A Survey on LLM-as-a-Judge
方法论、应用与未来研究方向探讨

分享人：

顾嘉炜（Jiawei Gu）

徐铖晋（Chengjin Xu）

王赛卓（Saizhuo Wang）

LLM-as-a-Judge 方法论

分享人：

顾嘉炜（Jiawei Gu）

A Survey on LLM-as-a-Judge

JIAWEI GU[1,*], XUHUI JIANG[1,*], ZHICHAO SHI[1,2,*], HEXIANG TAN[2], XUEHAO ZHAI[3], CHENGJIN XU[1], WEI LI[2], YINGHAN SHEN[2], SHENGJIE MA[1,4], HONGHAO LIU[1], YUANZHUO WANG[2], JIAN GUO[1,†],

[1]IDEA Research, International Digital Economy Academy, China

[2]Institute of Computing Technology, Chinese Academy of Sciences , China

[3]Department of Civil and Environmental Engineering, Imperial College London, UK

[4]Gaoling School of Artificial Intelligence, Renmin University of China , China

ABSTRACT

Accurate and consistent evaluation is crucial for decision-making across numerous fields, yet it remains a challenging task due to inherent subjectivity, variability, and scale. Large Language Models (LLMs) have achieved remarkable success across diverse domains, leading to the emergence of "LLM-as-a-Judge," where LLMs are employed as evaluators for complex tasks. With their ability to process diverse data types and provide scalable, cost-effective, and consistent assessments, LLMs present a compelling alternative to traditional expert-driven evaluations. However, ensuring the reliability of LLM-as-a-Judge systems remains a significant challenge that requires careful design and standardization. This paper provides a comprehensive survey of LLM-as-a-Judge, addressing the core question: How can reliable LLM-as-a-Judge systems be built? We explore strategies to enhance reliability, including improving consistency, mitigating biases, and adapting to diverse assessment scenarios. Additionally, we propose methodologies for evaluating the reliability of LLM-as-a-Judge systems, supported by a novel benchmark designed for this purpose. To advance the development and real-world deployment of LLM-as-a-Judge systems, we also discussed practical applications, challenges, and future directions. This survey serves as a foundational reference for researchers and practitioners in this rapidly evolving field. The associated resources can be accessed at https://github.com/IDEA-FinAI/LLM-as-a-Judge.

1 INTRODUCTION

> Judgment is the faculty of thinking the particular as contained under the universal. It involves the capacity to subsume under rules, that is, to distinguish whether something falls under a given rule.
>
> —— Kant, *Critique of Judgment* [43], *Introduction IV, 5:179*; *Critique of Pure Reason* [42], *A132/B171.*

Recently, Large Language Models (LLMs) have achieved remarkable success across numerous domains, ranging from technical fields to the humanities and social sciences. Building on their success, the concept of using LLMs as evaluators—commonly referred to as "LLM-as-a-Judge" [160]—has gained significant attention, where LLMs are tasked with determining whether something falls within the scope of a given rule [42, 43]. This growing interest stems from LLMs' ability to mimic human-like reasoning and thinking processes, enabling them to take on roles traditionally reserved for human experts while offering a cost-effective solution that can be effortlessly scaled to meet

Outline

Background

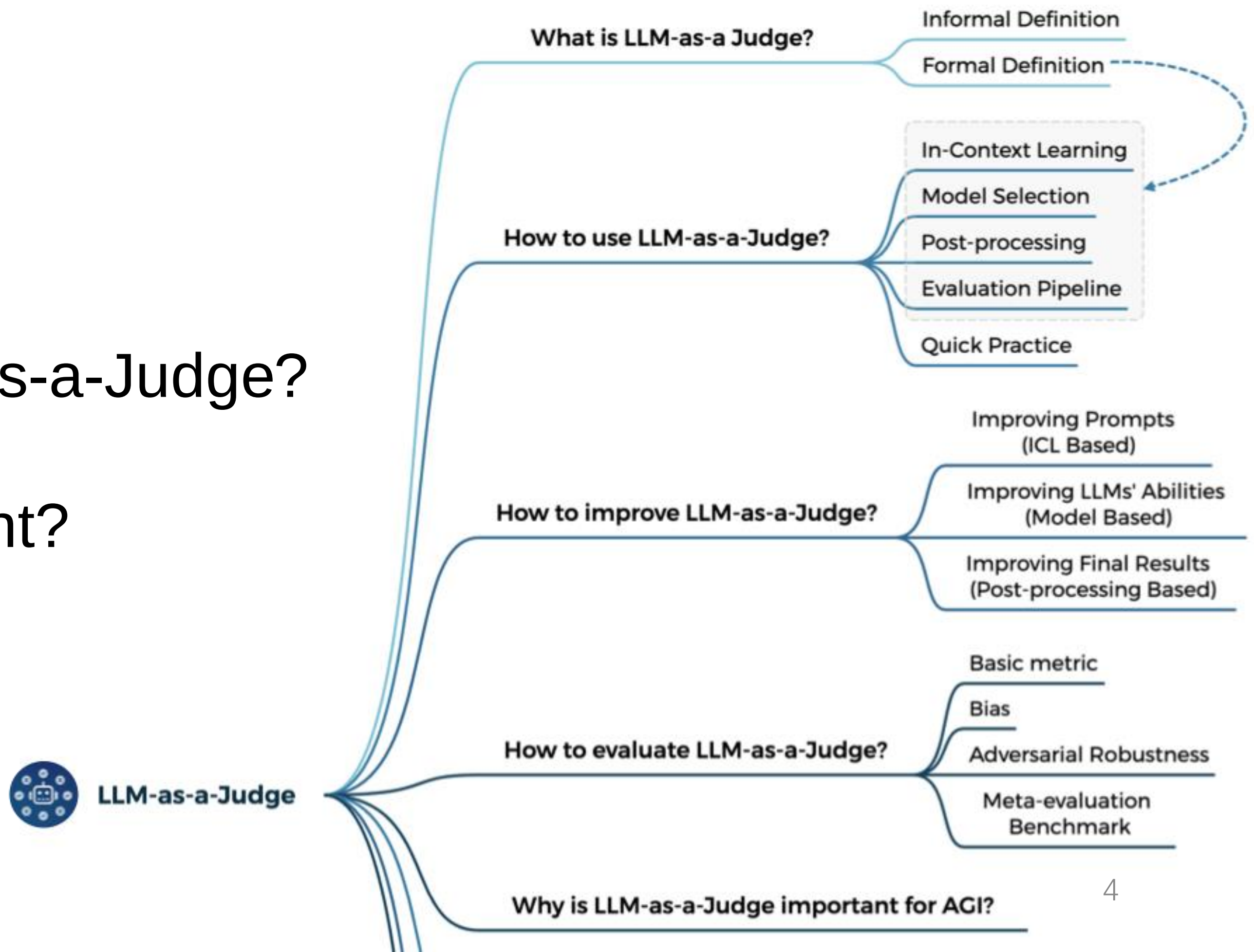
1 What is LLM-as-a-Judge?

2 How to use LLM-as-a-Judge?

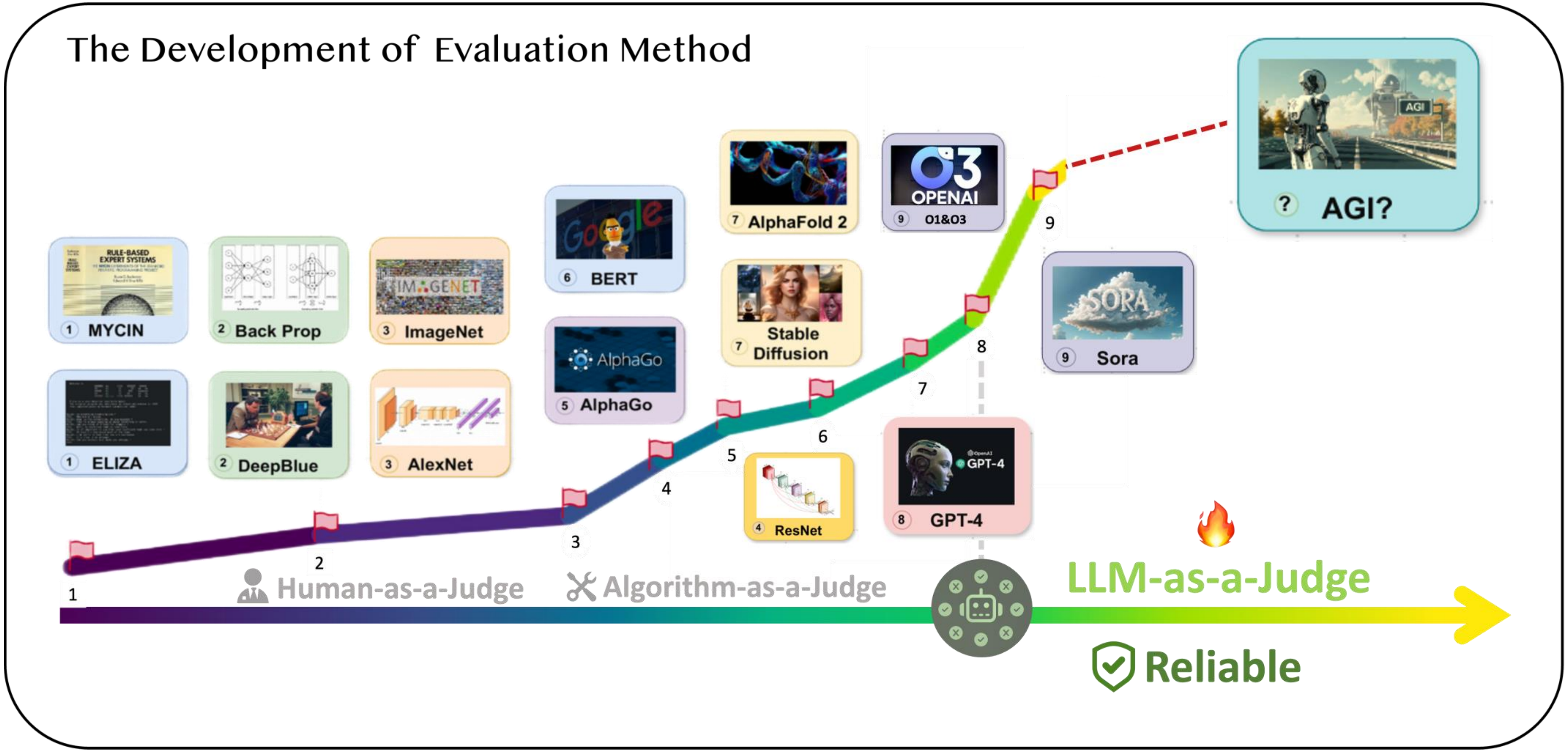
3 How to improve a reliable LLM-as-a-Judge?

4 Why is LLM-as-a-Judge important?

Conclusion

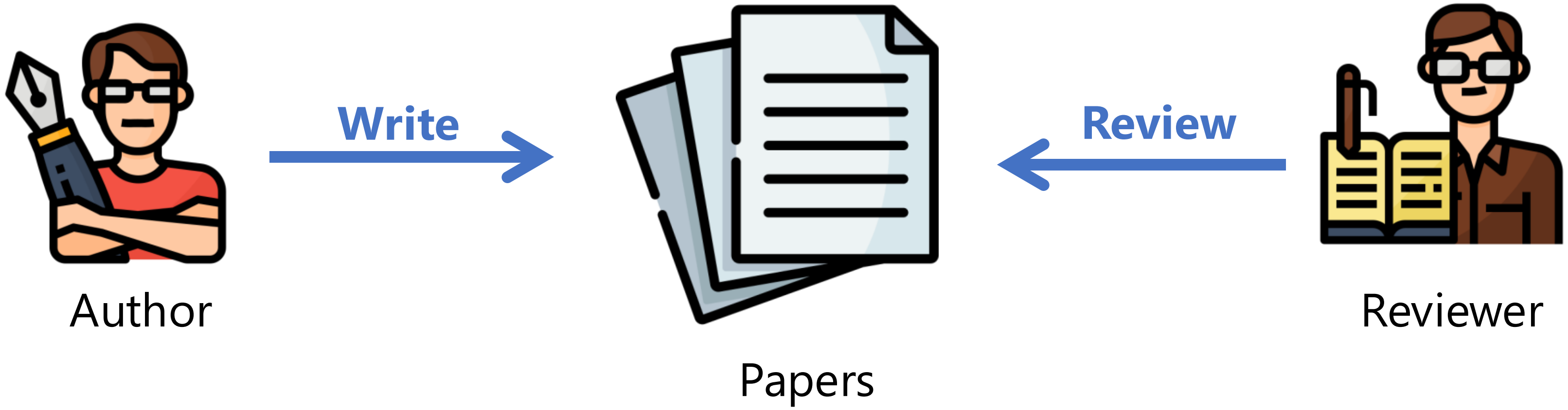


Background



The Development of Evaluation Method
1 MYCIN
1 ELIZA
2 Back Prop
2 DeepBlue
3 ImageNet
3 AlexNet
6 BERT
5 AlphaGo
7 AlphaFold 2
7 Stable Diffusion
4 ResNet
9 O1&O3
8 GPT-4
9 Sora
? AGI?
1
2
3
4
5
6
7
8
9
Human-as-a-Judge
Algorithm-as-a-Judge
LLM-as-a-Judge
Reliable

