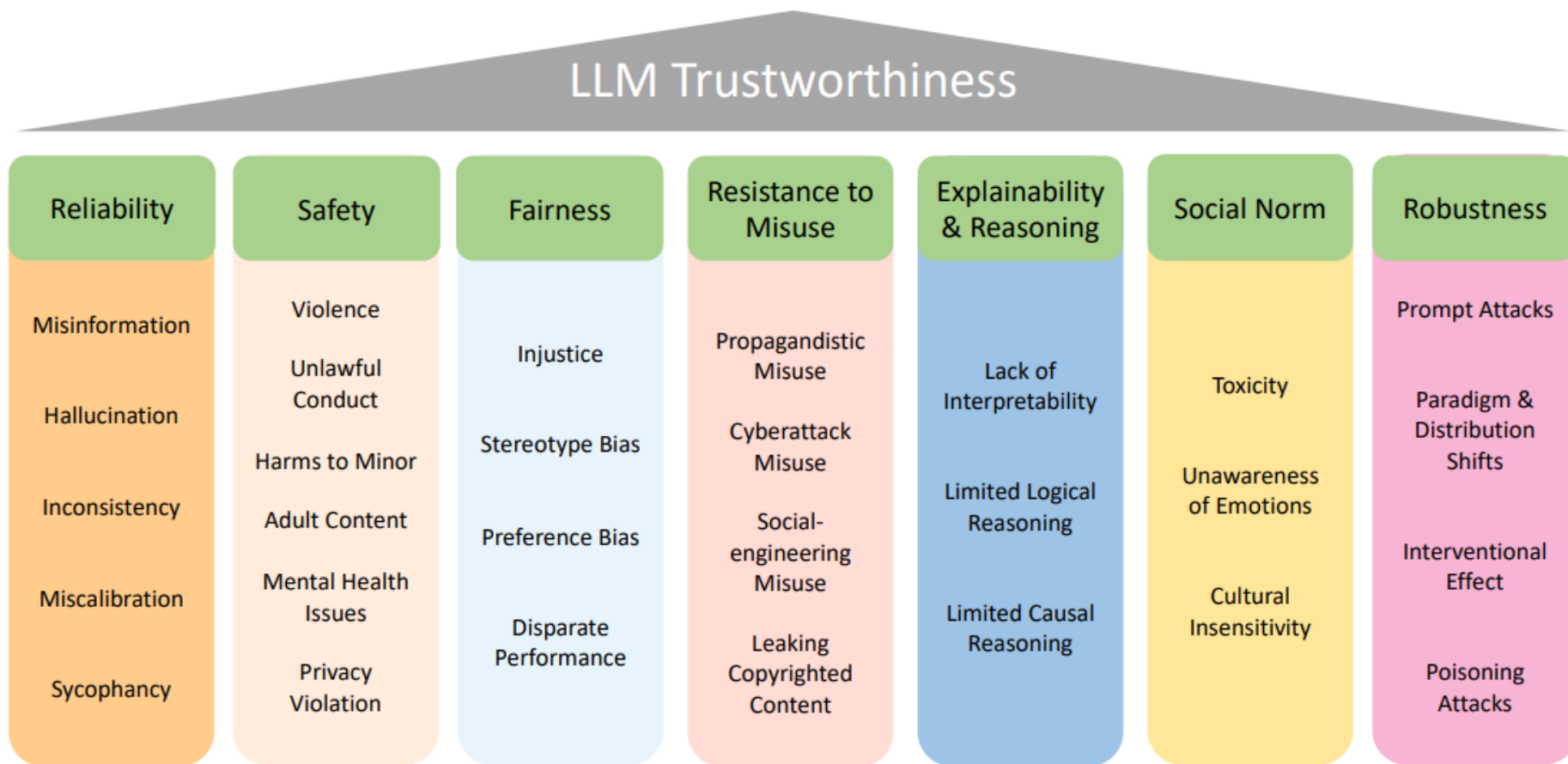
- ➡ 1. 背景简介
- 2. 大模型动态检索增强
- 3. 大模型自省校对方法
- 4. 结束语



大模型的可信度(Trustworthiness)

● 字节：7个维度：

可靠性、安全性、
公平性、
抗误用能力、
可解释推理能力、
遵循社会规范、
鲁棒性[1]。

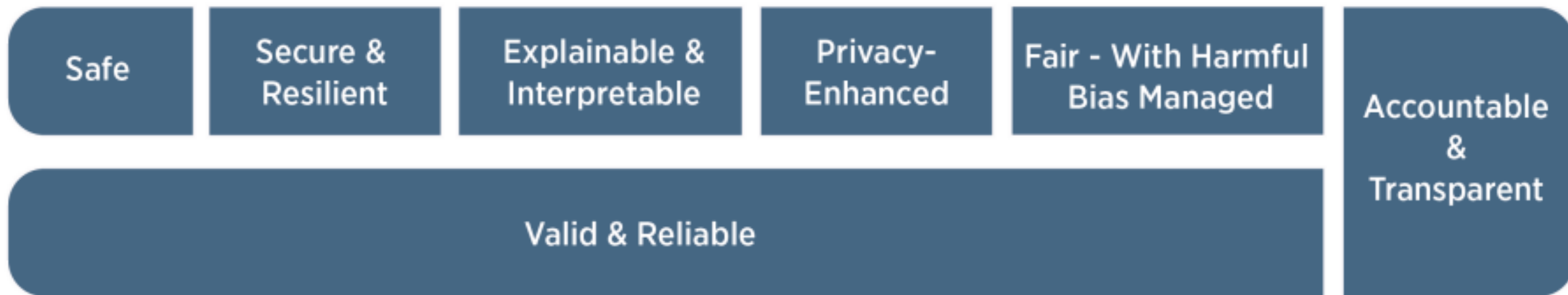




大模型的可信度(Trustworthiness)

- **NIST:** 7个维度:

