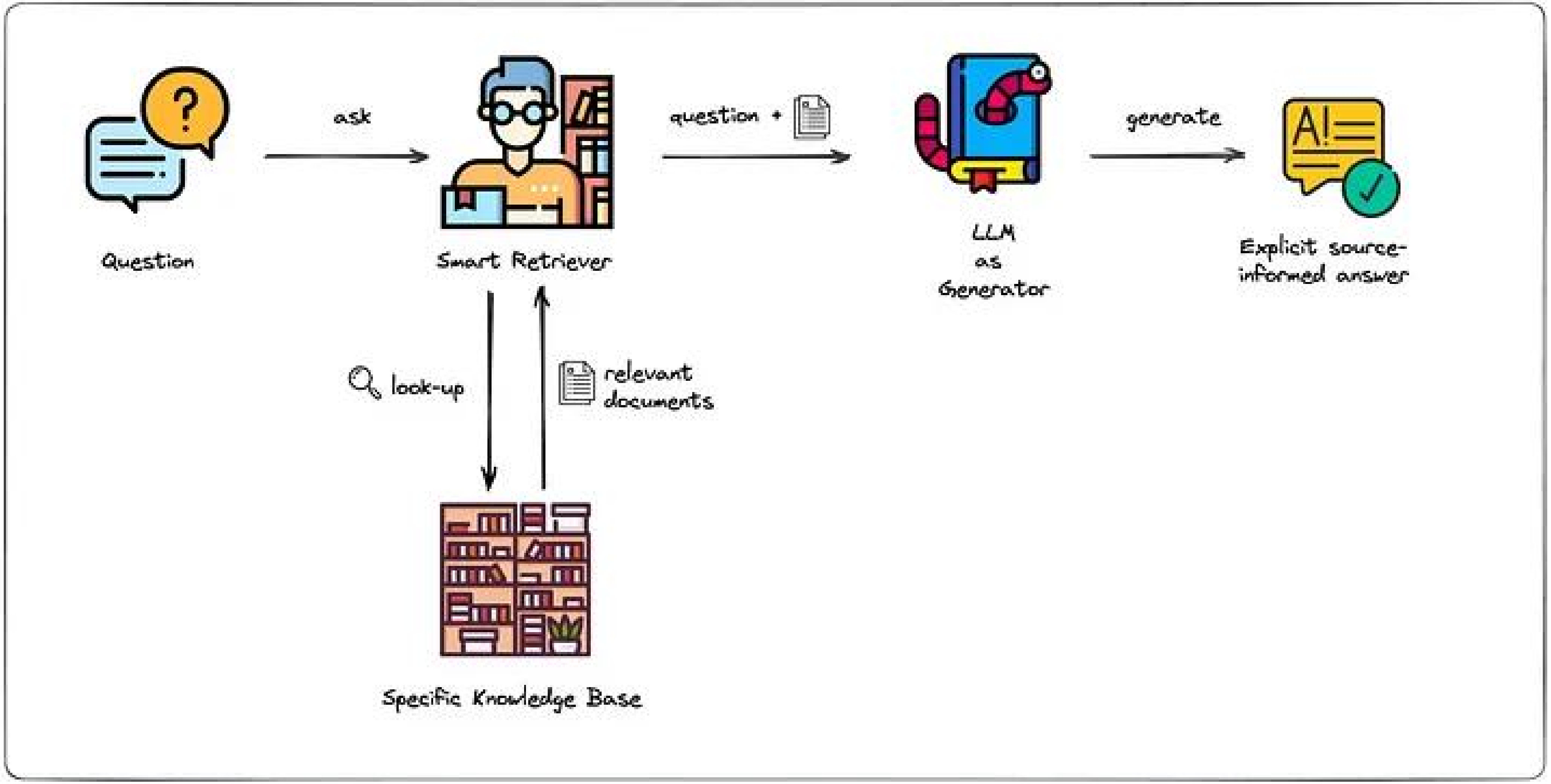
# LLM-as-a-Judge在RAG及合成数据中的价值与应用

讲解人：**徐铖晋 博士 深圳市特聘专家（B类）**

IDEA研究院AI研究科学家
DataArc联合创始人&CTO

The World Needs a Few Good IDEAs

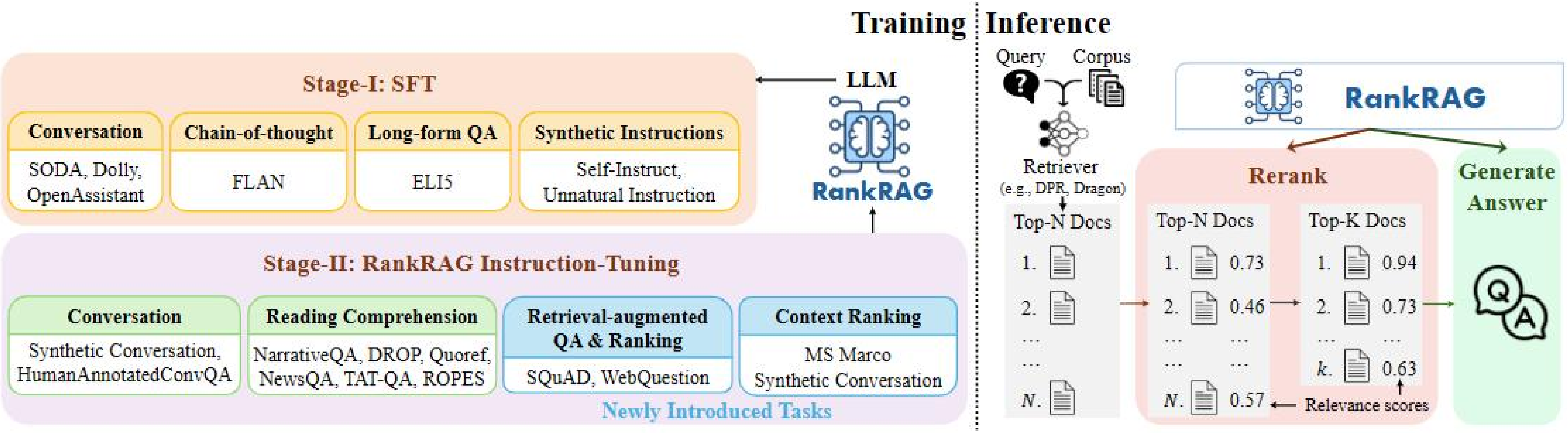
# 检索增强生成（LLM-as-a-generator）



# LLM-as-a-judge in RAG

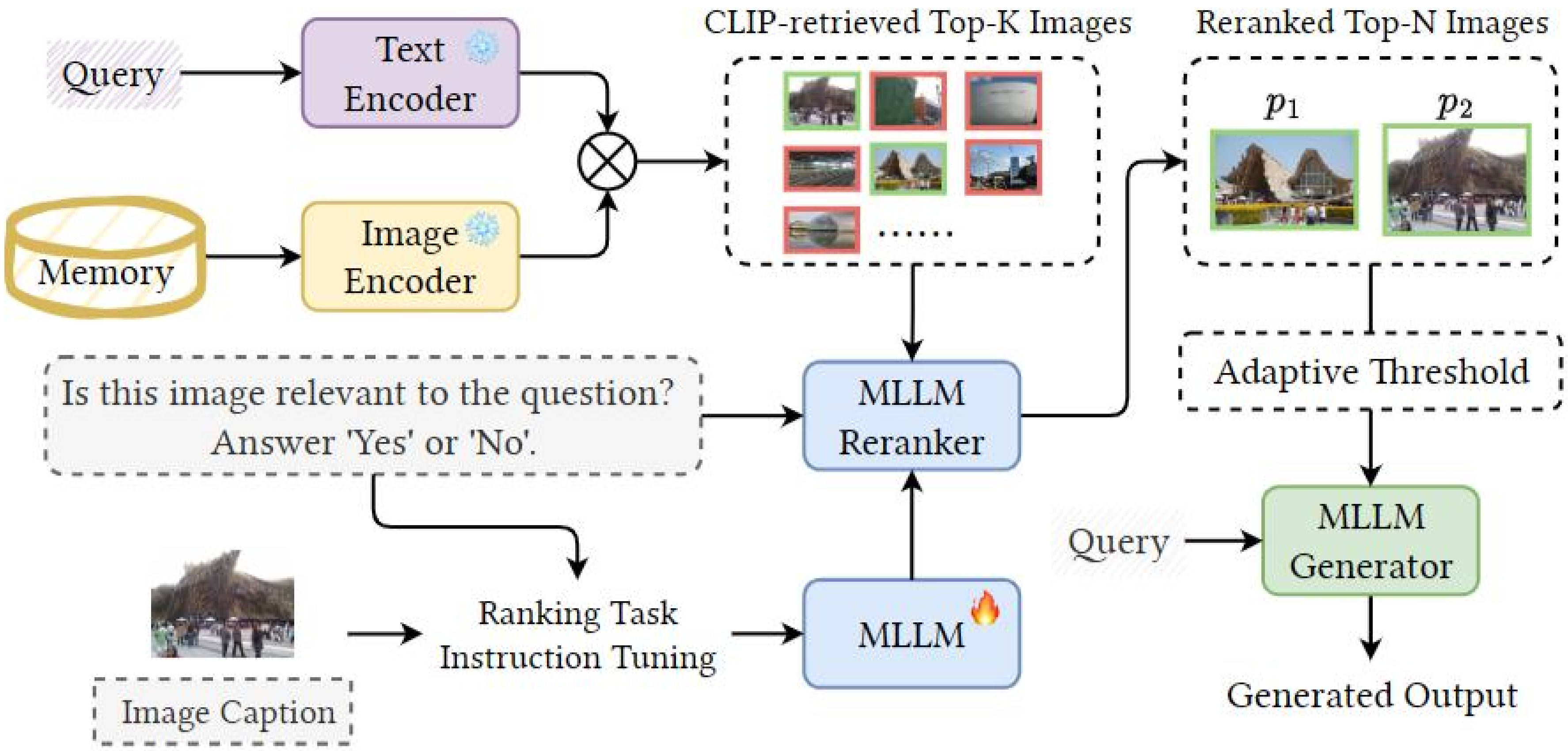
1. 检索信息和问题足够相关吗（Reranker）
2. 检索信息是否足够回答问题（Critic）
3. 回答是否和检索信息一致（Verifier）
4. 对回答质量进行评估（Evaluator）

# LLM-as-a-reranker




System: This is a chat between a user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user' s questions based on the