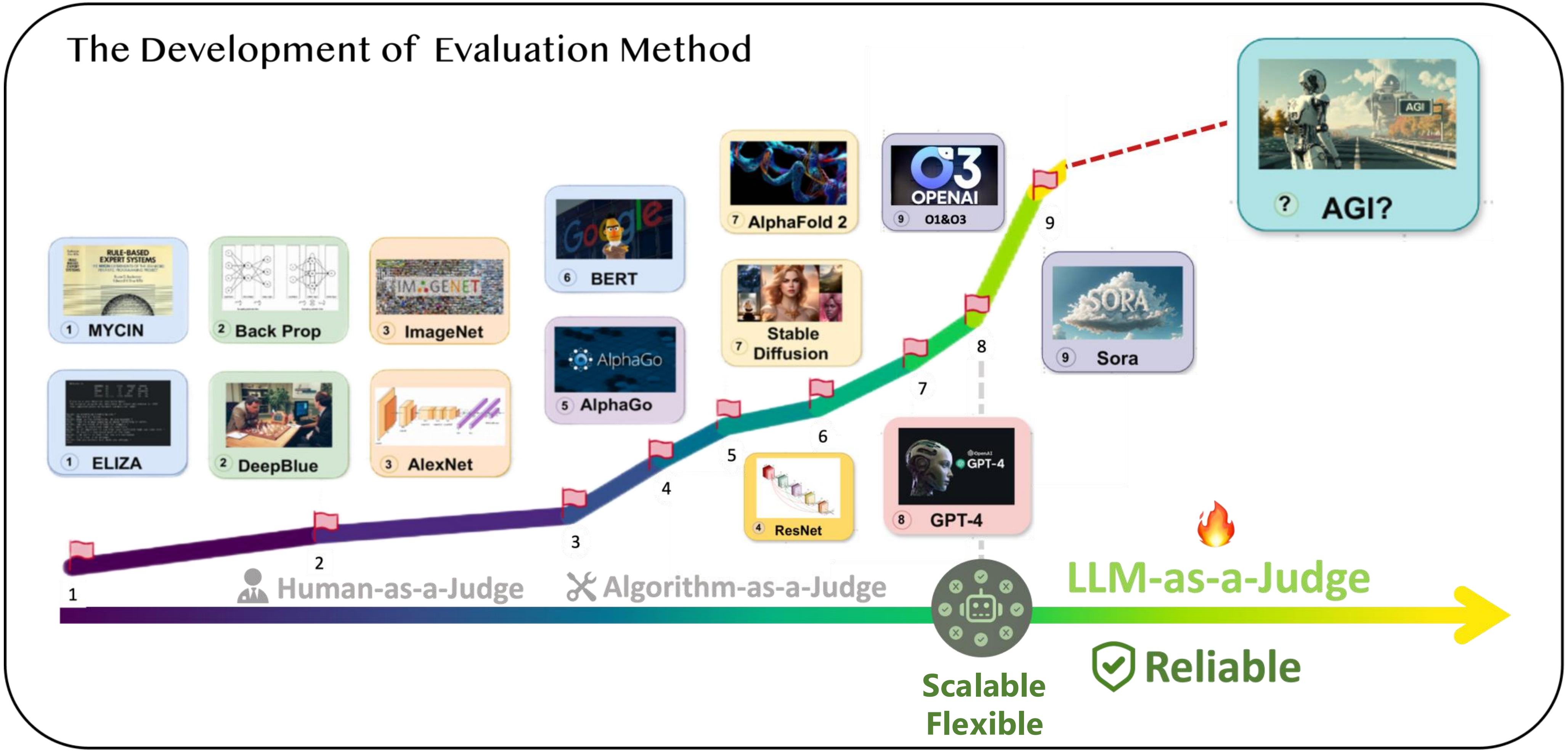
Background



Human (not scalable)	good enough?	Accept or not?
Algorithm (not flexible)	Plagiarism Rate	Plagiarism Rate
LLM-as-a-Judge (Scalable, Flexible)	good enough? Why?	Accept or not? Why?

Judge for
A score
A decision
A choice
A rate
……

Background



The Development of Evaluation Method
1 MYCIN
1 ELIZA
2 Back Prop
2 DeepBlue
3 ImageNet
3 AlexNet
6 BERT
5 AlphaGo
7 AlphaFold 2
7 Stable Diffusion
4 ResNet
9 O1&O3
8 GPT-4
9 Sora
? AGI?
Human-as-a-Judge
Algorithm-as-a-Judge
LLM-as-a-Judge
Scalable
Flexible
Reliable

Background

对创作idea的判断

对稿子质量的判断

对罪责下的裁定

对案情的判断

对病情的判断

对投资预期的判断

1 What is LLM-as-a-Judge

> 判断是将具体事物看作是普遍规律下的一部分的能力。它涉及将事物归纳到规则中的能力，即区分某物是否符合某一规则。
>
> —— 康德，《判断力批判》；《纯粹理性批判》

LLM-as-a-Judge refers to the use of LLMs to evaluate objects, actions, or decisions based on predefined rules criteria, or preferences. It encompasses a broad spectrum of roles, including **Graders, Evaluators/Assessors, Critics, Verifiers, Examiners, Reward/Ranking Models**, etc.

1 What is LLM-as-a-Judge

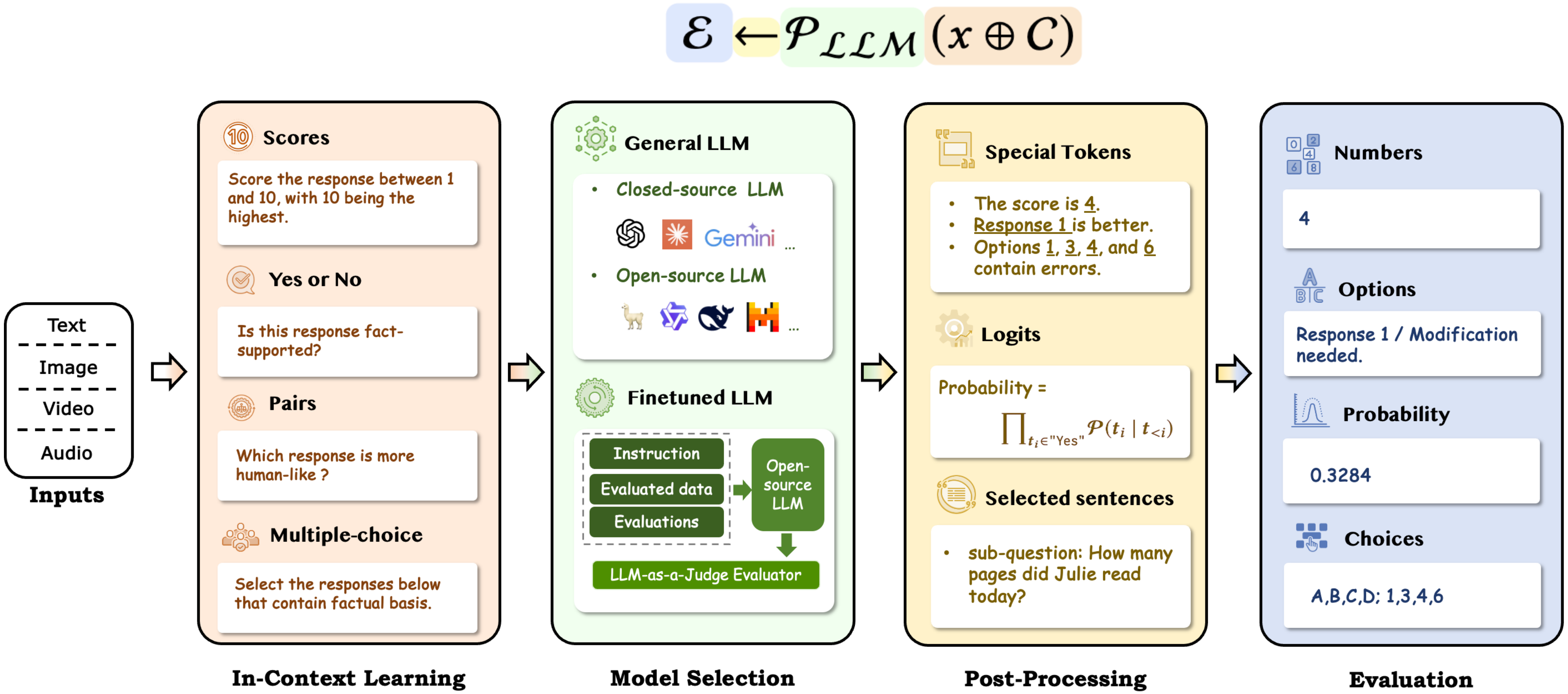
$$\mathcal{E} \leftarrow \mathcal{P}_{\mathcal{LLM}}(x \oplus \mathcal{C})$$

- $\mathcal{E}$: The final evaluation obtained from the whole LLM-as-a-Judge process in the expected manner. It could be a score, a choice, a label or a sentence, etc.
- $\mathcal{P}_{\mathcal{LLM}}$: The probability function defined by the corresponding LLM, and the generation is an auto-regressive process.
- x: The input data in any available types (text, image, video), which waiting to be evaluated.
- $\mathcal{C}$: The context for the input x, which is often prompt template or combined with history information in dialogue.
- $\oplus$: The combination operator combines the input x with the context $\mathcal{C}$, and this operation can vary depending on the context, such as being placed at the beginning, middle, or end.

The formulation of LLM-as-a-Judge reflects that LLM is a type of **auto-regressive generative model,** which generates subsequent content based on the context and then obtains target **evaluation** from it.

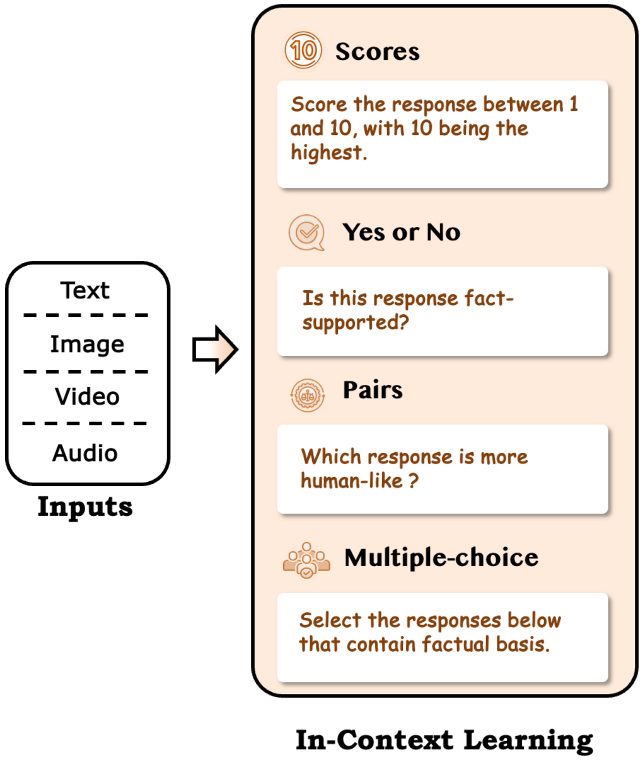
Gu, Jiawei, et al. "A Survey on LLM-as-a-Judge." *arXiv preprint arXiv:2411.15594* (2024).

2 How to use LLM-as-a-Judge



2.1 In-Context Learning

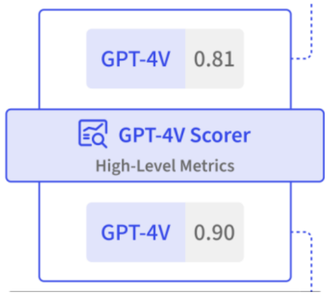
$$\mathcal{E} \leftarrow P_{\mathcal{LLM}}(x \oplus \mathcal{C})$$



2.1.1 Generating scores

Evaluate the quality of summaries written for a news article. Rate each summary on four dimensions: {Dimension_1}, {Dimension_2}, {Dimension_3}, and {Dimension_4}. You should rate on a scale from 1 (worst) to 5 (best).