有效可靠、安全(Safe)、安全稳健(Secure and Resilient)、可问责且透明、可解释、隐私保障、公平[2]。

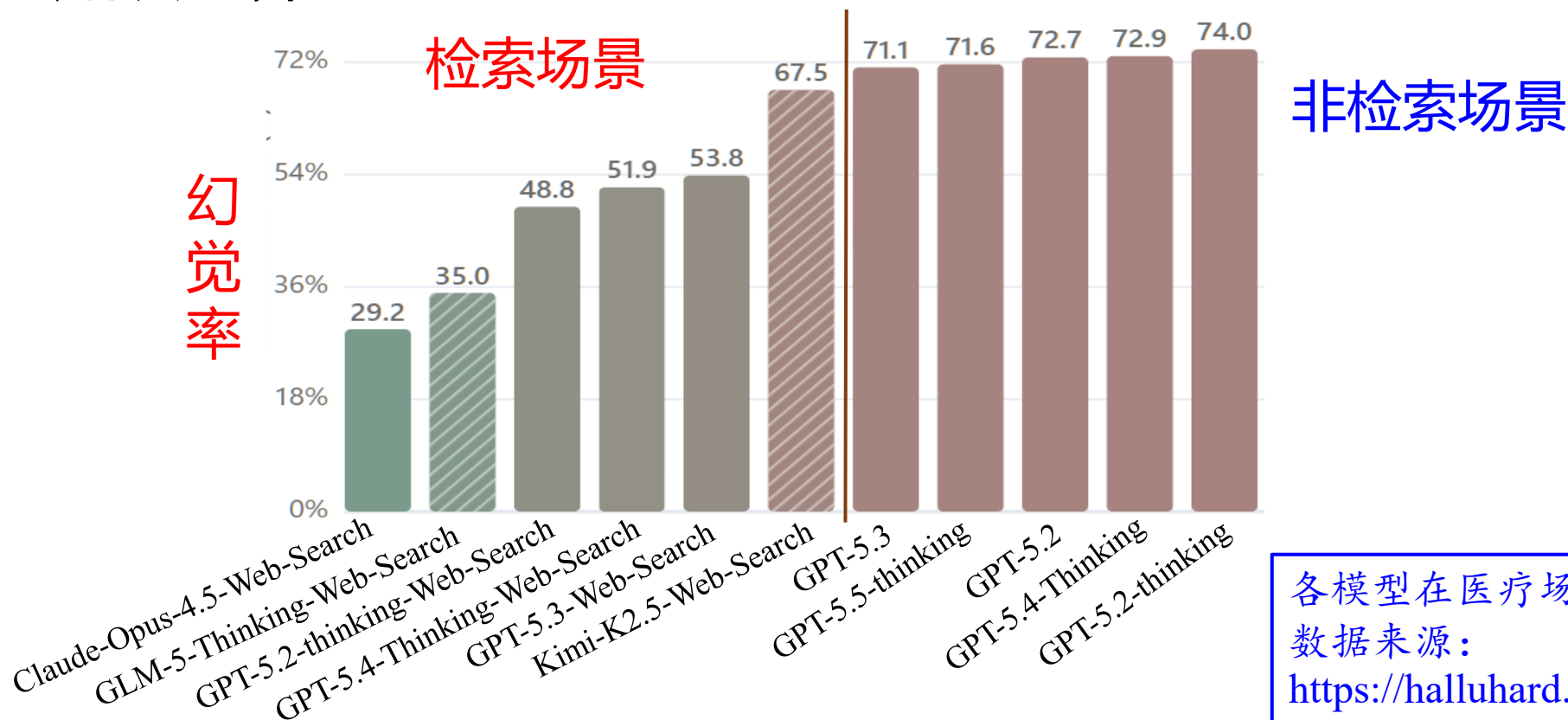


[2] <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>

两大关注点

- 输出内容准确性
- 模型知识边界

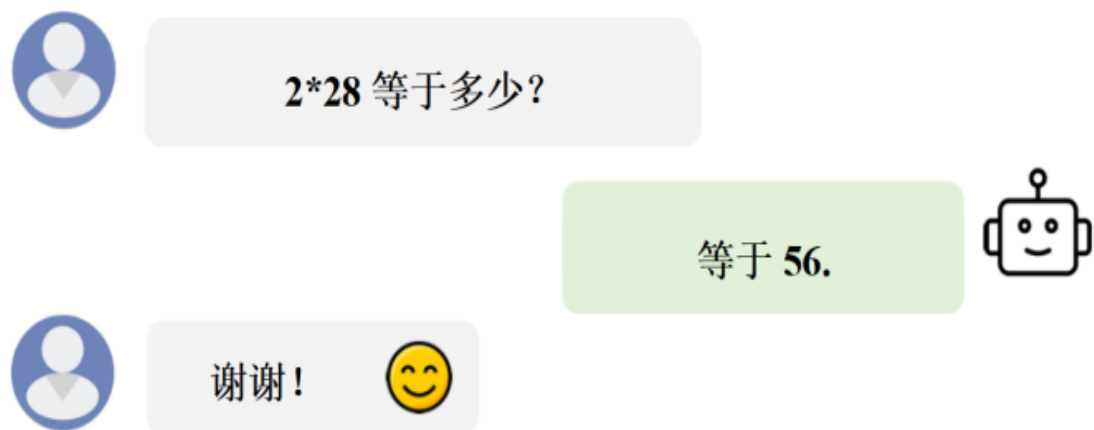
大模型在医疗、法律等真实复杂数据(halluhard 数据集)的幻觉率约为**30%**(检索场景)~**70%**(非检索场景)。



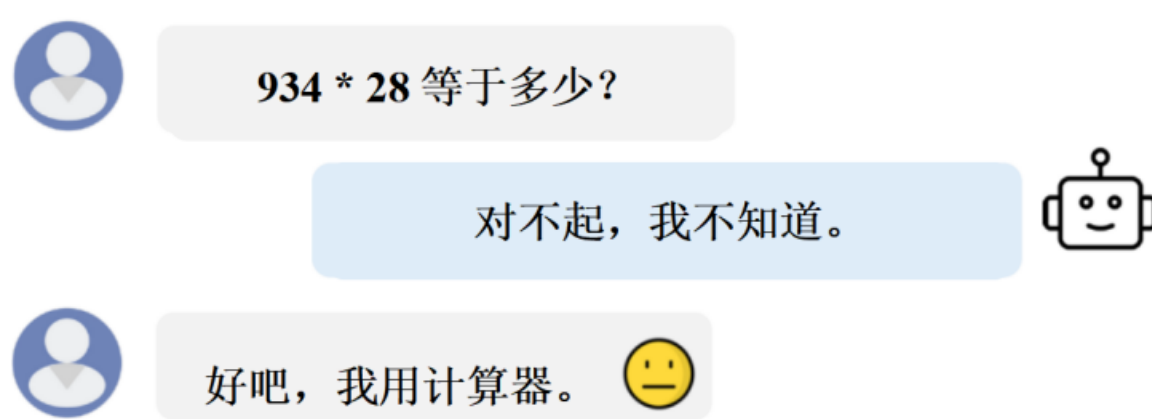
各模型在医疗场景幻觉率
数据来源：
<https://halluhard.com/index.html>

问题思考

- **模型知识边界：**大模型不具备全知能力，存在知识边界。
- **大模型的可信增强应当具备不确定性感知能力：**对于自身知识范围之外的问题，不必强行作答，如实表明未知，或通过检索查证获取准确信息，提升可信度。