context. The assistant should also indicate when the answer cannot be found in the context.
Passage 1: {(Shuffled) Passage 1}
Passage 2: {(Shuffled) Passage 2}
Passage 3: {(Shuffled) Passage 3}
User: {For the question <question>, access whether the above passages are relevant to the question. Return all the relevant passage id. }
Assistant:

# LLM-as-a-reranker

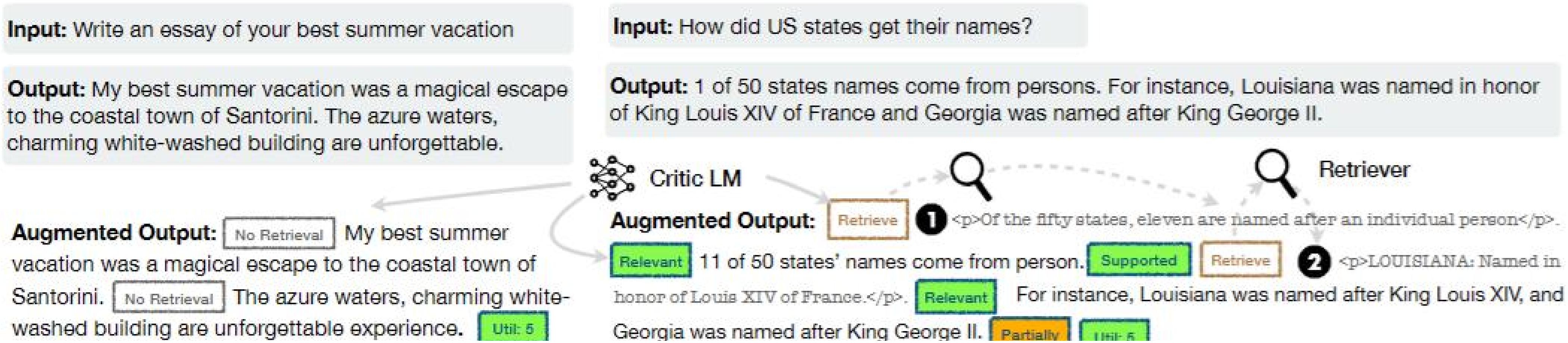



论文：MLLM is a strong reranker: Advancing multimodal retrieval-augmented generation via knowledge-enhanced reranking and noise-injected training