Article: {Article}
Summary: {Summary}



ChartMimic

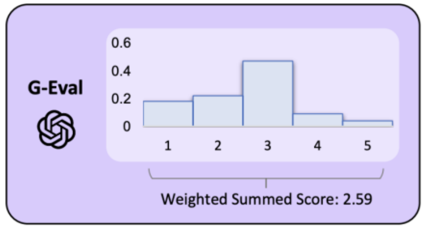
0-1

Continuous

LM as an Examiner

1-3,1-5

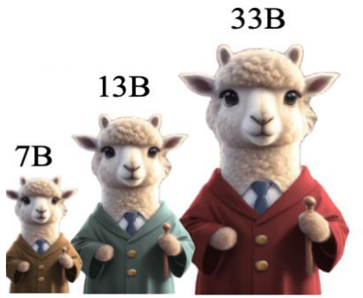
Likert



G-Eval

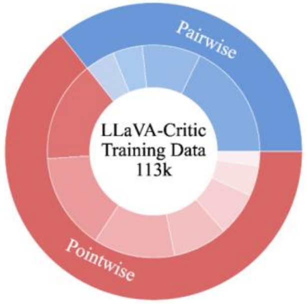
1-5

Likert



JudgeLM

1-10



LLaVA-Critic

0-100

Continuous

2.1.1 Generating scores

B.4 Likert Scale Scoring

You are a fair assessment expert, and you will be given a set of question-answer pairs. Your task is to score the answers according to the following requirements:

a. You should score the answer based on your knowledge of the corresponding question. You can assume your own answer to the corresponding question is the ground truth for the question.

b. You should rate the answer on 5 metrics, for the first 4 metrics, assign a score between 1 and 3, with 3 being the highest:

1. For accuracy, you will score whether the answer correctly answers the question.

2. For coherence, you will assess the structure and logic of the answer, and whether the answer is understandable by non-professionals.

3. For factuality, you will only evaluate whether the answer contains factual errors.

4. For comprehensive, you will determine if the answer covers multiple aspects of the question and provides a comprehensive response. For simple questions (when, which, where, etc), the plain answer itself suffices and should be rated 3.

5. Finally, you will provide an overall score between 1 and 5, with 5 being the highest.

You should only give the score, Format like: coherence: 3

DO NOT complete the answer!

Question: {Question} Answer: {Response}

1-3,1-5

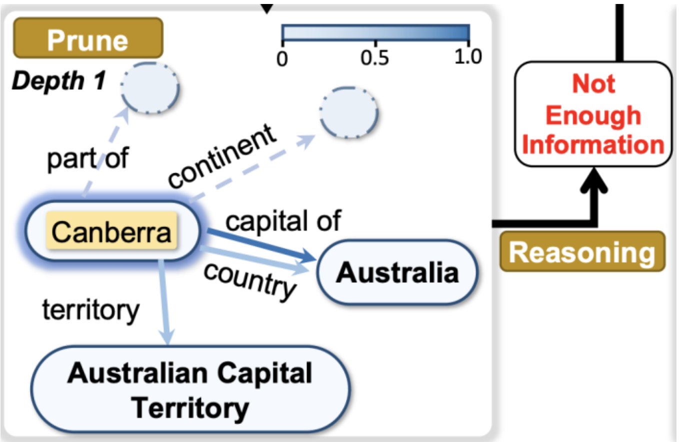
Likert

Bai, Yushi, et al. "Benchmarking foundation models with language-model-as-an-examiner." *Advances in Neural Information Processing Systems* 36 (2024).

2.1.2 Solving Yes/No question

Is the sentence supported by the article? Answer "Yes" or "No".

Article: {Article}
Sentence: {Sentence}

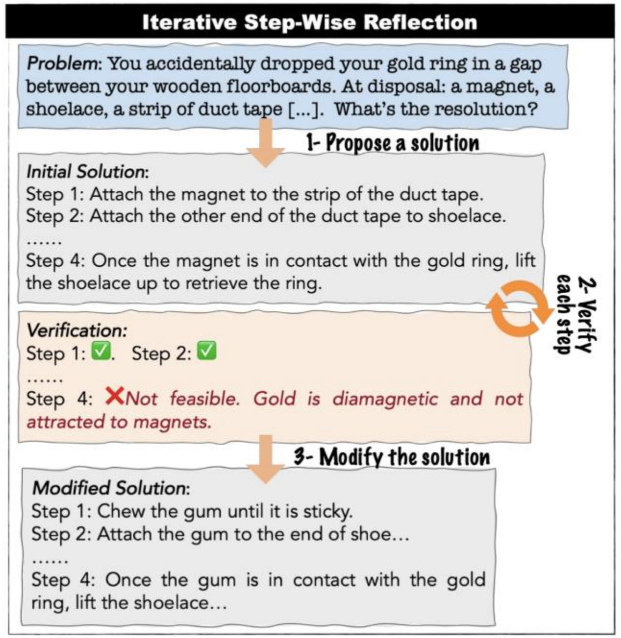


Think-on-Graph

Yes/No

MathVista

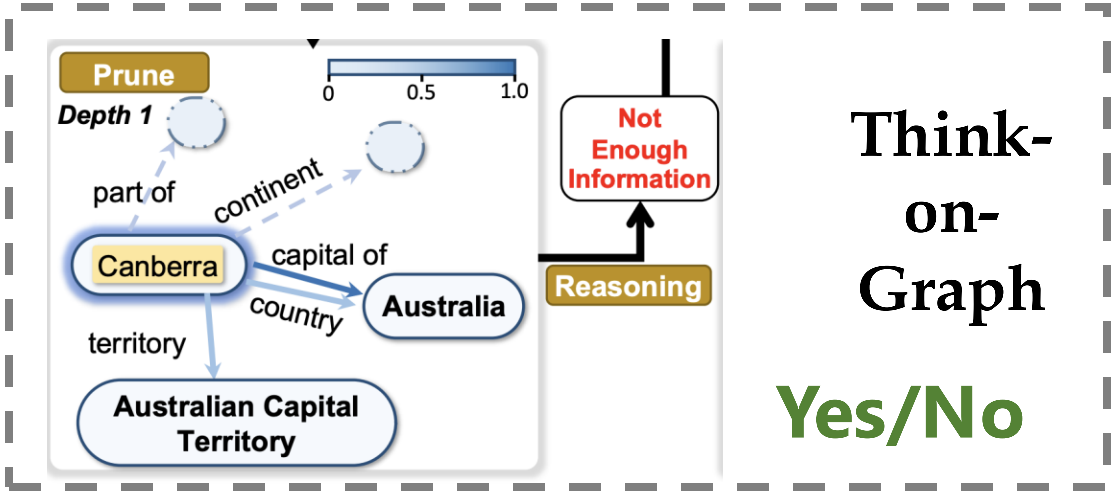
0/1



MacGyver

"Modification needed." / "No modification needed."

2.1.2 Solving Yes/No question



E.4 TOG-R

E.4.1 REASONING

Please answer the question using Topic Entity, Relations Chains and their Candidate Entities that contribute to the question, you are asked to answer whether it's sufficient for you to answer the question with these triples and your knowledge (Yes or No).

In-Context Few-shot

Q: {Query}

Topic Entity, with relations chains, and their candidate entities: {Explored Relation Chains}

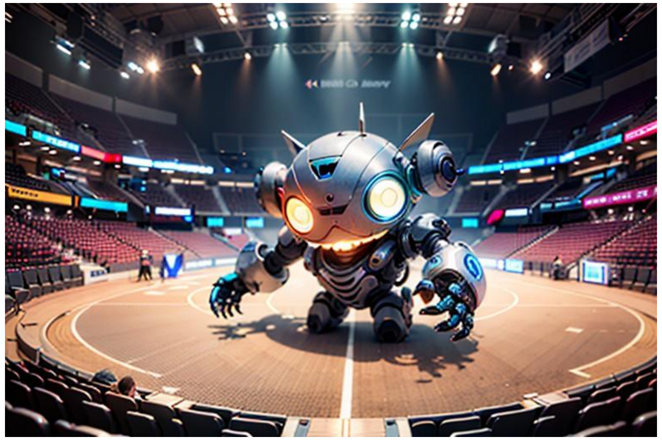
A:

Sun, Jiashuo, et al. "Think-on-graph: Deep and responsible reasoning of large language model with knowledge graph." *arXiv preprint arXiv:2307.07697* (2023).

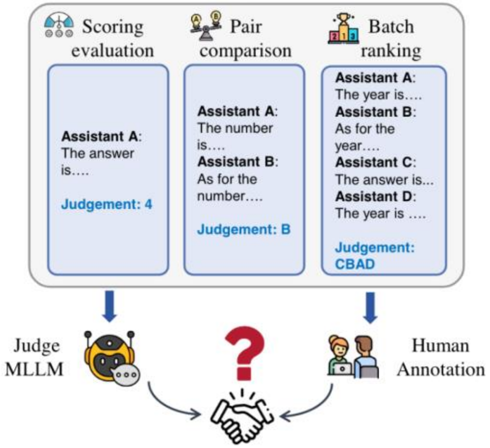
2.1.3 Conducting pairwise comparisons

Given a new article, which summary is better? Answer "Summary 0" or "Summary 1". You do not need to explain the reason.

Article: {Article}
Summary 0: {Summary_0}
Summary 1: {Summary_1}



Chatbot Arena



MLLM Arena

Two options

Three options:

win, tie, lose

Four options:

win, both good tie, both bad tie, lose

Batch Ranking

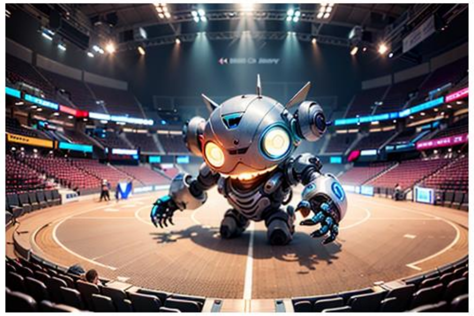
2.1.3 Conducting pairwise comparisons

```
[System]
Please act as an impartial judge and evaluate the quality of the responses provided by two
AI assistants to the user question displayed below. You should choose the assistant that
follows the user’s instructions and answers the user’s question better. Your evaluation
should consider factors such as the helpfulness, relevance, accuracy, depth, creativity,
and level of detail of their responses. Begin your evaluation by comparing the two
responses and provide a short explanation. Avoid any position biases and ensure that the
order in which the responses were presented does not influence your decision. Do not allow
the length of the responses to influence your evaluation. Do not favor certain names of
the assistants. Be as objective as possible. After providing your explanation, output your
final verdict by strictly following this format: "[[A]]" if assistant A is better, "[[B]]"
if assistant B is better, and "[[C]]" for a tie.

[User Question]
{question}

[The Start of Assistant A’s Answer]
{answer_a}
[The End of Assistant A’s Answer]

[The Start of Assistant B’s Answer]
{answer_b}
[The End of Assistant B’s Answer]
```



Chatbot Arena

Zheng, Lianmin, et al. "Judging llm-as-a-judge with mt-bench and chatbot arena." *Advances in Neural Information Processing Systems* 36 (2023): 46595-46623.

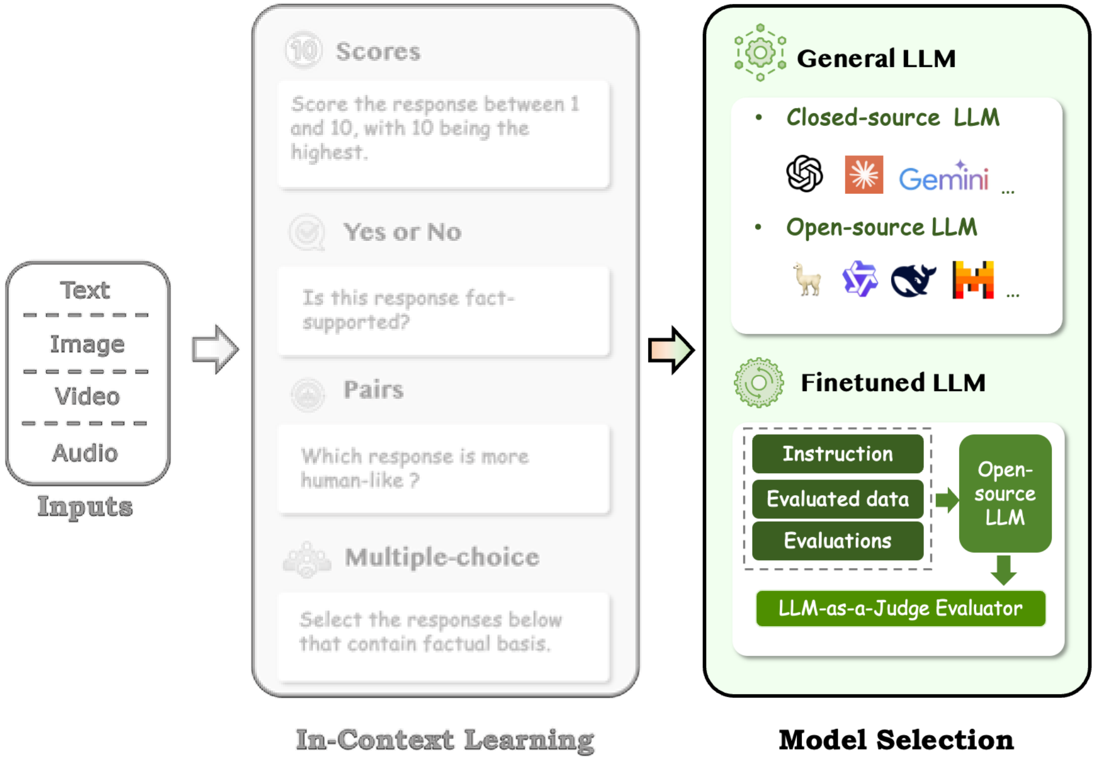
2.1.4 Making multiple-choice selections.