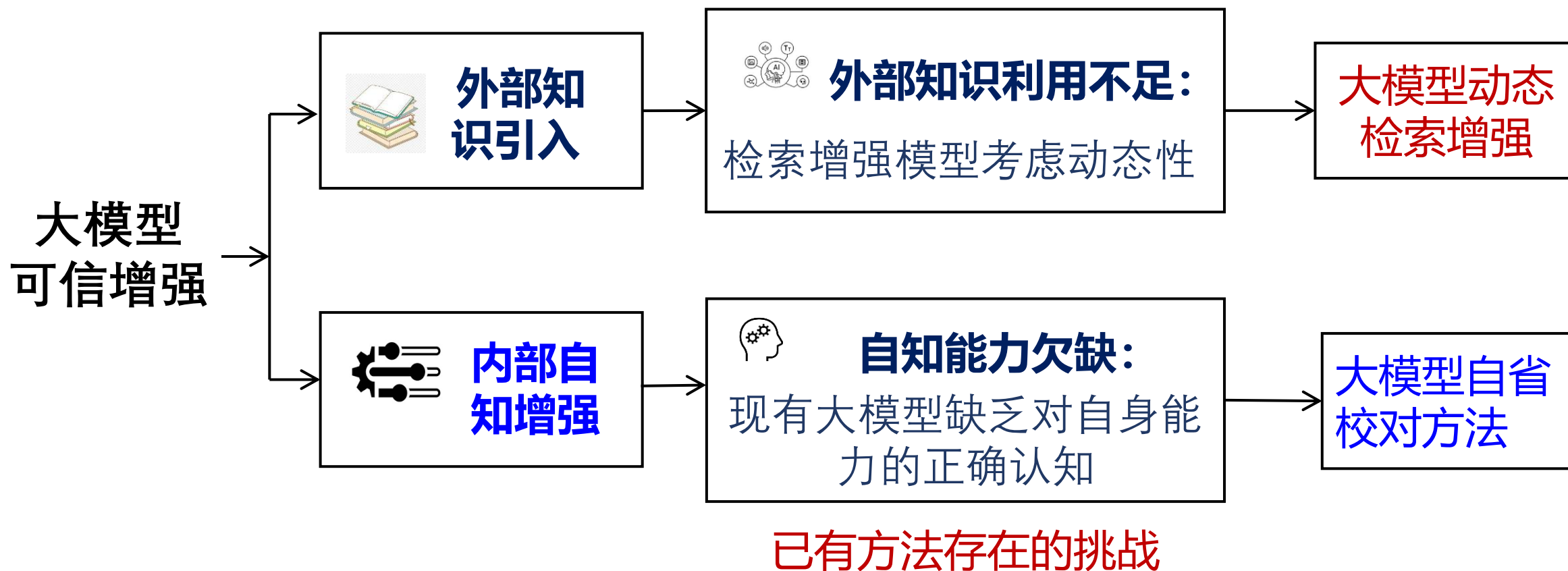


模型认为自己知道答案，正常回答。



模型认为自己不知道答案，拒绝回答。

研究思路





汇报提纲

1. 背景简介

➡ 2. 大模型动态检索增强

3. 大模型自省校对方法

4. 结束语

CoPRM: 协同奖励模型

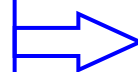
- **问题:** 现有检索智能体存在**检索-回答困境**, 重复搜索成本高; 过早回答质量差。
- **挑战:** 如何在检索过程中动态判断某一步是不是要检索、检索的内容是否合理?

用户问题: 居里夫人出生在哪座城市?

直接回答: 巴黎



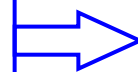
候选搜索1: “居里夫人原名玛丽·居里(Marie Curie), 出生于俄罗斯帝国统治下的**华沙**”。



回答: 华沙



候选搜索2: “波兰裔**法国**物理学家、化学家, 是首位诺贝尔奖女性得主”。



回答: 法国



思路: 在每一步选择更好的动作单元 (即回答或者检索某部分信息) 。

实现过程

Step 1: 生成候选步骤与大模型主观打分;



生成多个状态和候选动作，利用外部大模型判断每个候选动作的评估得分。

Step 2: 结果驱动的客观打分;

Step 3: 校准共识，得到偏好对;

Step 4: 训练奖励模型。

用户问题：居里夫人出生在哪座城市？

GPT-4o /
Qwen-qwq32B

动作P1: Search-居里夫人出生地

动作P2: Answer-华沙

动作P3: Search-国籍

动作P4: Search-居里夫人

.....

$P2 > P1 > P3 > P4$

实现过程

Step 1: 生成候选步骤与大模型主观打分;

Step 2: 结果驱动的客观打分;

Step 3: 校准共识, 得到偏好对;

Step 4: 训练奖励模型。

每个候选动作, 模型采样多个全程轨迹(得到最终答案), 再利用答案F1值作为该步骤的打分。

模型自己采样, 与标准答案对比评估回答步骤。

$P2 > P1 > P3 > P4$



实现过程

Step 1: 生成候选步骤与大模型主观打分;

Step 2: 结果驱动的客观打分;

Step 3: 校准共识, 得到偏好对; ➡

Step 4: 训练奖励模型。

结合外部模型主观打分和结果驱动客观打分，保留一致的偏好对，冲突样本丢弃，得到可信奖励偏好对。

实现过程

Step 1: 生成候选步骤与大模型主观打分;

Step 2: 结果驱动的客观打分;

Step 3: 校准共识, 得到偏好对;

Step 4: 训练奖励模型。

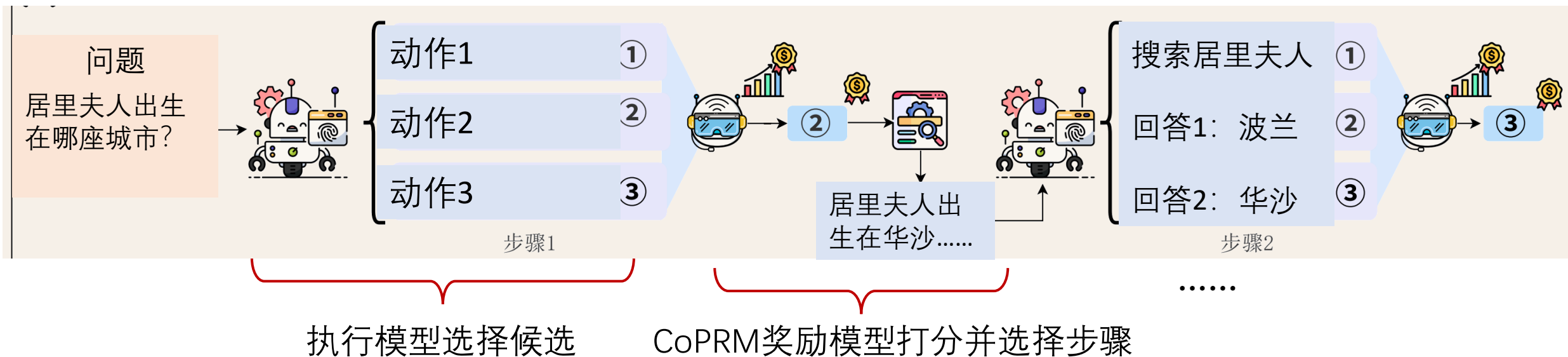
构建了 n 个(1万)偏好对 (s, p^+, p^-) , 采用成对排序损失优化奖励模型, 使得正例步骤得分高于负例步骤。