# LLM-as-a-reranker

| Methods | MultimodalQA | | | WebQA | | |
|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@10 | R@2 | R@5 | R@10 |
| CLIP-ViT-L/14-336px | **84.78** | 94.35 | 95.65 | 57.10 | 71.96 | 84.86 |
| *w/ SFT* | 83.04 | 94.35 | 94.78 | 55.09 | 73.23 | 81.94 |
| Vis-BGE-base | 49.57 | 74.78 | 82.61 | 28.78 | 43.62 | 54.56 |
| Vis-BGE-m3 | 43.48 | 66.52 | 72.17 | 26.69 | 40.75 | 51.14 |
| InternVL-C | 82.17 | **95.65** | 96.96 | **64.90** | **81.22** | 88.09 |
| InternVL-G | 82.17 | 95.22 | **97.39** | **64.90** | 80.23 | **88.28** |
| *Reranking Top-K from CLIP-ViT-L/14-336px* | | | | | | |
| LLaVA-v1.5-13B | 72.61 | 90.87 | 95.22 | 45.35 | 65.87 | 80.56 |
| *w/ caption-aware* IT | **98.26** | **98.26** | **98.26** | 79.74 | 88.14 | 89.77 |
| mPLUG-Owl2 | 67.83 | 87.39 | 93.91 | 43.26 | 63.80 | 79.38 |
| *w/ caption-aware* IT | 90.87 | 96.09 | 97.39 | 71.27 | 85.08 | 88.97 |
| Qwen-VL-Chat | 68.26 | 89.57 | 92.61 | 47.64 | 67.22 | 80.42 |
| *w/ caption-aware* IT | 91.30 | 95.65 | 97.39 | 80.12 | 88.53 | **89.96** |
| InternVL2-1B | 47.39 | 84.78 | 93.91 | 34.99 | 57.49 | 74.72 |
| *w/ caption-aware* IT | **98.26** | **98.26** | **98.26** | **82.00** | 88.78 | 89.94 |
| InternVL2-2B | 66.52 | 88.70 | 93.91 | 42.79 | 62.48 | 77.97 |
| *w/ caption-aware* IT | **98.26** | **98.26** | **98.26** | 81.91 | **88.94** | 89.94 |
| *Reranking Top-K from Different Retrievers* | | | | | | |
| LLaVA-v1.5-13B | | | | | | |
| *w/ Vis-BGE-base* | 88.70 | 88.70 | 88.70 | 59.61 | 64.71 | 65.70 |
| *w/ Vis-BGE-m3* | 84.78 | 84.78 | 84.78 | 57.57 | 62.26 | 63.03 |
| *w/ InternVL-C* | **98.70** | **98.70** | **98.70** | **82.08** | **90.79** | **92.72** |
| *w/ InternVL-G* | 97.83 | 97.83 | 97.83 | 81.91 | 90.24 | 92.31 |

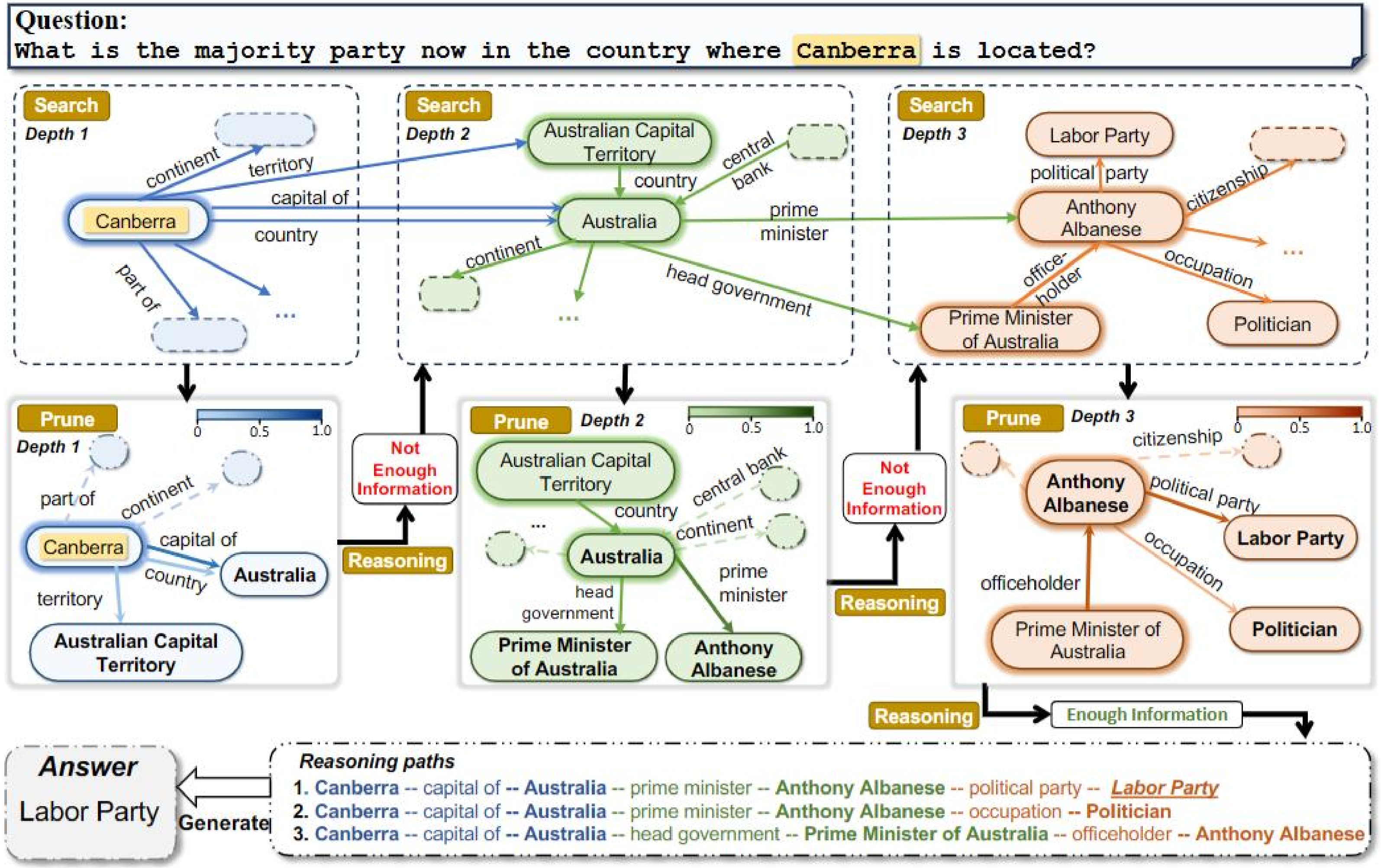
# Critic: 是否需要检索




| Type | Input | Output | Definitions |
|---|---|---|---|
| Retrieve | $x$ / $x, y$ | {yes, no, continue} | Decides when to retrieve with $\mathcal{R}$ |
| IsRel | $x, d$ | {**relevant**, irrelevant} | $d$ provides useful information to solve $x$. |
| IsSup | $x, d, y$ | {**fully supported**, partially supported, no support} | All of the verification-worthy statement in $y$ is supported by $d$. |
| IsUse | $x, y$ | {**5**, 4, 3, 2, 1} | $y$ is a useful response to $x$. |

论文：Self-rag: Learning to retrieve, generate, and critique through self-reflection （ICLR 2024）