You are given a summary and some semantic content units. For each semantic unit, choose those can be inferred from the summary, return their number.

Summary: {Summary}
Semantic content units:
1. {SCU_1}
2. {SCU_2}
......
n. {SCU_n}

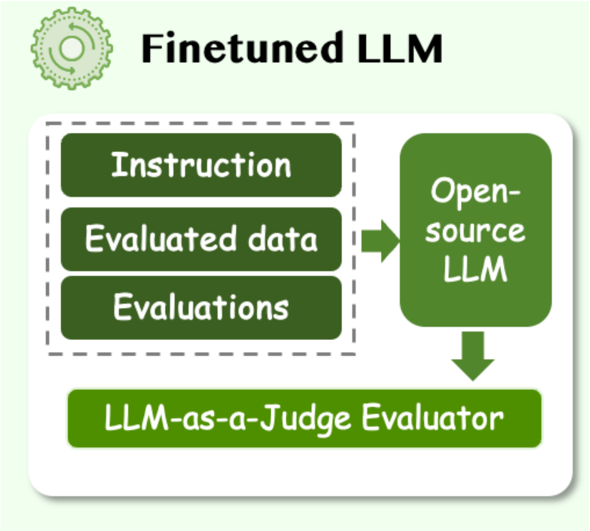
2.2 Model Selection

$$\mathcal{E} \leftarrow \mathcal{P}_{\mathcal{LLM}}(x \oplus \mathcal{C})$$



2.2 Model Selection

$$\mathcal{E} \leftarrow \mathcal{P}_{\mathcal{LLM}}(x \oplus \mathcal{C})$$

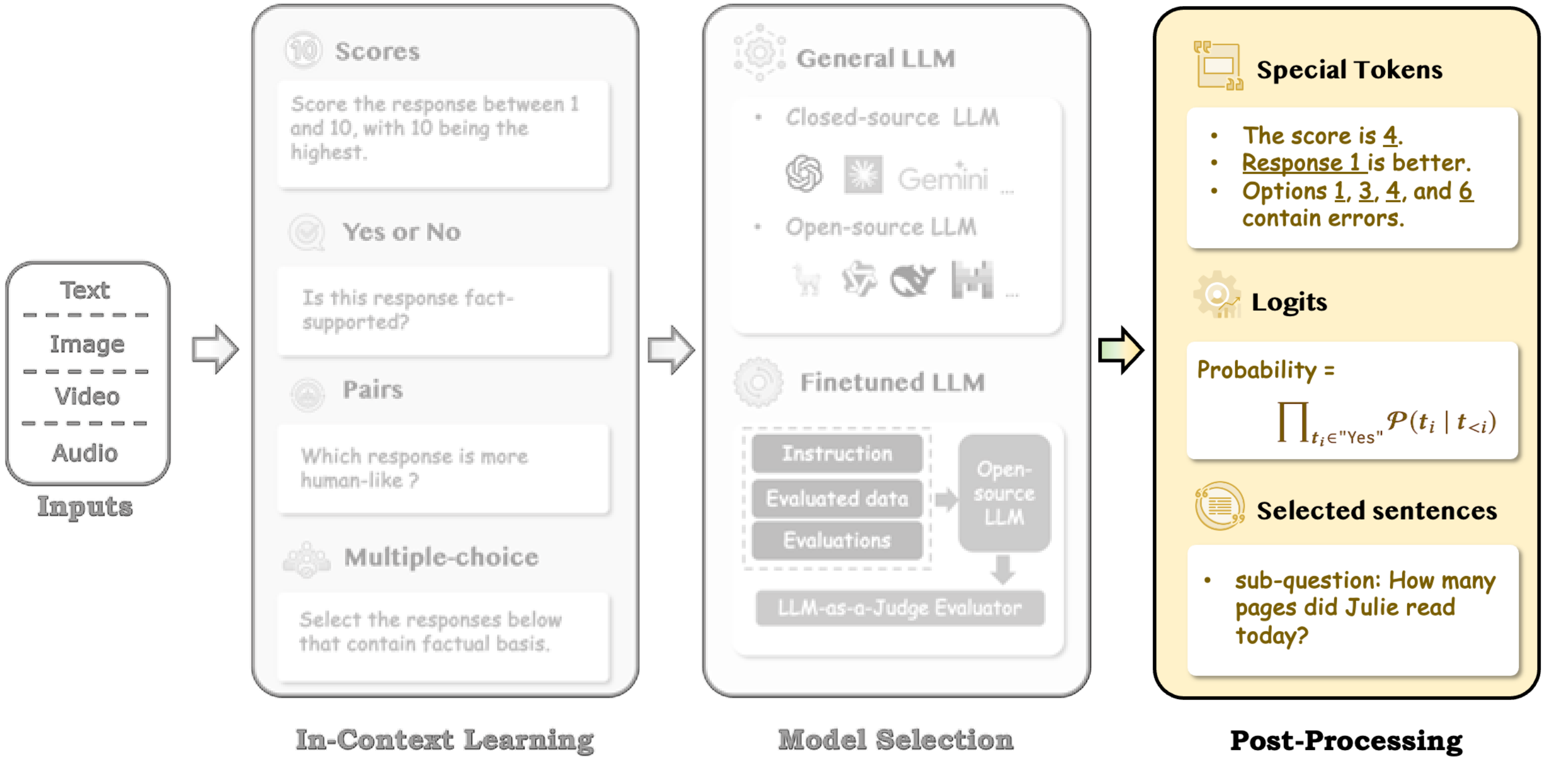


reasoning

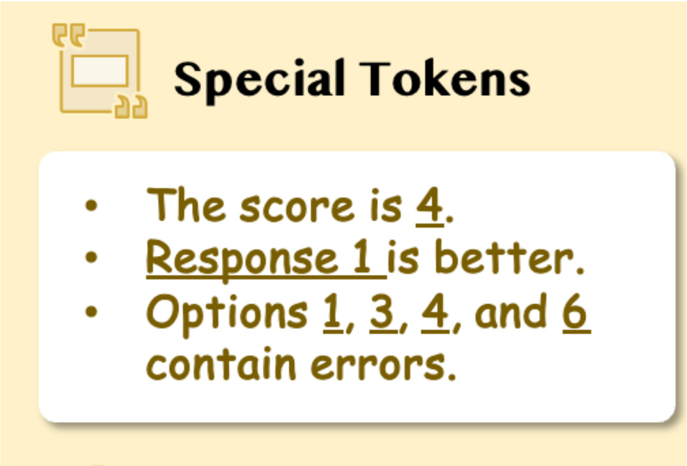
Instruction Following

2.3 Post-Processing

$$\mathcal{E} \leftarrow \mathcal{P}_{\mathcal{LLM}}(x \oplus \mathcal{C})$$



2.3.1 Extracting specific tokens

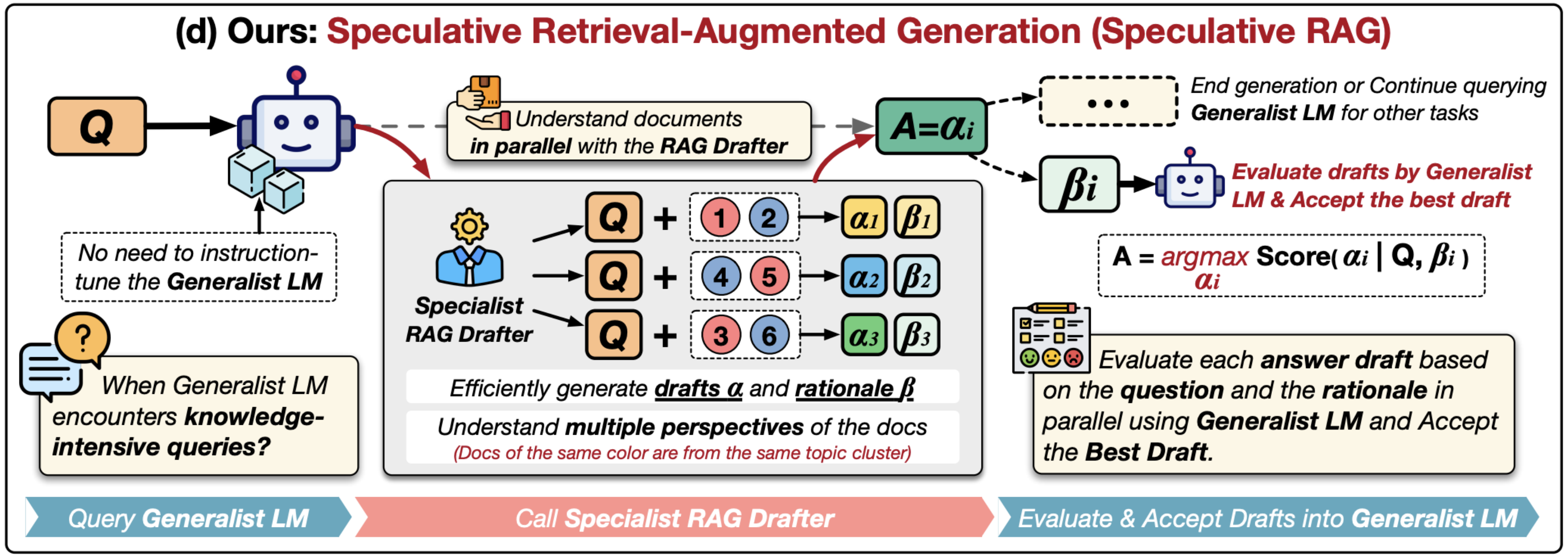


It is common to apply a **rule-match** to extract the corresponding token from the response generated during probability distribution iteration.

Variability in phrasing can complicate consistent parsing. This is particularly true when the evaluator model lacks sufficient instruction-following ability. For exmaple:
"Response 1 is better" vs "The better one is response 1”
“Five” vs “5”

- **Clear instructions**
- **few-shot strategies**
- **Model with strong instruction-following ability**

2.3.2 Normalizing the output logits



Logits

Probability =

$$\prod_{t_i \in \text{"Yes"}} \mathcal{P}(t_i \mid t_{<i})$$

“Do you think the rationale supports the answer, yes or no?”

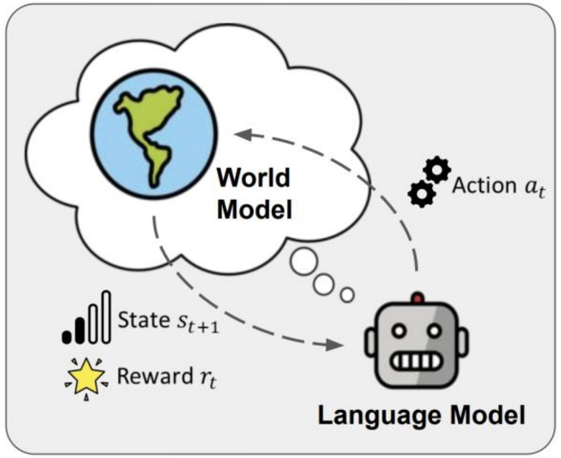
Self-reflection score = conditional probability of the positive answer ("Yes") to the self-reflection statement.