- ✓ s 是状态(输入+之前动作)
- ✓ p^+ 为较好的选择
- ✓ p^- 为较差的选择

实现过程

● 推理阶段：

- 模型每轮生成 k 个候选过程单元，对搜索步和回答步统一打分，选择 argmax 执行
- 搜索步：更新历史并进入下一轮；
- 回答步：终止并输出答案；减少无效循环和过早终止。





实验：性能对比

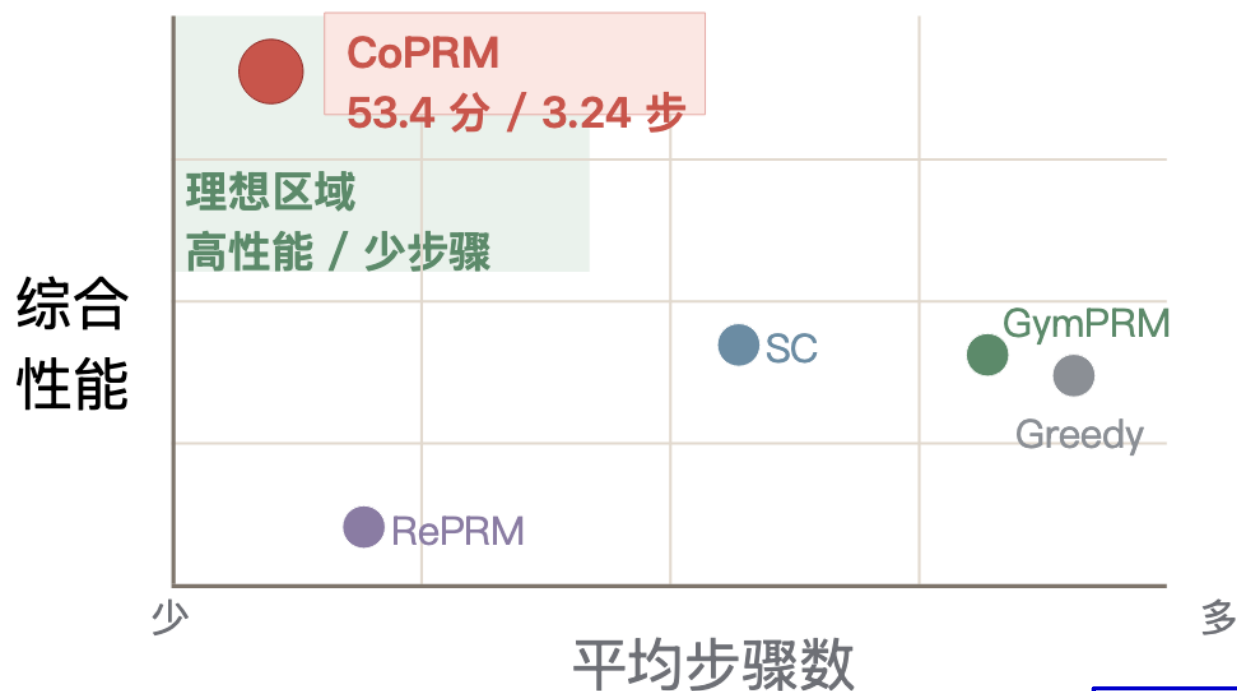
在不同的检索智能体上（提示工程、SFT、RL 等多种训练方法得到的智能体），CoPRM 在均实现了 3.1% ~ 9.6% 的提升。

跨智能体的相对提升

检索智能体	最强基线	CoPRM	提升(%)
PE-ReSearch	40.39	44.25	+9.6
ReSearch	49.34	53.38	+8.2
Search-R1	47.17	49.61	+5.1
R1-Searcher	48.05	49.58	+3.1
RS-ReSearch	50.92	52.70	+3.5

**平均提升
5.9%**

实验：推理步骤的对比



横轴为检索步骤，纵轴为性能得分。
理想区域位于左上方：综合性能高，
平均步骤少。

CoPRM 在多个检索智能体
上平均检索步数减少 12.9%。

X. Zhang et al., CoPRM: A Collaborative Process
Reward Model for Guiding Search Agent
Reasoning. Submitted to EMNLP 2026.