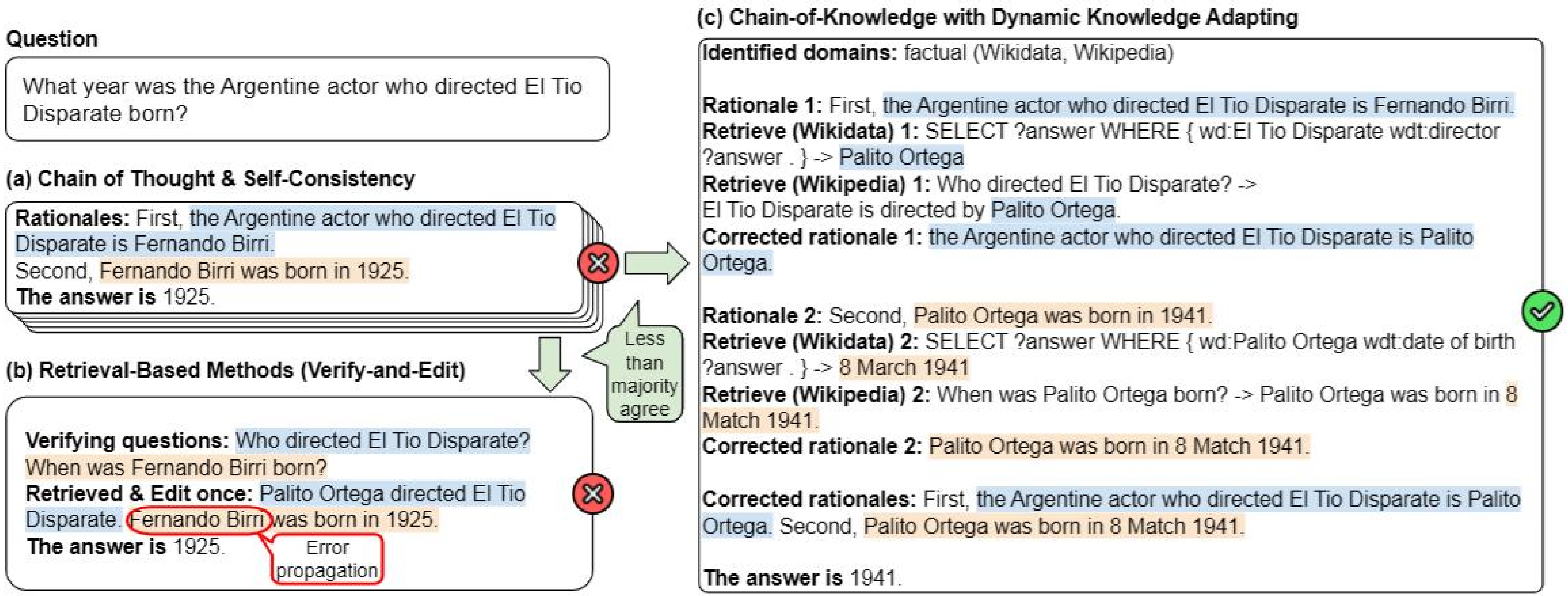
# Critic: 是否需要检索




论文：Think-on-Graph: Deep and responsible reasoning of large language model on knowledge graph （ICLR 2024）

Think-on-Graph 2.0: Deep and Faithful Large Language Model Reasoning with Knowledge-guided Retrieval Augmented Generation

# Verifier: 回答是否和检索信息一致




论文：Chain-of-Knowledge: Grounding Large Language Models via Dynamic Knowledge Adapting over Heterogeneous Sources （ICLR 2024）

# Evaluator: 回答是否和标准答案一致

Query: 谁赢得了2024年的美国总统选举?

Context: 唐纳德·川普在2024年总统大选中取得令人瞩目的胜利，于11月6日击败副总统卡玛拉·哈里斯....

Response: 唐纳德·川普

Golden Answer: 唐纳德·特朗普

Exact Match：False

LLM Eval: True

# Evaluator: 评估哪个答案更好

---Role---
You are an expert tasked with evaluating two answers to the same question based on four criteria: **Comprehensiveness**, **Diversity**, and **Empowerment.**

---Goal---
You will evaluate two answers to the same question based on four criteria: **Comprehensiveness, Diversity, and Empowerment.**

- **Comprehensiveness**: How much detail does the answer provide to cover all aspects and details of the question?
- **Diversity**: How varied and rich is the answer in providing different perspectives and insights on the question?
- **Empowerment**: How well does the answer help the reader understand and make informed judgments about the topic?

**For each criterion, choose the better answer (either Answer 1 or Answer 2) and explain why. Then, select an overall winner based on these three categories.**

***Evaluation Instruction Prompt***

Here is the question: {query}

Here are the two answers: **Answer 1**: {answer1}; **Answer 2**: {answer2}

Evaluate both answers using the three criteria listed above and provide detailed explanations for each criterion.

Output your evaluation in the following **JSON format**:

```
{{
    "Comprehensiveness": {{ "Winner": "[Answer 1 or Answer 2]", "Explanation": "[Provide explanation here]" }},
    "Diversity": {{ "Winner": "[Answer 1 or Answer 2]", "Explanation": "[Provide explanation here]" }},
    "Empowerment": {{ "Winner": "[Answer 1 or Answer 2]", "Explanation": "[Provide explanation here]" }},
    "Overall Winner": {{ "Winner": "[Answer 1 or Answer 2]", "Explanation": "[Summarize why this answer is the overall winner based on the three criteria]" }}
}}
```

***Evaluation Input Prompt***

论文：Lightrag: Simple and fast retrieval-augmented generation

# 数据-人工智能时代的石油

idea

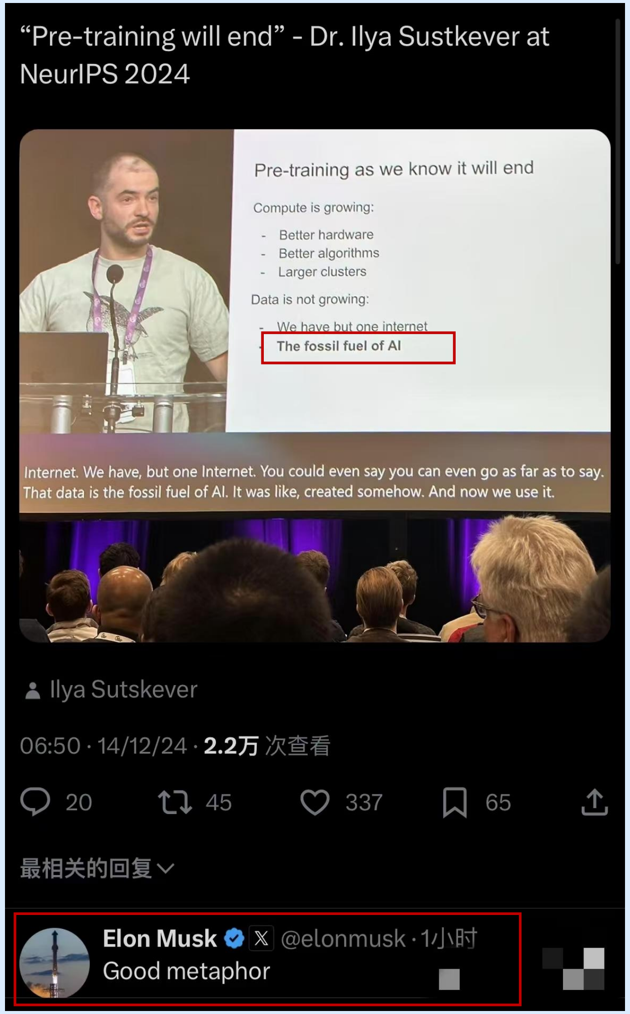
## 石油

BP《世界能源统计年鉴》（BP Statistical Review of World Energy），基于已探明石油储量与年产量的比例，原油将在至少50年后才可能被耗尽[1]

[1] BP. (2023). BP Statistical Review of World Energy 2023. https://www.bp.com/en/global/corporate/energy-economics/statistical-review-of-world-energy.html.