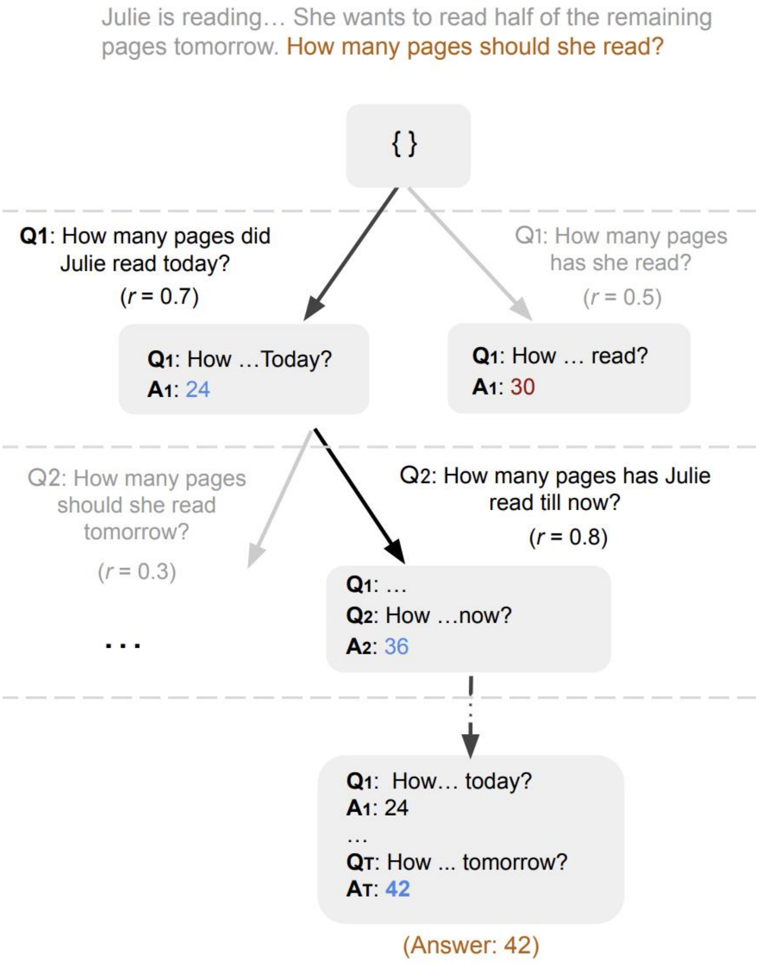
Wang, Zilong, et al. "Speculative rag: Enhancing retrieval augmented generation through drafting." *arXiv preprint arXiv:2407.08223* (2024).

2.3.3 Selecting sentences

Selected sentences

- *sub-question: How many pages did Julie read today?*

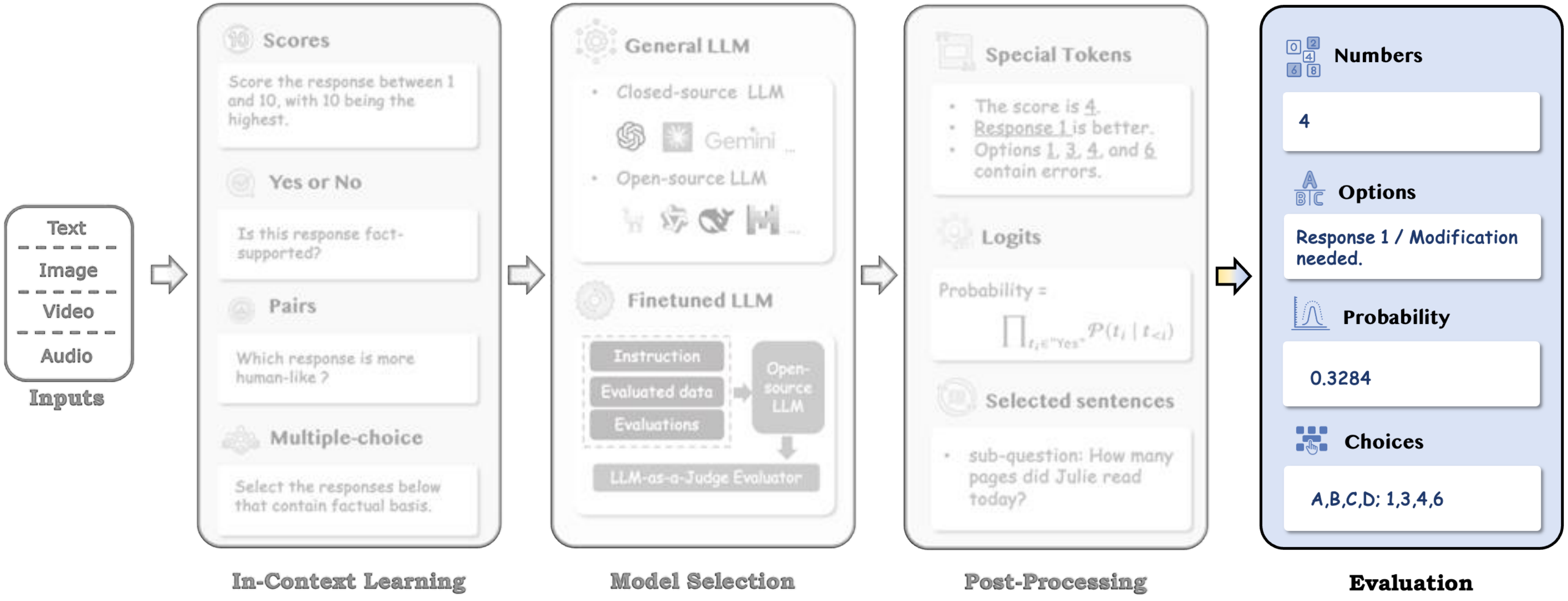




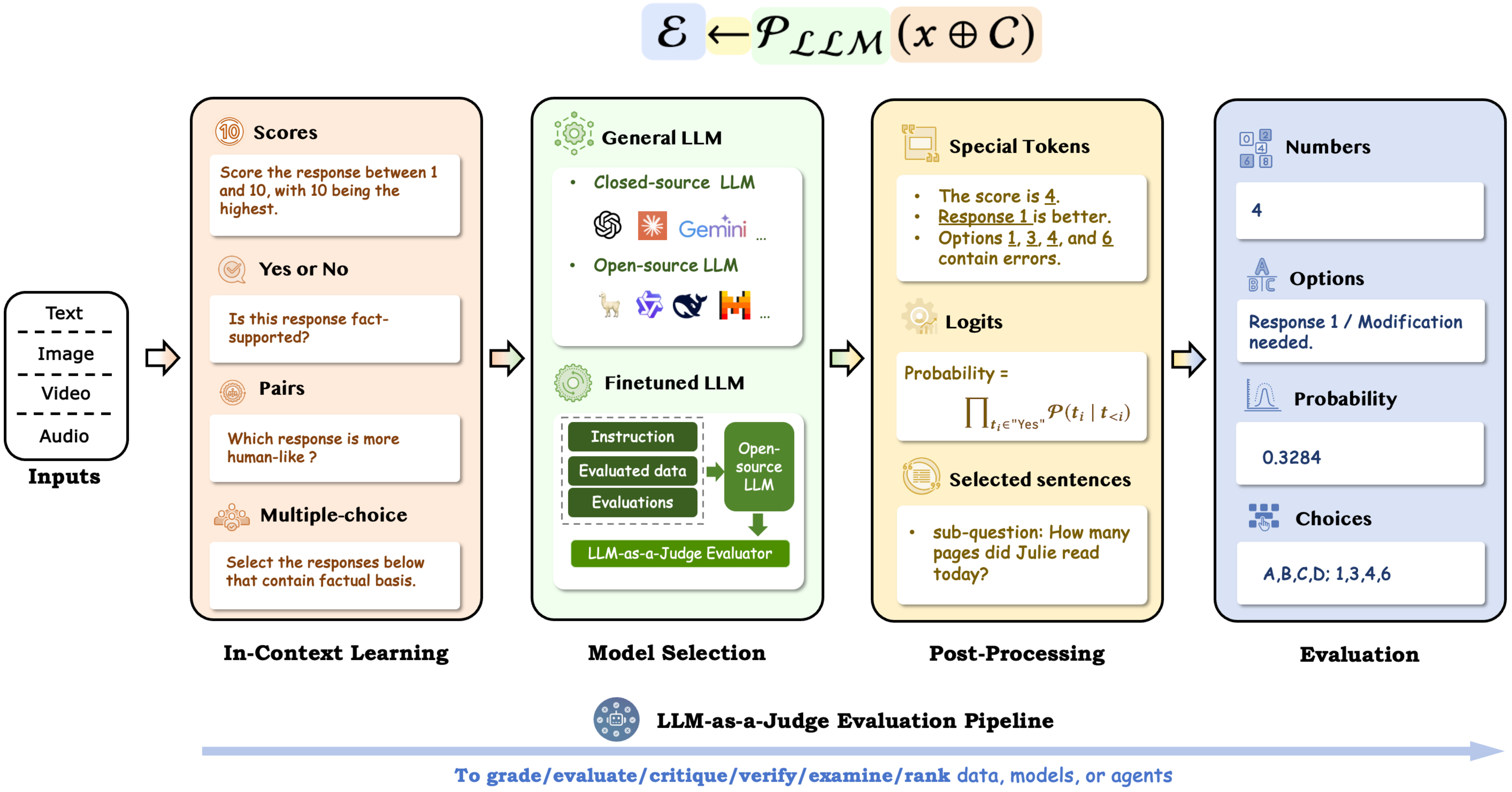
Hao, Shibo, et al. "Reasoning with language model is planning with world model." *arXiv preprint arXiv:2305.14992* (2023).

2.4 Evaluation Pipeline

$$\mathcal{E} \leftarrow P_{\mathcal{LLM}}(x \oplus \mathcal{C})$$



2.4 Evaluation Pipeline

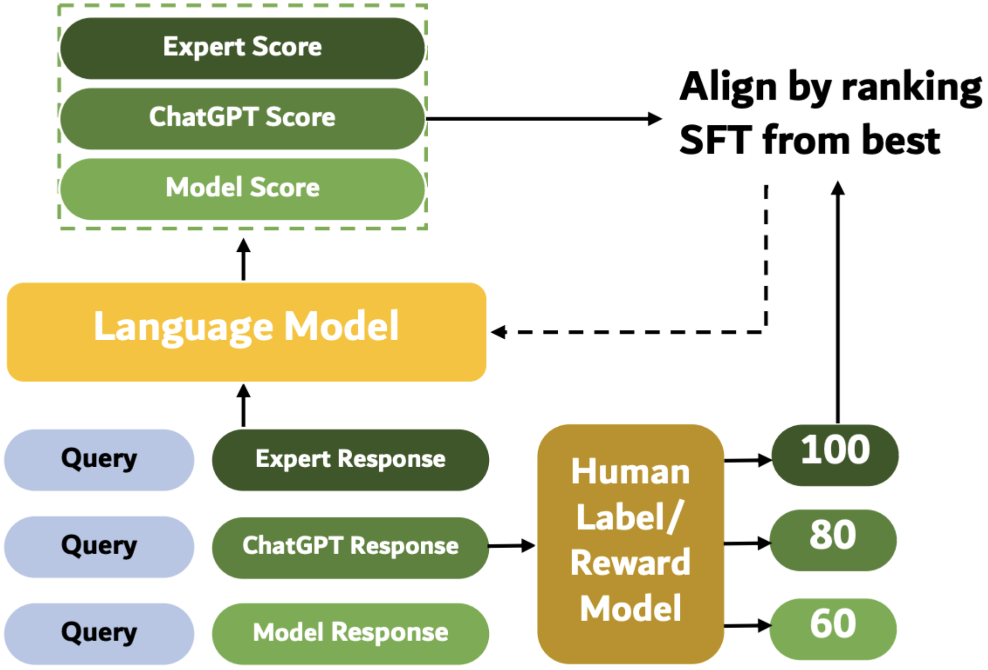


2.4 Evaluation Pipeline

2.4 Evaluation Pipeline

for Data

RRHF



Language models align with human preferences via supervised fine-tuning and **RLHF**, though PPO's complexity drives interest in simpler methods. Using LLM-as-a-Judge, like ChatGPT, **RRHF** provides a straightforward alternative for evaluation and alignment.

Yuan, Zheng, et al. "Rrhf: Rank responses to align language models with human feedback without tears." *arXiv preprint arXiv:2304.05302* (2023).