汇报提纲

1. 背景简介

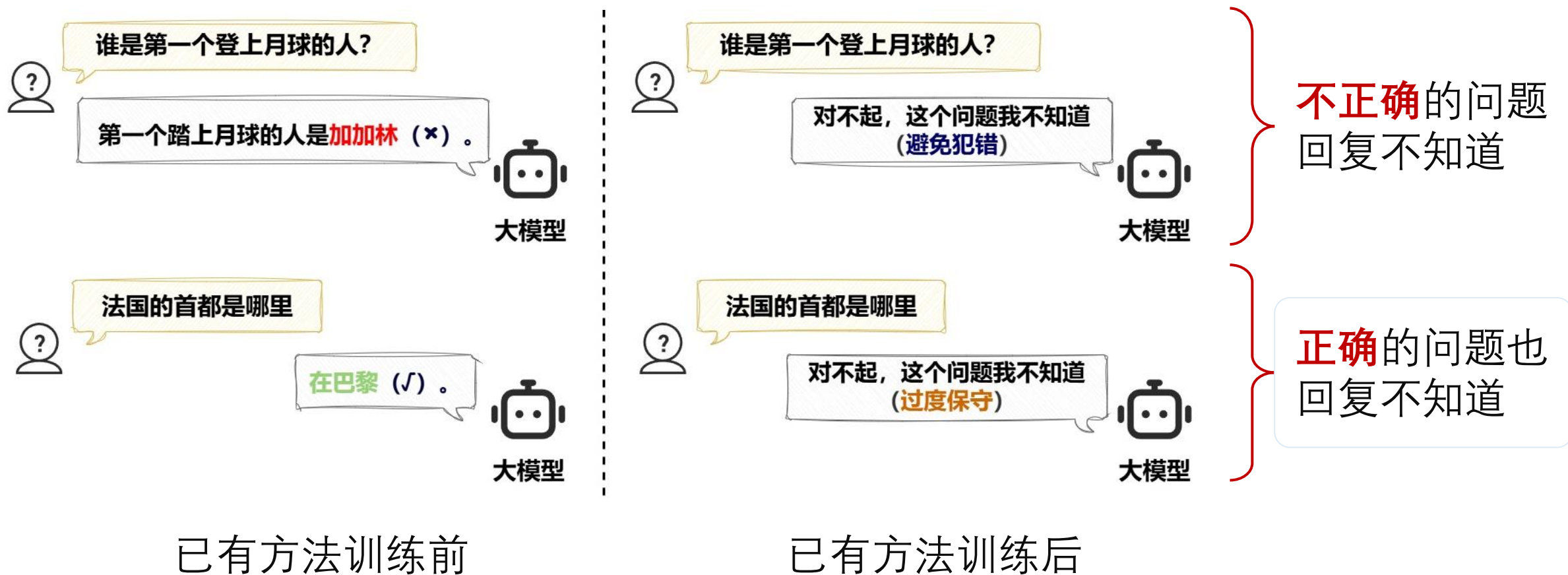
2. 大模型动态检索增强

➡ 3. 大模型自省校对方法

4. 结束语

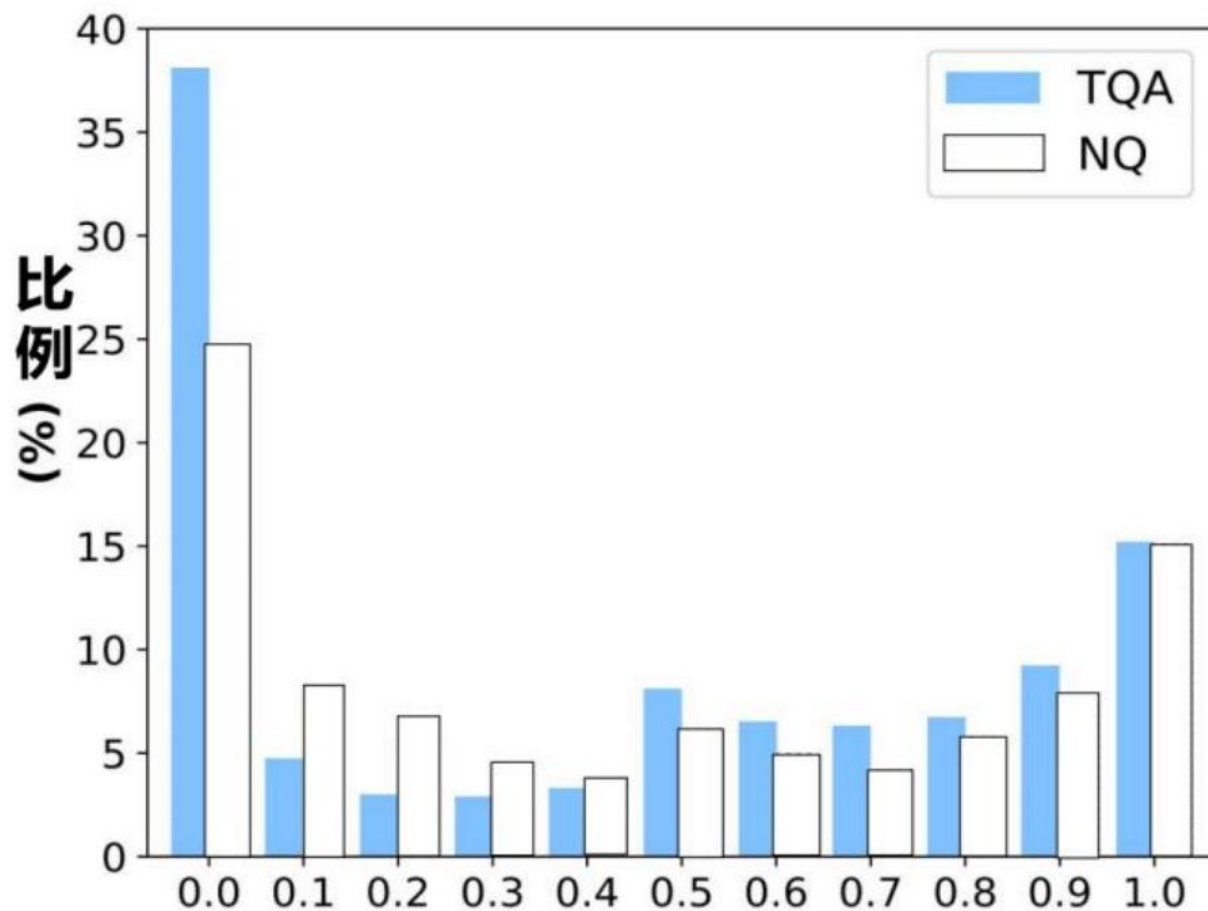
问题提出

已有自省/自知方法往往采样二元分类（知道、不知道）训练模式，模型**过度保守**。



原因分析

已有自省方法忽略了模型“**不确定**”的情况，将置信度 <1 （或者阈值）的问题都视为模型不知道，忽略了不同置信度的区别。



置信度系数CRes分布

Mistral-7B-Instruct 模型在 TQA 和 NQ 数据集的置信度分布



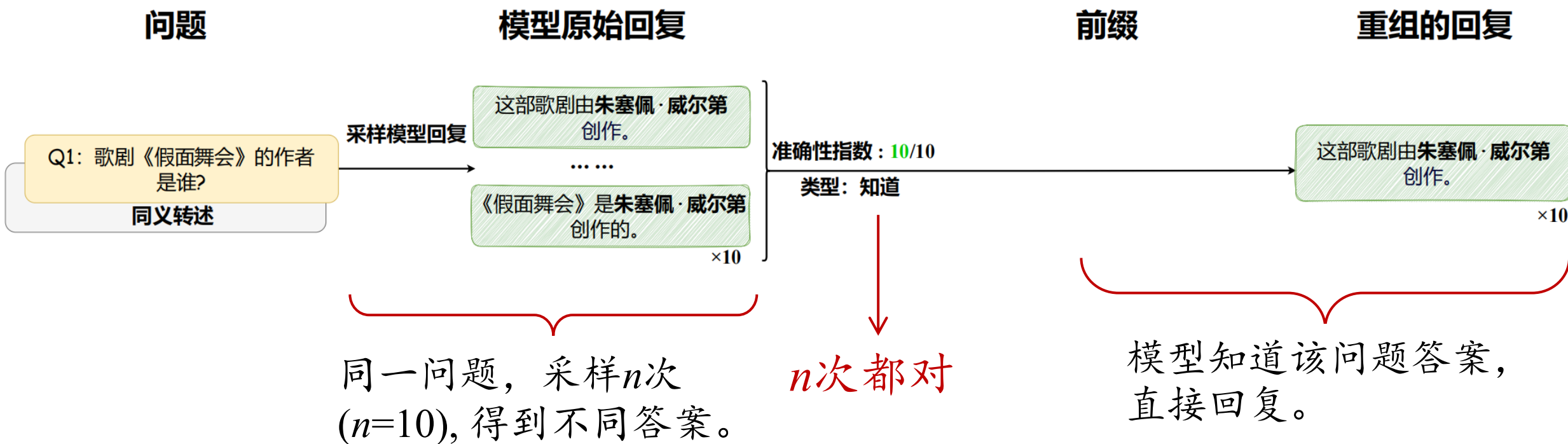
解决问题的思路

按照模型“知道”，“不确定”，“不知道”三类进行划分，并通过前缀区分高置信度、中等置信度和低置信度问题。

类型	CRes 的范围	前缀(子前缀)
知道	1.0	
不确定	[0.1 , 0.9]	我不确定该问题的答案，以下是我的回答。
		+ 我的信心很低。
		+ 我的信心中等。
		+ 我的信心较高。
不知道	0.0	我不知道该问题的答案，以下是我知道的内容。