## 数据 多久被耗尽?



# 合成数据成为持续不断获得高质量数据的新范式

## 破局点：合成数据

**是什么：**通过算法和数学模型构建，模拟真实数据中的统计属性和关系。

如果把数据比作AI的“石油”，那么合成数据就是AI的“可再生能源”。

**英伟达**开源3400亿参数大模型使用**98%**合成数据训练。

（Adler B, Agarwal N, Aithal A, et al. Nemotron-4 340B Technical Report[J]. arXiv preprint arXiv:2406.11704, 2024.）

**OpenAI** CEO Sam Altman曾公开表达他相信很快所有数据都会是合成数据。

(The promise and perils of synthetic data., Kyle Wiggers Oct 13, 2024)

**ScaleAI** CEO Alexandr Wang指出合成数据的生成需要像对待芯片生产一样看待。

(A16Z: Unlocking AI’s Future: Alexandr Wang on the Power of Frontier Data David George and Alexandr Wang)

# Synthetic Data & LLM-as-a-judge

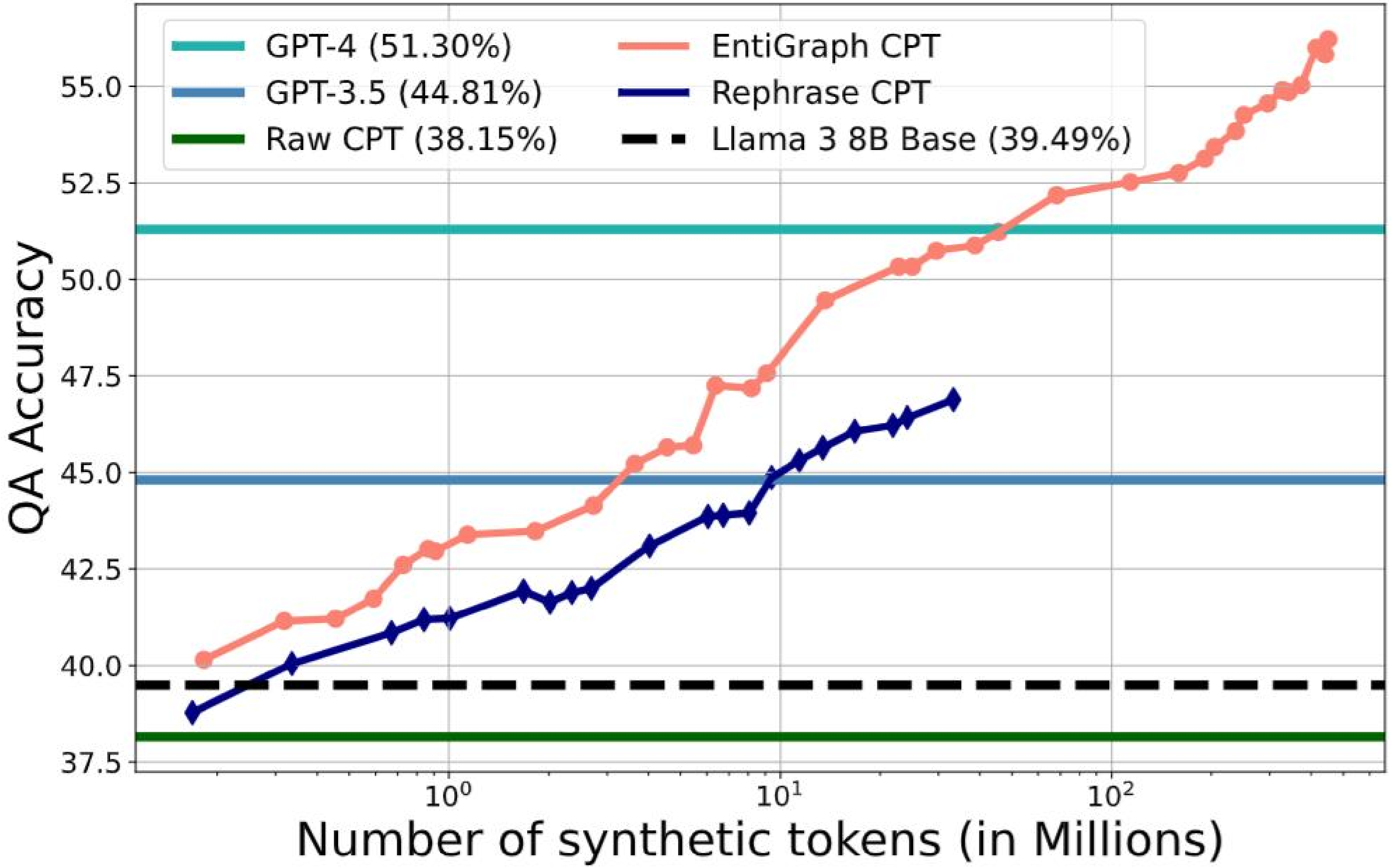

# Judge: 使用模型直接评估合成数据质量

| Indicator | Notation | Explanation |
|---|---|---|
| Length | $Len$ | The average length of every response in the dataset. |
| Reward score | $Rew$ | The average reward model inference score of every pair in the dataset. (Köpf et al., 2023) |
| Perplexity | $PPL$ | The exponentiated average negative log-likelihood of response. |
| MTLD | $MTLD$ | Measure of Textual Lexical Diversity McCarthy and Jarvis (2010) |
| KNN-i | $KNN_i$ | Distance to approximate $i^{th}$-nearest neighbors (Dong et al., 2011) in SentenceBERT(Reimers and Gurevych, 2019) embedding space. |
| Unieval-naturalness | $Nat$ | The score of whether a response is like something a person would naturally say, provided by the UniEval (Zhong et al., 2022) dialogue model. |
| Unieval-coherence | $Coh$ | The score of whether this response serves as a valid continuation of the previous conversation, provided by the UniEval (Zhong et al., 2022) dialogue model. |
| Unieval-understandability | $Und$ | The score of whether the response is understandable, provided by the UniEval (Zhong et al., 2022) dialogue model. |

Table 1: Natural language indicators for instruction quality evaluation. Every data example is viewed as a pair of instruction(input) and response(output). Unless otherwise specified, the indicator value on a dataset is the average value of each sample on the indicator.