2.4 Evaluation Pipeline

for Models

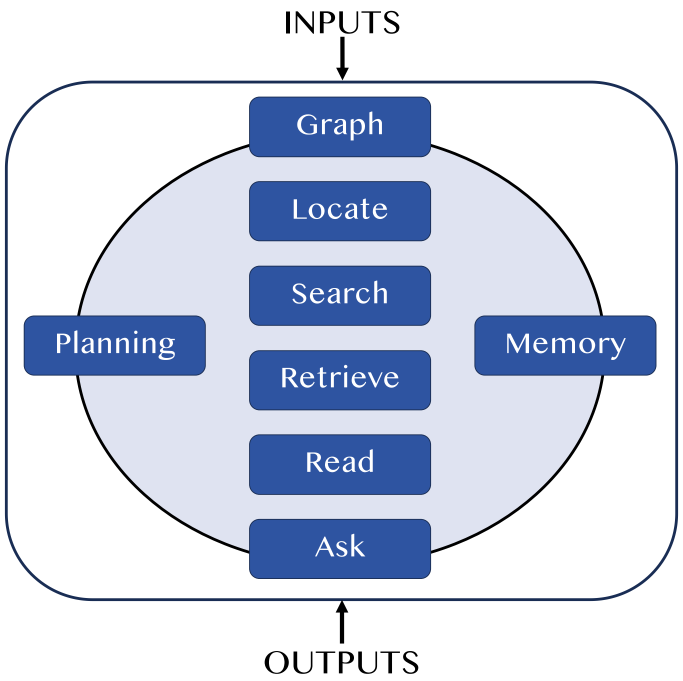
Chatbot Arena

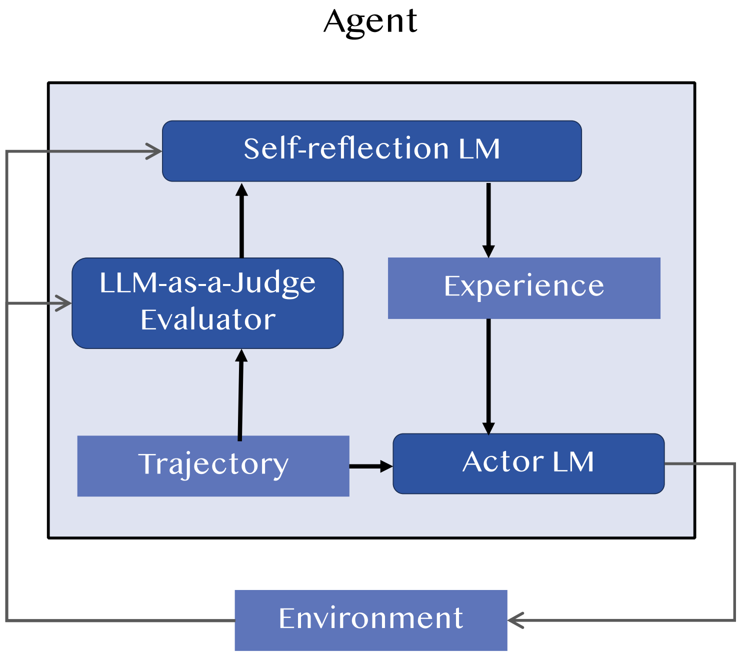
LLM-as-a-Judge

- for Models
- for Agents
- for Data
- for Reasoning/Thinking

2.4 Evaluation Pipeline

for Agents



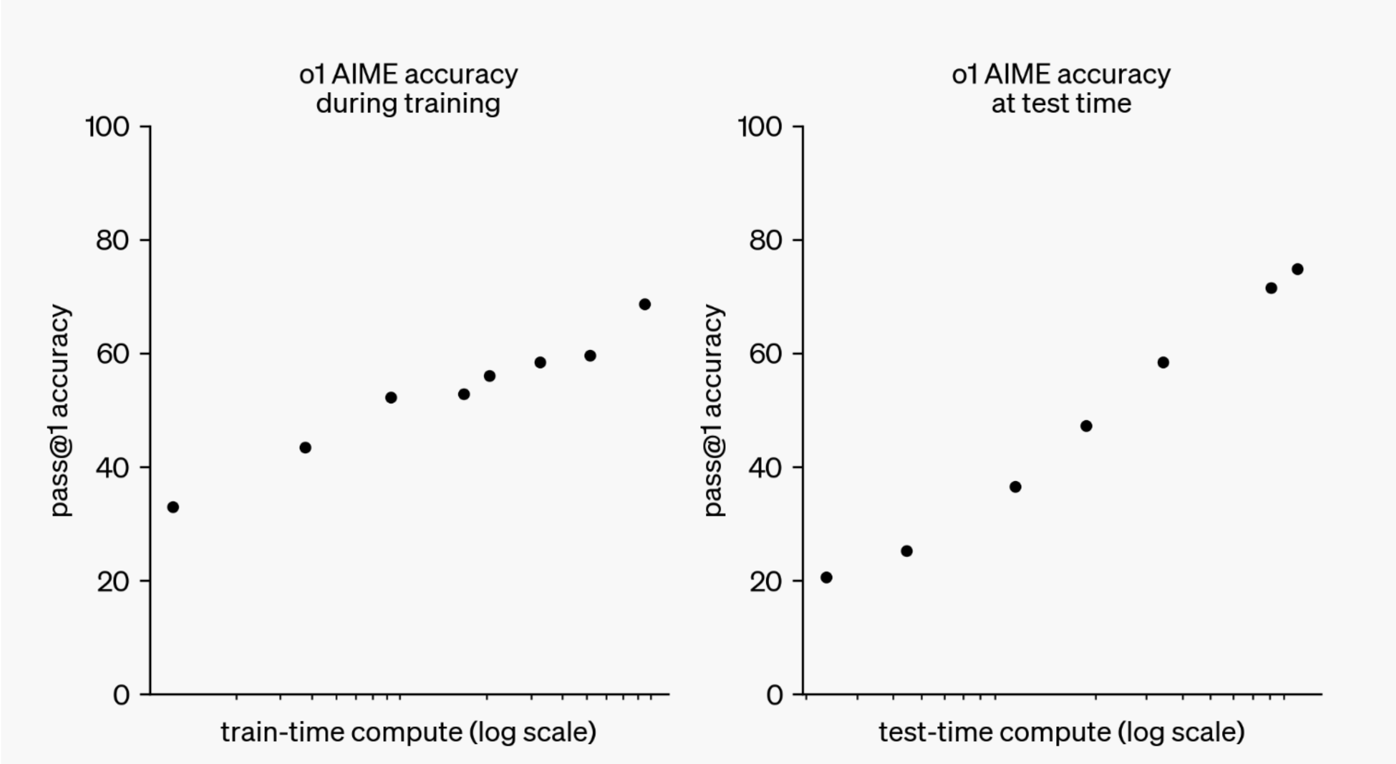


LLM-as-a-Judge can be applied in two ways for agents: **evaluating the entire process** or **assessing specific stages** within the framework. As the agent's brain, LLMs can evaluate like humans, reducing human involvement and balancing thoroughness with effort. Additionally, agents can interact with environments via language and use LLM feedback to guide their actions.

Shinn, Noah, et al. "Reflexion: Language agents with verbal reinforcement learning." *Advances in Neural Information Processing Systems* 36 (2024).
Zhuge, Mingchen, et al. "Agent-as-a-judge: Evaluate agents with agents." *arXiv preprint arXiv:2410.10934* (2024).

2.4 Evaluation Pipeline

for Reasoning/Thinking



LLM-as-a-Judge & Test time compute scaling
(More Reasoning/Thinking)

2.4 Evaluation Pipeline

for Reasoning/Thinking

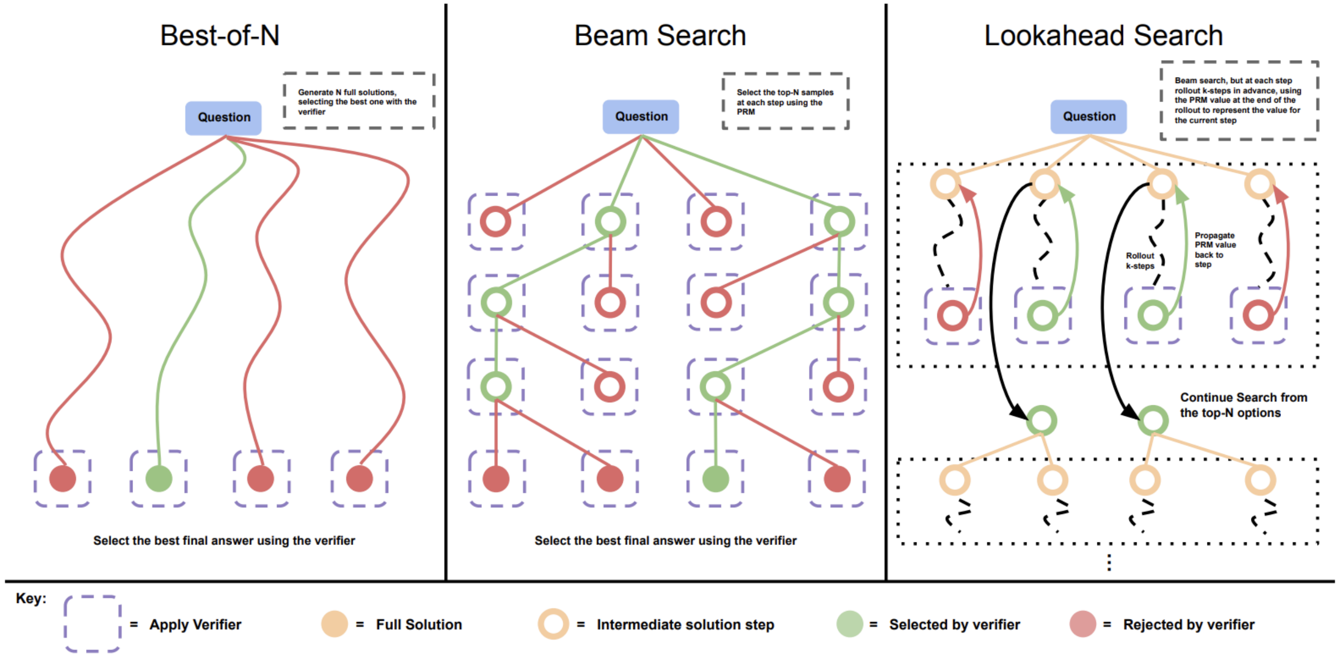
LLM-as-a-Judge & Test time compute scaling
(More Reasoning/Thinking)

Reasoning, central to tasks like decision-making and problem-solving, **often relies on judgments for logical coherence and clarity.**

LLM-as-a-Judge enhances reasoning in two ways: **during training**, it acts as a reward model in reinforcement learning, helping create high-quality reasoning datasets through verification, preference optimization, and self-refinement; **during testing**, it evaluates and selects the best reasoning paths, such as in "Best-of-N" scenarios. This dual role is essential for improving reasoning systems.

2.4 Evaluation Pipeline

for Reasoning/Thinking



Every node needs a judge.

When the number of judges is sufficient, the thinking process will be more effective.

Snell, Charlie, et al. "Scaling llm test-time compute optimally can be more effective than scaling model parameters." *arXiv preprint arXiv:2408.03314* (2024).

2.4 Evaluation Pipeline

for Reasoning/Thinking

Model	Method	Reward Model	N=1	N=2	N=4	N=8	N=16
	Majority Voting	–		–	37.25	36.25	38.25
	BoN	GPT-4o (Self)		35.50	35.75	36.75	–
GPT-4o	BoN	Gemini Flash Thinking	36.00	40.75	36.25	36.5	–
	Tournament	Gemini Flash Thinking		40.75	39.25	**41.25**	35.25
	Pass@N	–		45.00	53.25	65.75	74.00
	Majority Voting	–		–	37.75	39.25	39.75
	BoN	Gemini Flash (Self)		38.25	36.50	36.00	–
Gemini 2.0 Flash	BoN	Gemini Flash Thinking	36.25	36.75	37.00	40.25	–
	Tournament	Gemini Flash Thinking		36.75	37.25	**40.75**	38.75
	Pass@N	–		45.25	56.25	64.50	75.00
	Majority Voting	–		–	48.00	49.00	**50.75**
Gemini 2.0 Flash Thinking	Tournament	Gemini Flash Thinking (Self)	43.50	45.50	47.25	47.25	48.00
	Pass@N	–		53.75	64.50	71.50	81.50
o1	–	–	**45.75**	–	–	–	–

Testing the scaling with LLM-as-a-Judge is effective.

Training with reinforcement signals from LLM-as-a-Judge can reach higher levels of reasoning.

Hao, Yunzhuo, et al. "Can MLLMs Reason in Multimodality? EMMA: An Enhanced MultiModal ReAsoning Benchmark." *arXiv preprint arXiv:2501.05444* (2025).