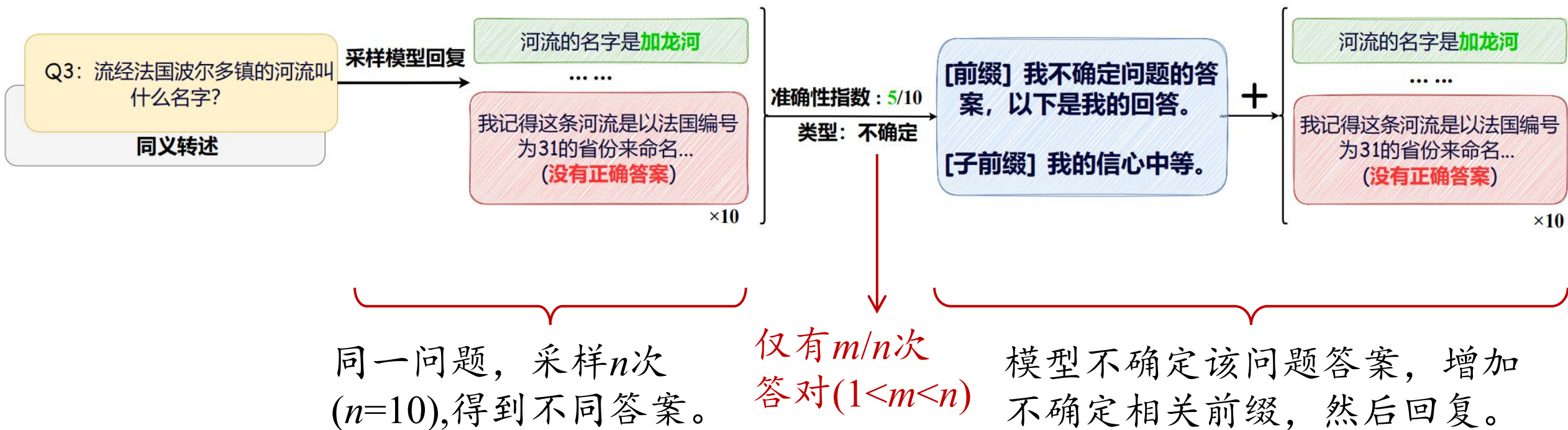
实现方法

- **重组回复**：前缀+完整的原始回复，维持模型原始提供信息的能力，尤其是面对“不知道”或“不确定”问题。



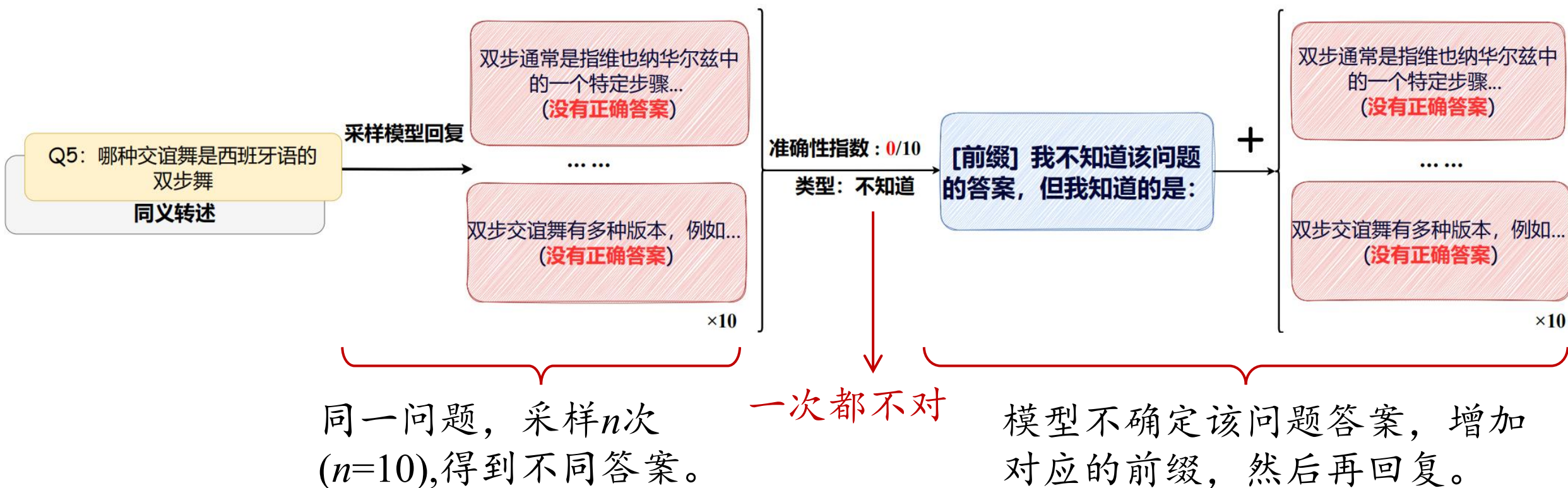
实现方法

- **重组回复：** 前缀+完整的原始回复，维持模型原始提供信息的能力，尤其是面对“不知道”或“不确定”问题。



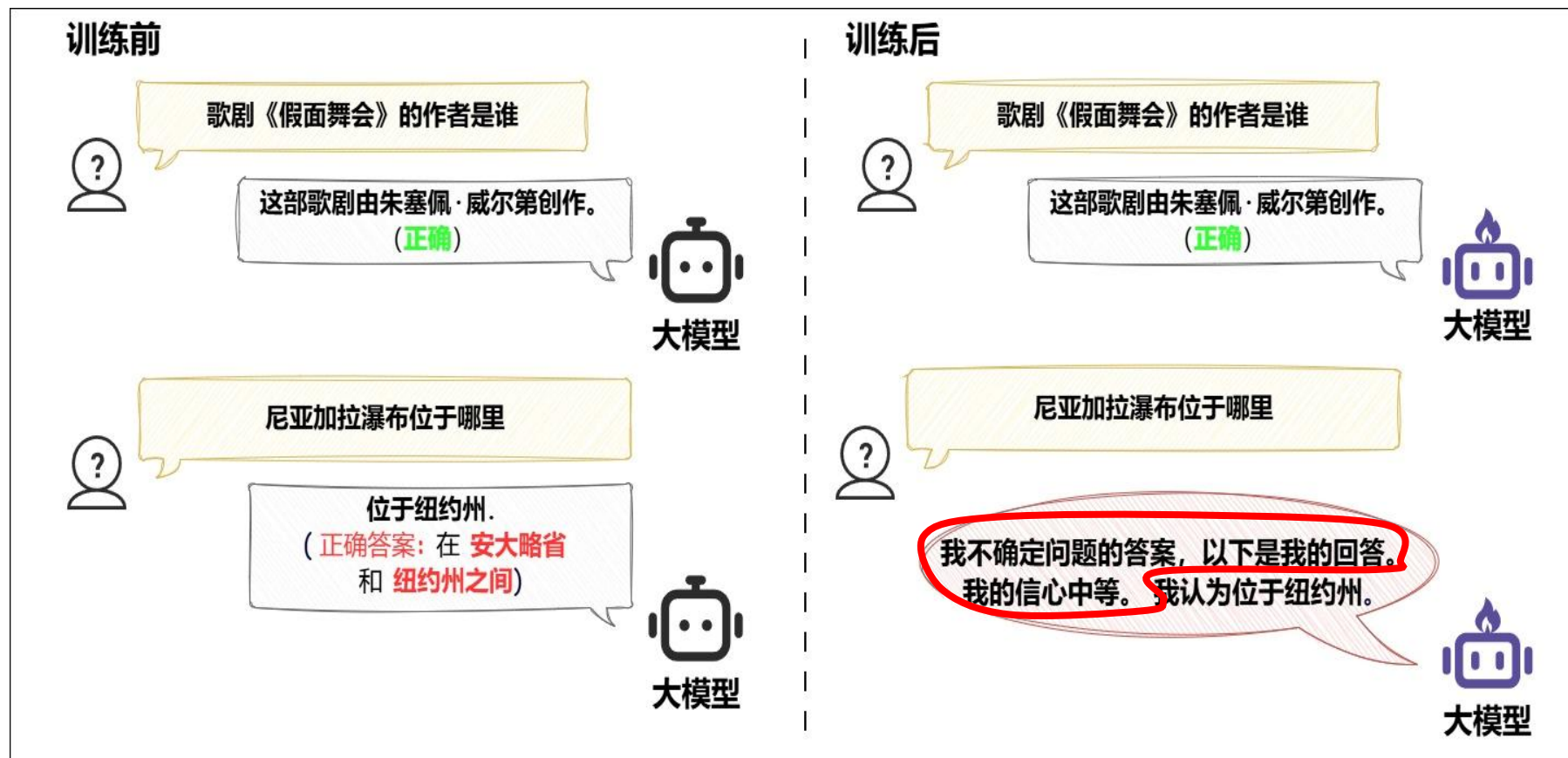
实现方法

- **重组回复**：前缀+完整的原始回复，维持模型原始提供信息的能力，尤其是面对“不知道”或“不确定”问题。



实验验证

- **数据：**基于TriviaQA数据集，模型Mistral-7B/ LLaMA2-7B，构建各5万条的置信度自省训练集。
- **训练：**有监督微调





实验验证

- 在准确性指标 (ACC, F1), 校准效果指标 (ECE), **CIT**方法都取得了最好的效果。

方法类别	方法	TriviaQA			Natural Questions			GSM8K		
		ACC	F1	ECE	ACC	F1	ECE	ACC	F1	ECE
Mistral-7B-Instruct										
无监督方法	I-Prompt	22.72	13.33	0.35	21.28	11.11	0.42	6.02	3.71	0.65
	UEP	20.02	14.55	0.29	20.52	16.10	0.30	23.61	11.00	0.69
	S-Token	21.62	13.78	0.31	21.40	13.59	0.32	20.76	8.04	0.49
	TOPK-S	21.06	16.25	0.27	22.60	20.09	0.28	21.71	10.84	0.63