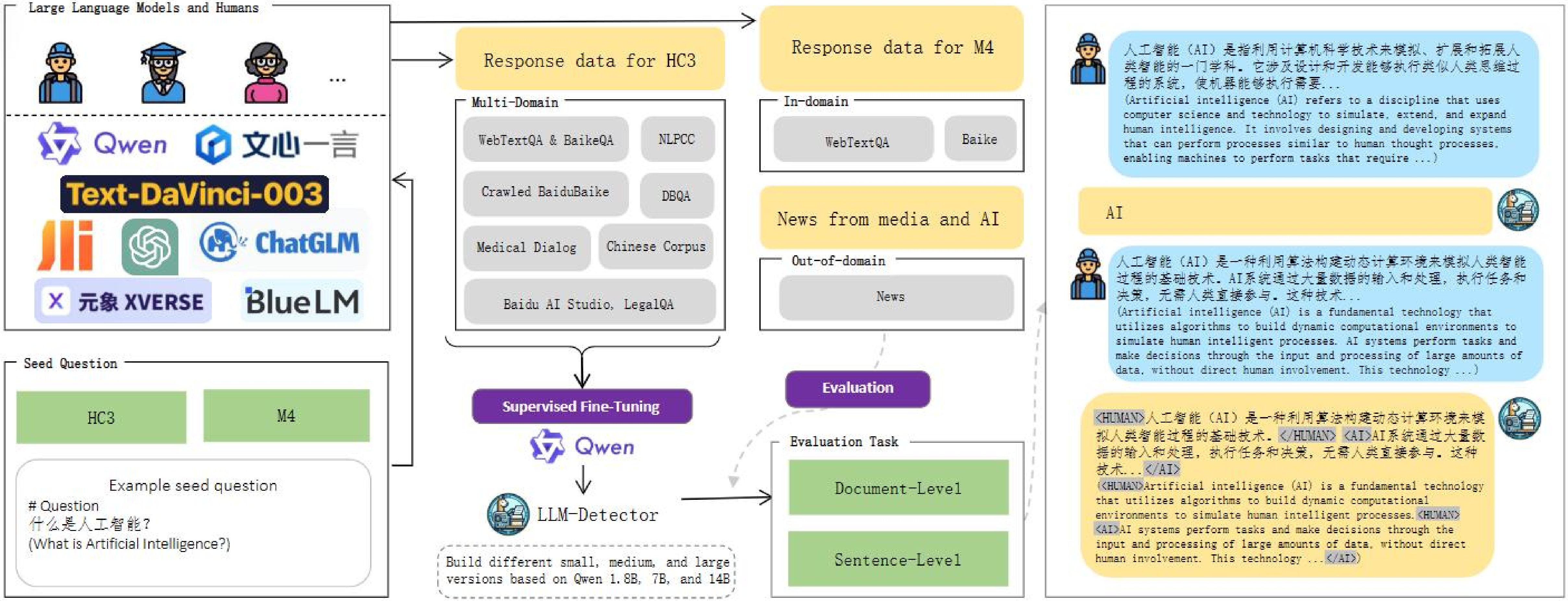
论文：Instruction Mining: High-Quality Instruction Data Selection for Large Language Models