3 How to improve a reliable LLM-as-a-Judge?

$$\mathcal{E} \leftarrow \mathcal{P}_{\mathcal{LLM}}(x \oplus \mathcal{C})$$

① **位置偏差**
评估器倾向于偏好提示中特定位置的响应

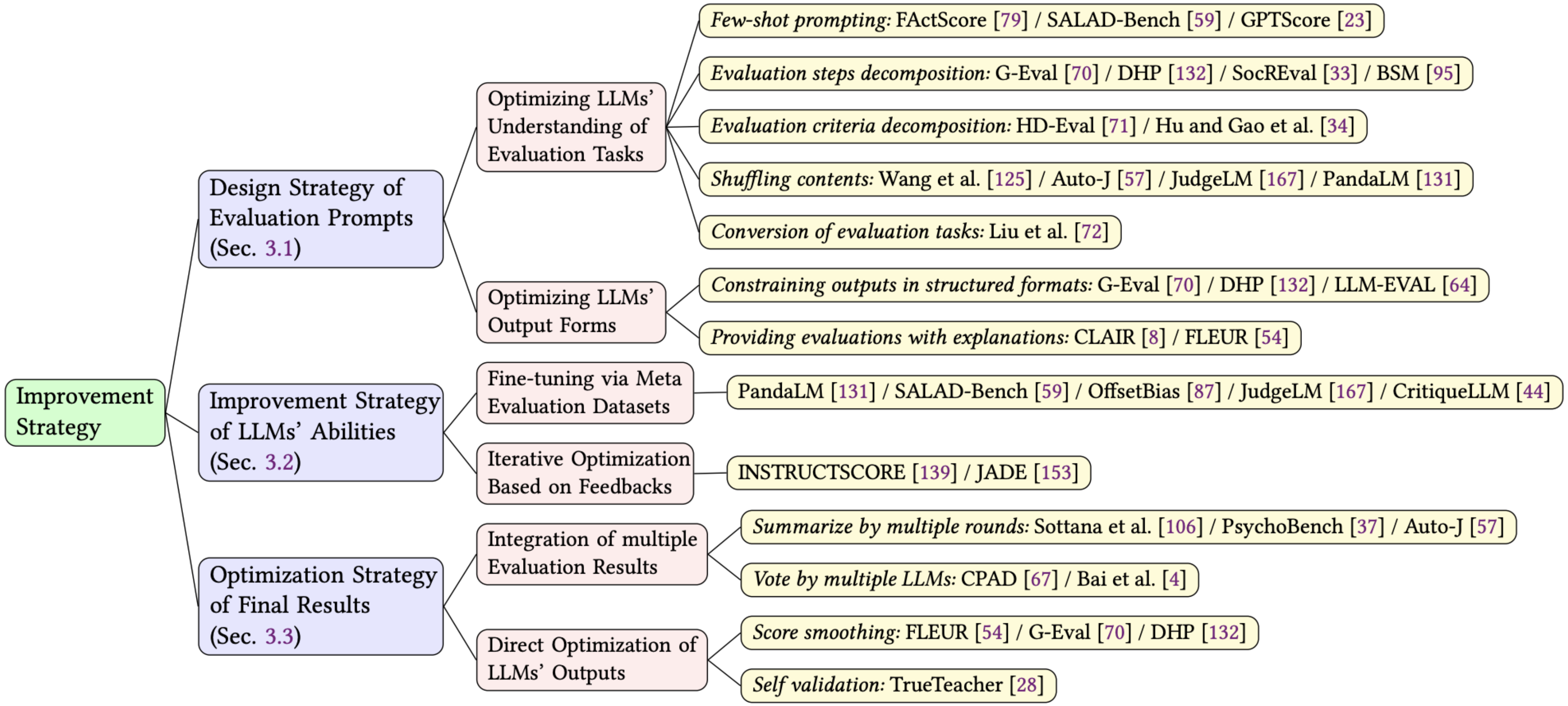
② **长度偏差**
评估器倾向于更冗长的响应

③ **自增强偏差**
LLM评估器可能更偏好由自身生成的响应，尤其是在评估与自身生成的文本进行比较时。

④ **其他偏差**
- **多样性偏差**：对特定群体（如性别、种族等）的偏见。
- **具体性偏差**：倾向于包含权威来源、复杂术语的具体性回答。
- **情绪偏差**：偏好带有特定情绪（如积极、悲伤、愤怒）的响应。

3 How to improve a reliable LLM-as-a-Judge?

$$\mathcal{E} \leftarrow \mathcal{P}_{\mathcal{LLM}}(x \oplus \mathcal{C})$$



3 How to improve a reliable LLM-as-a-Judge?

Design Strategy of Evaluation Prompts (Sec. 3.1)

1. 提示工程优化

① 优化LLM对评估任务的理解
- **提供高质量的few-shot案例：需人工确认**
- 分解评估步骤：G-Eval、DHP使用CoT方法，SocREval采用苏格拉底式方法
- 细化评估标准：**Likert量表**
- **应对未知偏差：**通过内容随机交换，消除位置偏差影响等
- **绝对评分与相对比较**：由于相对比较比绝对评分更可靠，通过局部配对比较实现全局排序

② 优化LLM的输出形式
- 提升输出形式的鲁棒性：特定格式模板（如“X: Y”）,\box{}, JSON格式等
- 增强输出的可解释性：输出评分及相关理由

3 How to improve a reliable LLM-as-a-Judge?

Improvement Strategy of LLMs' Abilities (Sec. 3.2)

2. 提升LLM能力的改进策略

① 通过元评估数据集微调模型
② 基于评估反馈的迭代优化：自动反馈(Reward Model)，人工反馈
③ 选择推理能力和指令跟随能力更强的通用模型

3 How to improve a reliable LLM-as-a-Judge?

Optimization Strategy of Final Results (Sec. 3.3)

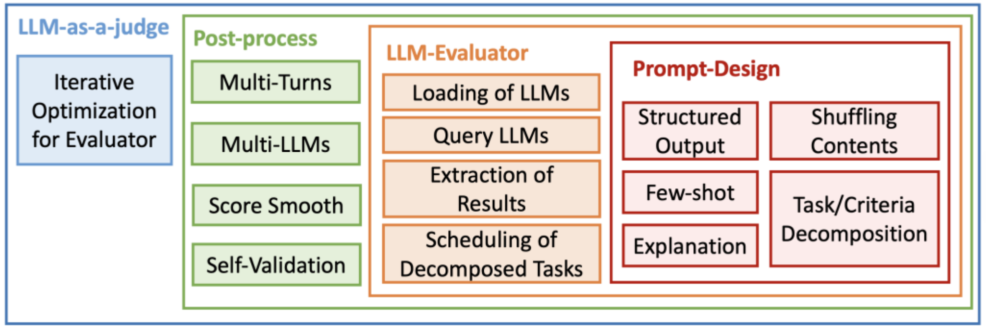
3. 提升LLM能力的改进策略

① 整合多次评估结果
- **多轮评估整合**：对同一内容多次评估并取均值
- **多模型评估整合**：由多个LLM同时评估同一内容并整合结果

② 优化随机性
- **分数平滑**（Score Smoothing）**：** 隐式概率和显式输出的加权平滑
- **自验证（Self-Verification）**：通过自检筛选出稳健的评估结果。

③ 选择推理能力和指令跟随能力更强的通用模型

3 How to improve a reliable LLM-as-a-Judge?



Benchmark	Release Year	Size	Annotation Format	Evaluation Dimension			
				Agreement	Position Bias	Length Bias	Bias Types
MTBench [162]	2023	80	Pairwise	✓	✓	✓	3
Chatbot Arena [162]	2023	30k	Pairwise	✓	✓	✓	3
FairEval [126]	2023	80	Pairwise	✓	✓	✗	1
PandaLM [131]	2023	-	Pairwise	✓	✓	✗	0
LLMEval^2 [156]	2023	2553	Pairwise	✓	✗	✗	0
Shepherd [128]	2023	1317	Score	✓	✗	✗	0
EvalBiasBench [87]	2023	80	Pairwise	✓	✓	✓	6
CALM [142]	2024	4356	Pairwise & Score	✗	✓	✓	12
JudgeBench [113]	2024	-	Pairwise	✓	✗	✗	0
MLLM-as-a-Judge [9]	2024	30k	Pairwise & Score	✓	✗	✗	0
CodeJudge [159]	2024	1860	Score	✓	✗	✗	0
KUDGE [103]	2024	3324	Pairwise & Score	✓	✗	✗	0

Table 1. Benchmark for meta-evaluation of LLM-judge.

3 How to improve a reliable LLM-as-a-Judge?

LLMs	Alignment with Human (n=5106)	Biases Position (n=2633)	Length (n=34)	Concreteness (n=28)	Empty Reference (n=26)	Content Continuation (n=24)	Nested Instruction (n=24)	Familiar Knowledge (n=24)
GPT-4-turbo	61.56	80.49	91.18	89.29	65.38	95.83	70.83	100.0
GPT-3.5-turbo	54.72	68.78	20.59	64.29	23.08	91.67	58.33	54.17
Qwen2.5-7B-Instruct	56.54	63.50	64.71	71.43	69.23	91.67	45.83	83.33
LLaMA3-8B-Instruct	50.72	38.85	20.59	57.14	65.38	75.00	45.83	54.17
Mistral-7B-Instruct-v0.3	55.42	59.78	26.47	67.86	53.85	66.67	37.50	41.67
Mixtral-8×7B-Instruct-v0.1	56.29	59.06	50.00	78.57	42.31	83.33	29.17	83.33
gemini-2.0-thinking	60.80	76.84	94.12	89.29	50.00	100.00	83.33	100.00
o1-mini	60.19	76.73	91.18	89.29	53.85	95.83	75.00	95.83

Table 2. The meta-evaluation results for different LLMs. All the values are percentages.

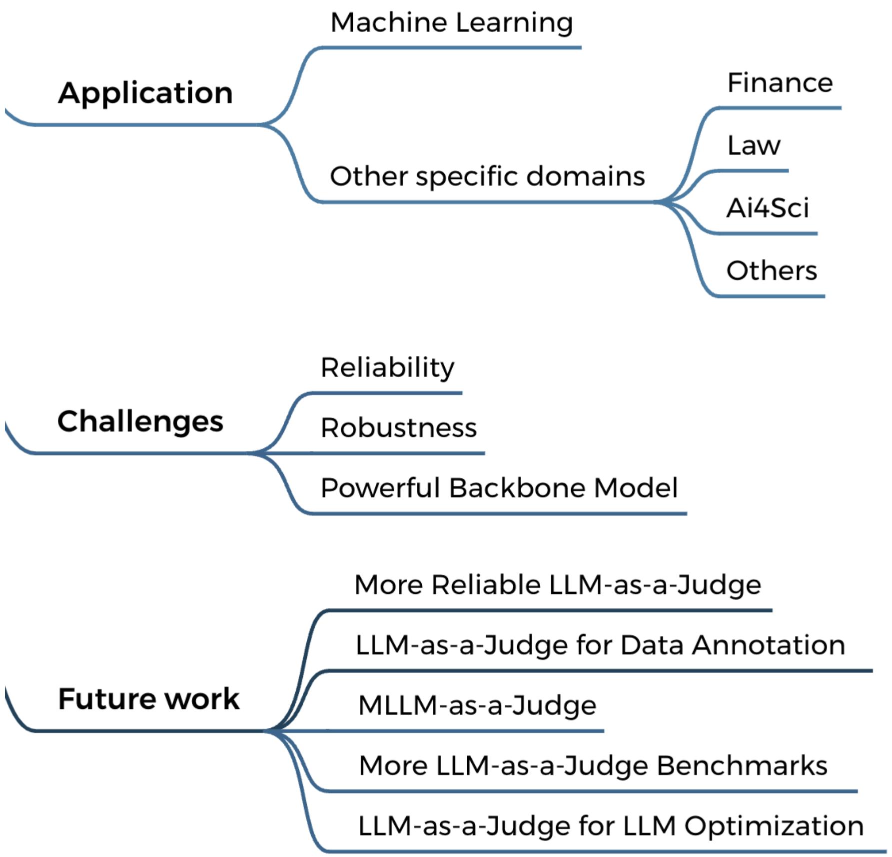
- **LLM能力差异**：GPT-4在自动评估中表现最优，开源模型中Qwen2.5-7B表现突出
- **策略有效性**：多数投票策略（w/ majority@5）在缓解随机性和偏差方面最有效
- 建议在**成对比较任务中**采用**更强大的LLM、内容位置交换和多轮多数投票策略**

3 How to improve a reliable LLM-as-a-Judge?

Although reasoning capabilities are considered foundational for effective judgment, evaluations of reasoning-focused LLMs like o1-mini and Gemini-thinking revealed that they **did not outperform in aligning with human preferences.**

Surprisingly, their performance was slightly inferior to GPT-4, indicating that **advanced reasoning does not necessarily lead to better judgment alignment.**

4 **Why** is LLM-as-a-Judge important?



4 **Why** is LLM-as-a-Judge important?