	CEF	21.32	11.05	0.33	20.80	8.82	0.41	4.91	1.89	0.49
监督训练方法	R-Tuning	21.02	13.25	0.28	21.64	14.91	0.28	20.60	11.67	0.56
	Pf-SFT	26.86	22.06	0.25	25.64	22.85	0.29	6.50	4.76	0.57
	IDK-DPO	27.12	23.97	---	22.52	19.54	---	22.03	14.79	---
CIT*		29.52	28.59	0.14	25.96	24.25	0.22	25.20	14.94	0.34
LLaMA2-7B-Chat										
无监督方法	I-Prompt	20.28	7.37	0.55	20.56	8.20	0.67	1.64	0.65	0.68
	UEP	20.68	12.12	0.26	21.76	15.84	0.42	1.64	0.66	0.92
	S-Token	20.40	14.81	0.42	21.76	13.80	0.32	17.21	6.92	0.40
	TOPK-S	23.34	19.74	0.15	23.28	19.83	0.27	9.02	6.72	0.61
	CEF	22.18	14.86	0.21	21.92	16.52	0.35	24.59	14.93	0.75
监督训练方法	R-Tuning	22.46	15.99	0.19	22.80	17.32	0.29	19.67	8.07	0.64
	Pf-SFT	22.52	15.25	0.23	21.68	13.62	0.38	18.85	13.70	0.48
	IDK-DPO	21.40	13.76	---	23.20	17.79	---	23.77	7.73	---
CIT*		24.22	19.27	0.14	25.60	21.25	0.17	24.59	16.49	0.38