# Judge: 通过模型训练的效果评估合成数据质量



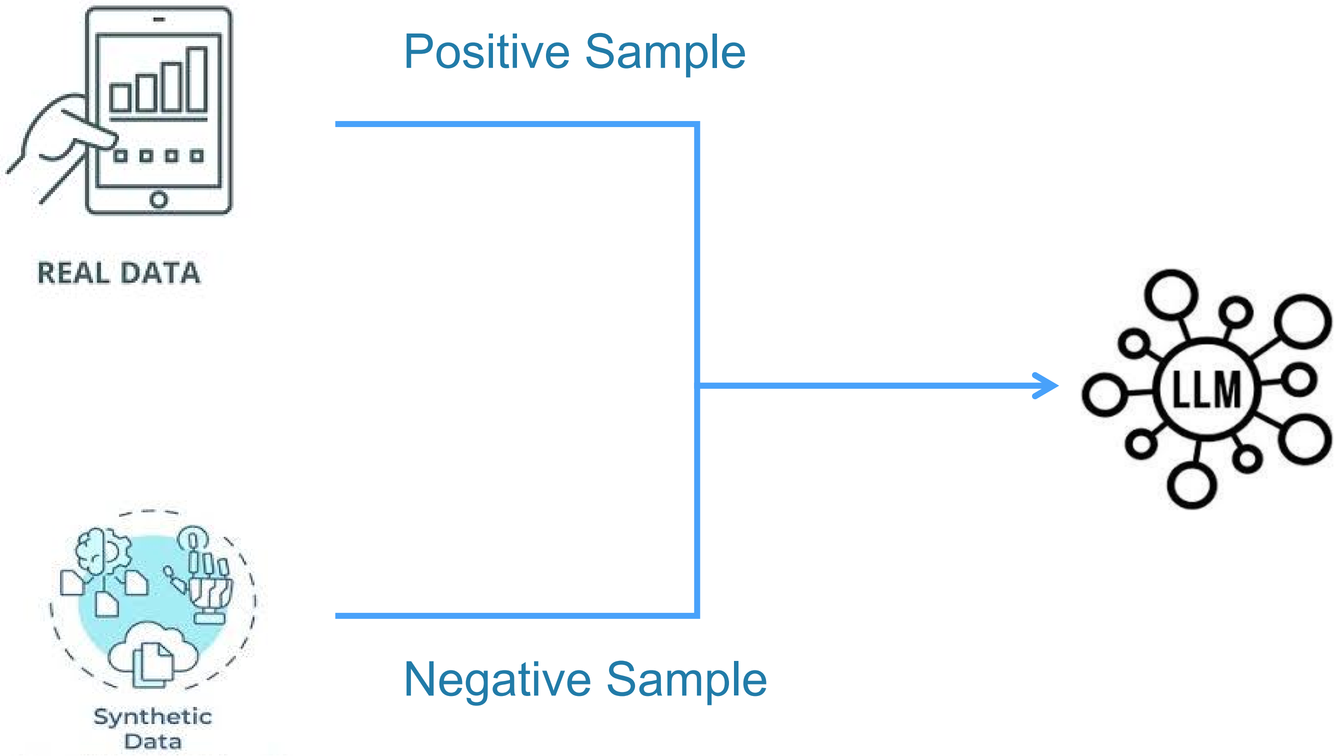

论文：Synthetic Continued Pretraining

# Judge: 通过模型分辨真实数据和合成数据

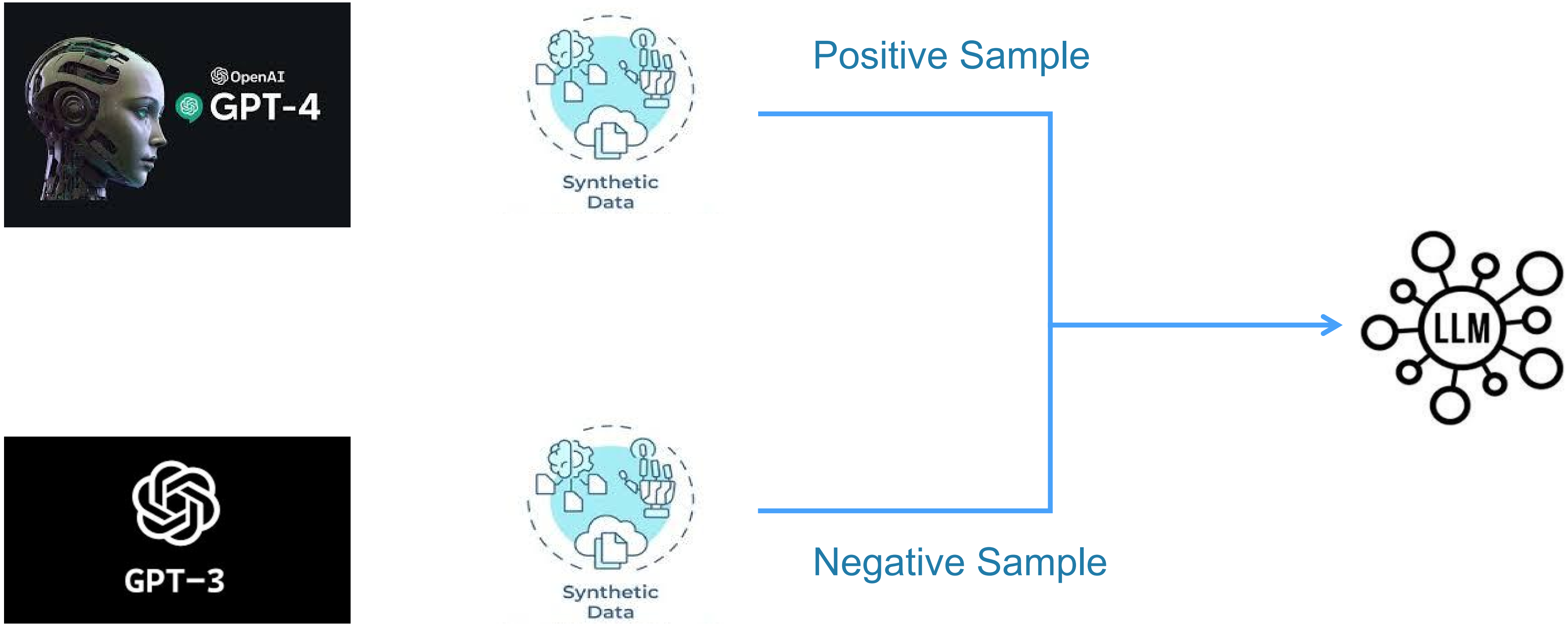



论文：LLM-Detector: Improving AI-Generated Chinese Text Detection with Open-Source LLM Instruction Tuning