对创作idea的判断


对稿子质量的判断

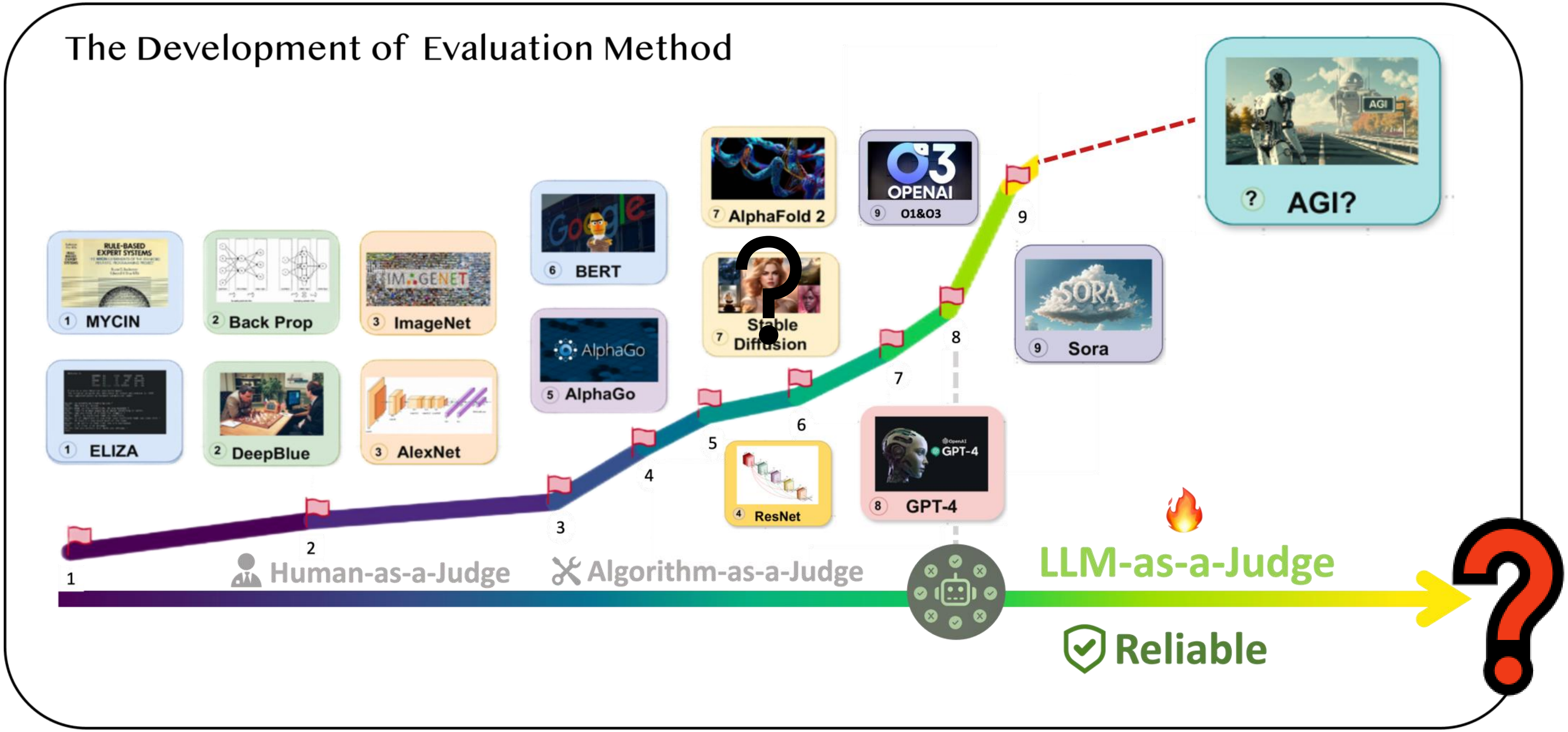

对罪责下的裁定


对案情的判断

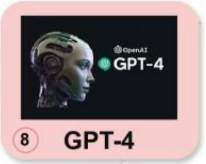

对病情的判断


对投资预期的判断

4 **Why** is LLM-as-a-Judge important?



4 **Why** is LLM-as-a-Judge important?



LLM-as-a-Judge

Scalable Flexible

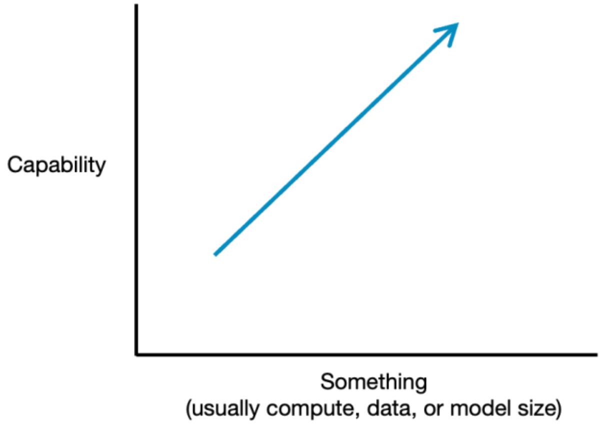
World Model-as-a-Judge

Reliable

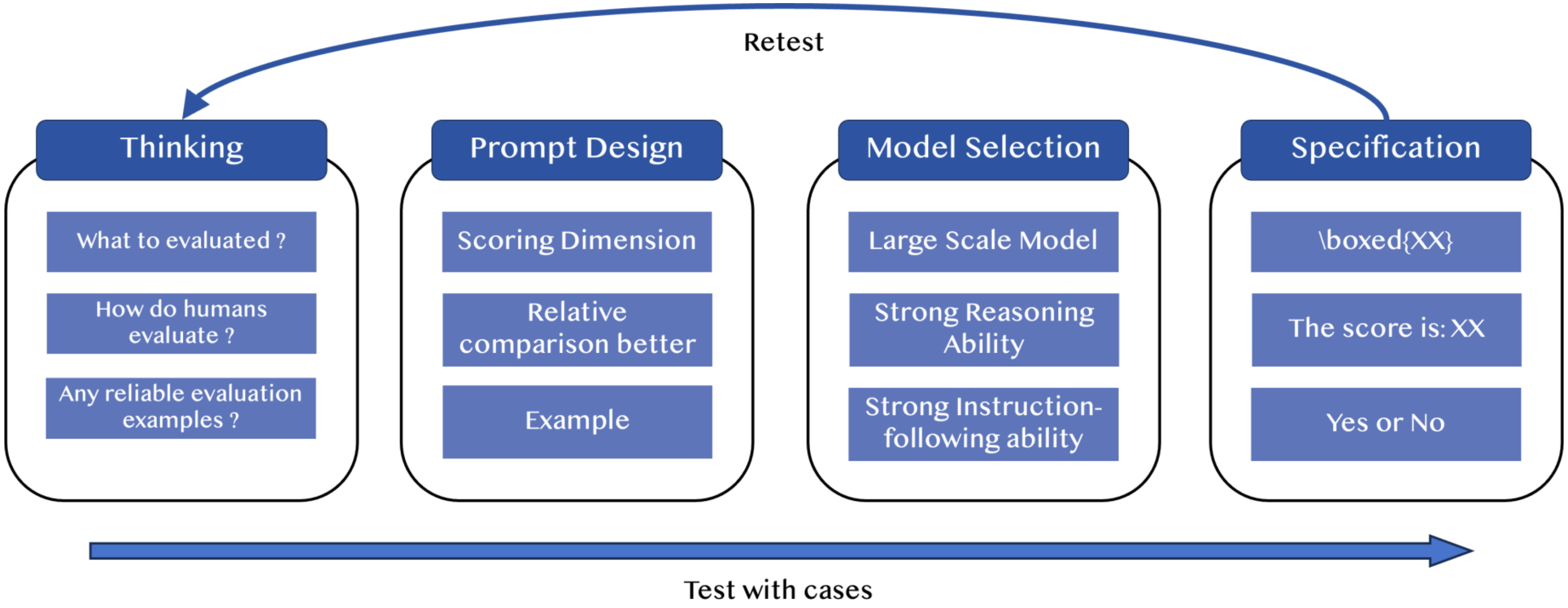
ENVIRONMENT

Feedback

self-evolution



Quick Practice



Conclusion

LLM-as-a-Judge is an LLM-based evaluation framework that excels in **scalability and flexibility** for tasks like scoring and rating, meeting the growing demand for efficient evaluation systems across various fields. However, to fully realize its potential, challenges related to reliability, such as consistency, bias mitigation, and contextual adaptability, must be addressed.

Building a **reliable LLM-as-a-Judge system** requires careful design and optimization at various stages, including dataset creation, model fine-tuning, and standardization of evaluation metrics, to ensure outputs align with human standards and evaluation needs.

As an important AI tool, **LLM-as-a-Judge is expected to see widespread use in academic research, industrial applications, and various societal roles in the future.**

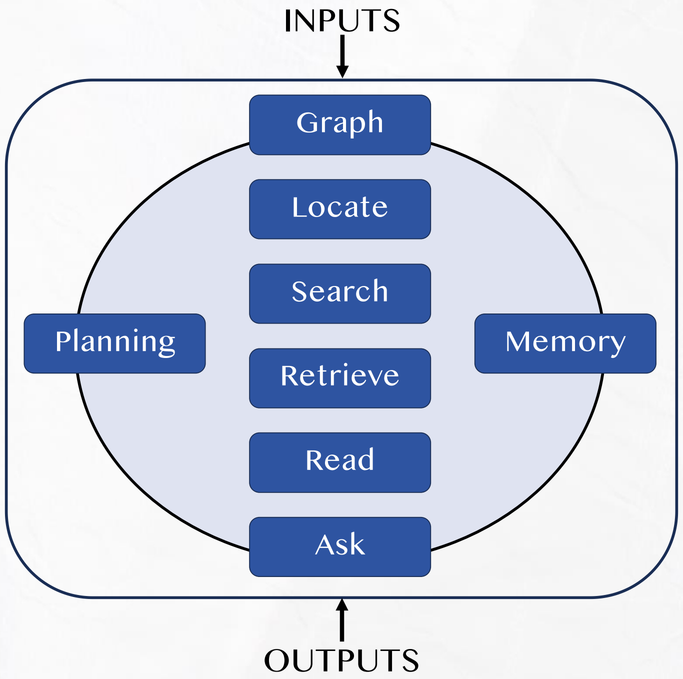
QuantAgent
量化智能体的自我改进策略

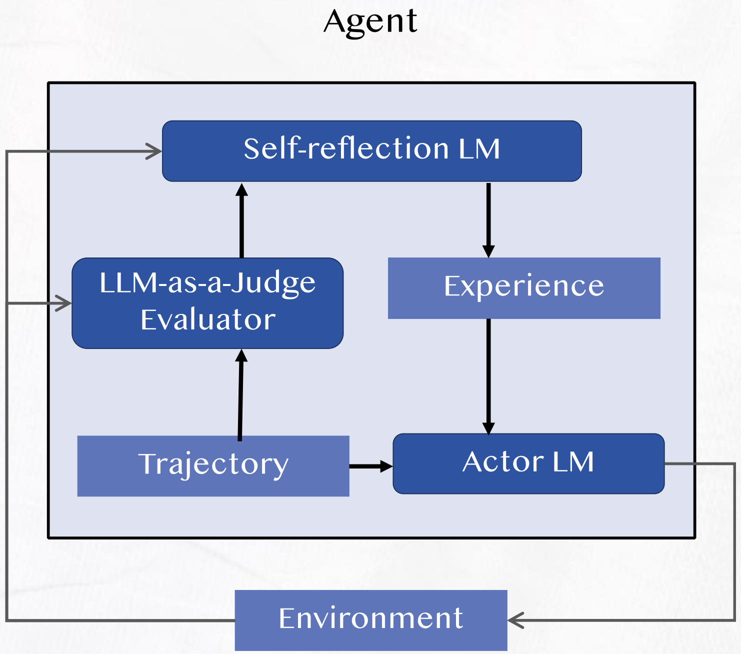
分享人：

王赛卓（Saizhuo Wang）

2.4 Evaluation Pipeline

for Agents





LLM-as-a-Judge can be applied in two ways for agents: **evaluating the entire process** or **assessing specific stages** within the framework. As the agent's brain, LLMs can evaluate like humans, reducing human involvement and balancing thoroughness with effort. Additionally, agents can interact with environments via language and use LLM feedback to guide their actions.

Shinn, Noah, et al. "Reflexion: Language agents with verbal reinforcement learning." *Advances in Neural Information Processing Systems* 36 (2024).
Zhuge, Mingchen, et al. "Agent-as-a-judge: Evaluate agents with agents." *arXiv preprint arXiv:2410.10934* (2024).

LLM-as-a-Judge
在RAG及合成数据中的价值与应用

分享人：

徐铖晋（Chengjin Xu）

Thank You!

顾嘉炜（Jiawei Gu） 徐铖晋（Chengjin Xu） 王赛卓（Saizhuo Wang）

论文 热门推特 热门解读

Contact us: kuvvius@gmail.com

swangeh@connect.ust.hk