实验验证

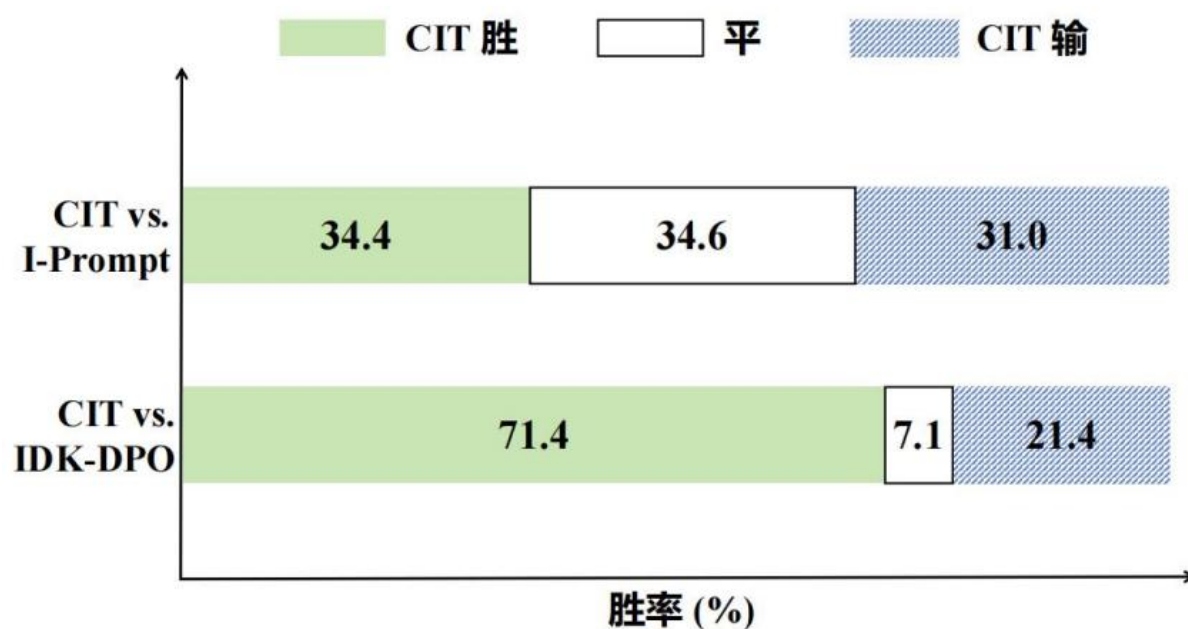
● **CIT**方法在高置信度问题上取得了最好的准确率 (LEM-K, LEM-U)

方法类别	方法	TriviaQA		NQ		TriviaQA		NQ	
		LEM-K	LEM-U	LEM-K	LEM-U	LEM-K	LEM-U	LEM-K	LEM-U
无监督方法	I-Prompt	61.18	66.29	55.51	39.00	44.21	0.00	31.39	0.00
	UEP	62.78	55.97	50.81	44.76	30.66	60.12	30.65	46.49
	S-Token	65.66	51.17	62.26	49.56	59.02	61.37	57.14	58.14
	TOPK-S	56.80	58.82	40.10	52.50	53.85	59.85	13.38	48.39
	CEF	66.39	63.20	58.85	57.92	61.73	66.28	54.52	50.74
监督训练方法	R-Tuning	70.85	28.26	59.81	48.17	27.67	67.56	49.87	57.42
	Pf-SFT	65.63	60.00	57.86	53.00	67.89	56.28	59.52	38.74
	IDK-DPO	69.23	---	27.44	---	71.64	---	69.21	---
	CIT*	86.78	66.86	63.57	60.81	76.77	67.92	72.95	58.51

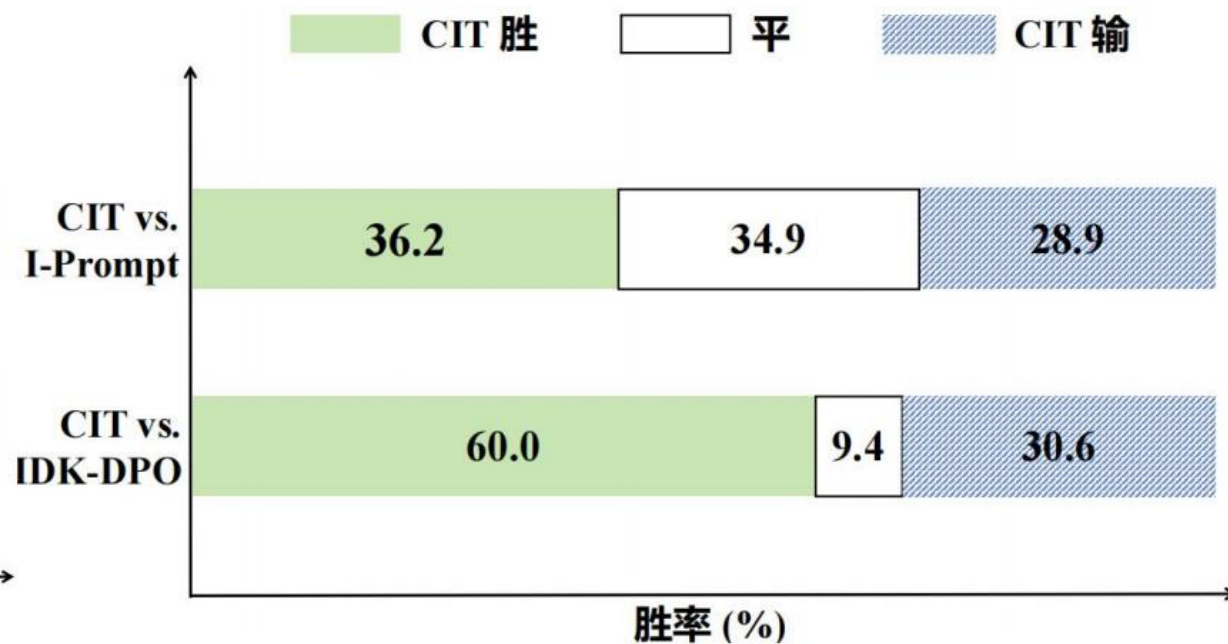


人工评价

● **CIT**方法超过了基线方法，提升了提供实用建议的能力。



(a) TriviaQA 上 CIT 与之前方法的胜率



(b) NQ 上 CIT 与之前方法的胜率

T. Xi et al., Confidence Introspection: A Self-Reflection Method for Reliable and Helpful Large Language Models. *IEEE Transactions on Audio, Speech and Language Processing*, vol. 34, pp. 834-847, 2026.



汇报提纲

1. 背景简介
2. 大模型动态检索增强
3. 大模型自省校对方法

➡ 4. 结束语



可信 + 可靠 + 可控

◆ 外部知识引入 + 内部自知增强

- 解决传统方案存在检索-回答困境：重复检索或过早回答。
- 解决模型自知能力欠缺问题，提高模型的实用性。

知之为知之，不知为不知，是知也。

- 知识边界
- 自我认知

知识源 + 计算模型 + 逻辑推理

可计算、可推断

感谢团队赵阳老师和张兴泉、
席腾霄两位博士生提供的帮助！

谢谢！

Thanks!