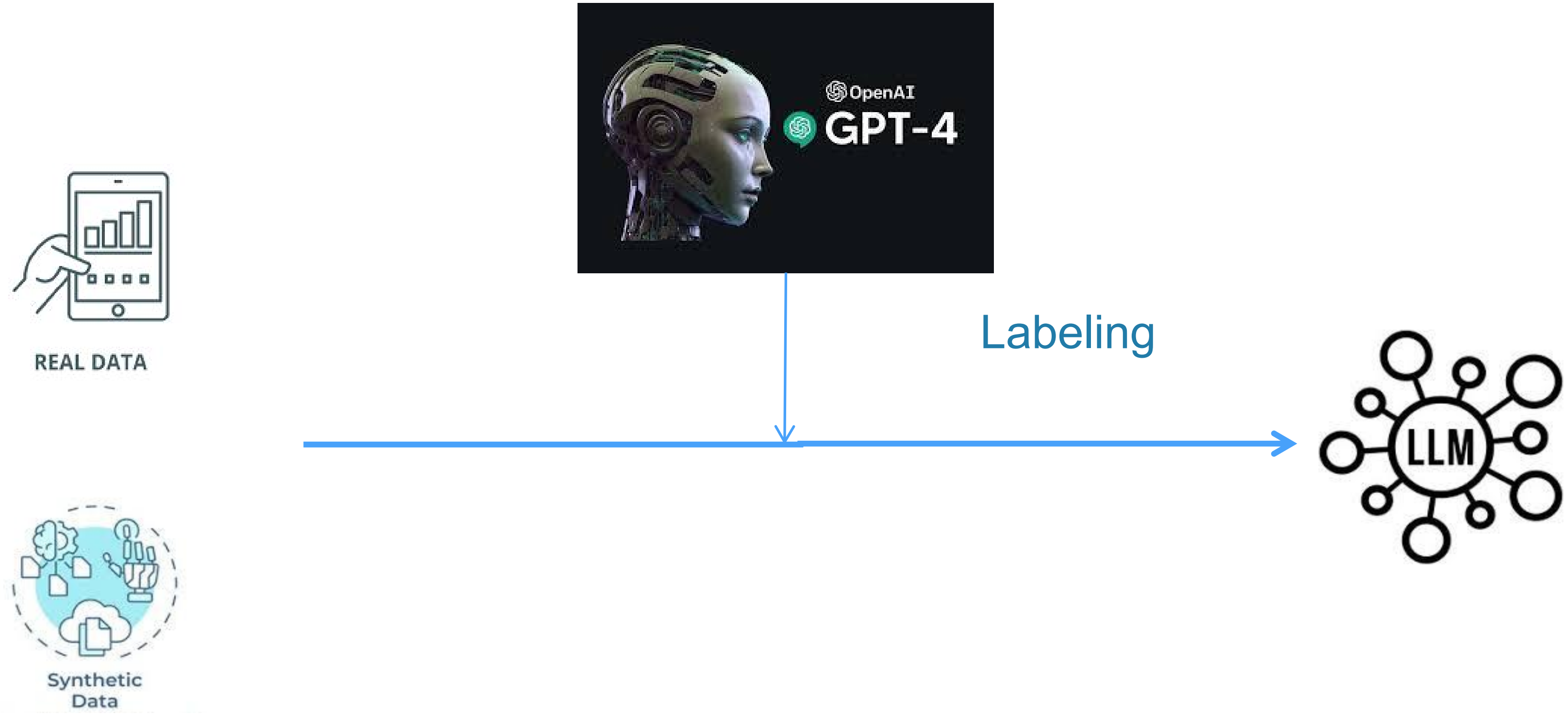
# 用合成数据训练LLM-as-a-judge：蒸馏



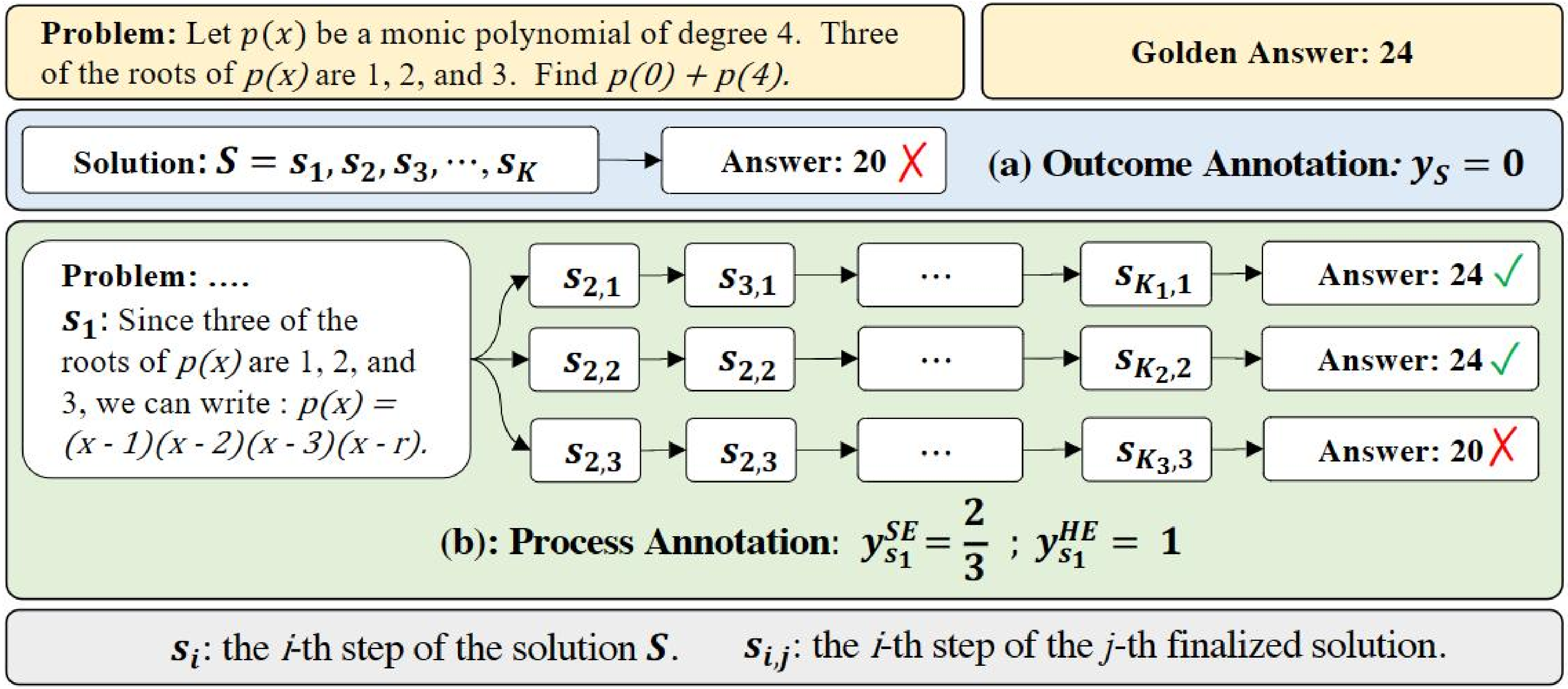
# 用合成数据训练LLM-as-a-judge：蒸馏



# 用合成数据训练LLM-as-a-judge：蒸馏



# 用合成数据训练LLM-as-a-judge：蒙特卡洛 inspired by AlphaGo




论文：MATH-SHEPHERD: VERIFY AND REINFORCE LLMS STEP-BY-STEP WITHOUT HUMAN ANNOTATIONS

# 用合成数据训练LLM-as-a-judge：蒙特卡洛 + GAN

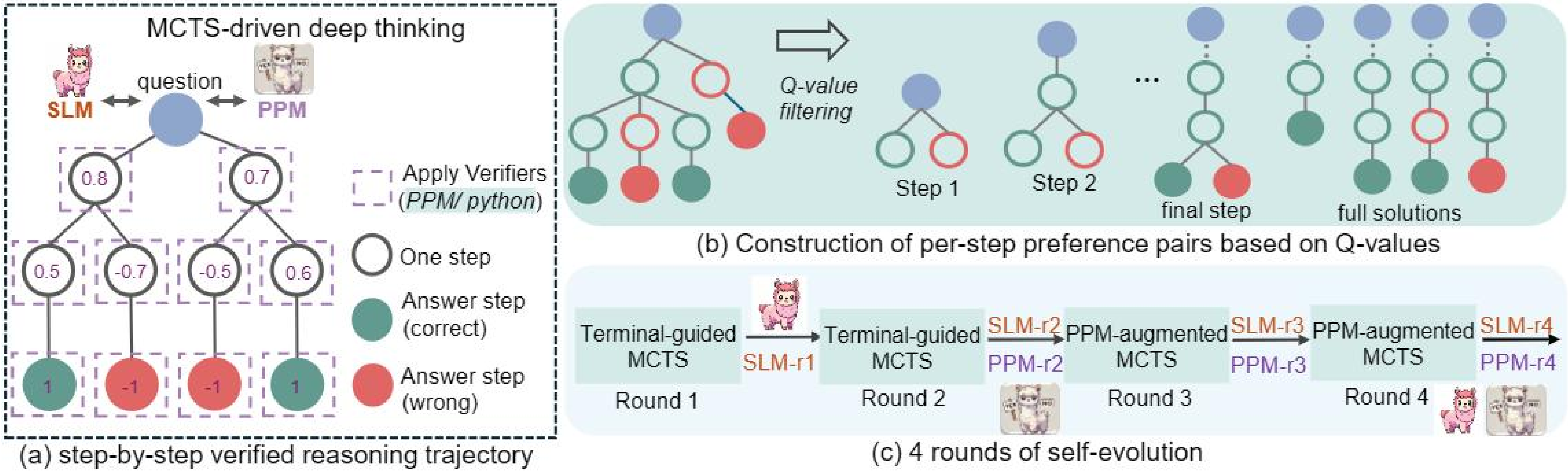


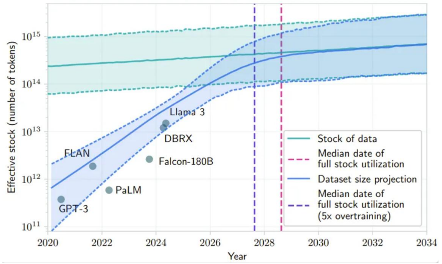

Figure 1: The overview of rStar-Math.

论文：rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking

# Why 合成数据&LLM-as-a-judge?

idea

## 1. 大模型Scaling Law困局，合成数据重要性将逐渐成为共识




但似乎我们不必等到 2028 年了。昨天，The Information 发布了一篇独家报道《随着 GPT 提升减速，OpenAI 改变策略》，其中给出了一些颇具争议的观点：

- OpenAI 的下一代旗舰模型的质量提升幅度不及前两款旗舰模型之间的质量提升；
- AI 产业界正将重心转向在初始训练后再对模型进行提升；
- OpenAI 已成立一个基础团队来研究如何应对训练数据的匮乏。

GPT-5?


## 2. 数据层面先发优势：合成数据积累 + 流程标准化

scale

## 3. 国内缺少具备对应能力与技术的企业，市场蓝海

# 谢谢！

**Contact Us：**


**Tel：** +86 18768113611

**E-mail：** xuchengjin@idea.edu.cn