SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

北京大学前沿计算研究中心
Center on Frontiers of Computing Studies, Peking University

Benchmarking LLMs' Judgments with No Gold Standard

ICLR 2025




Shengwei Xu*
shengwei@umich.edu

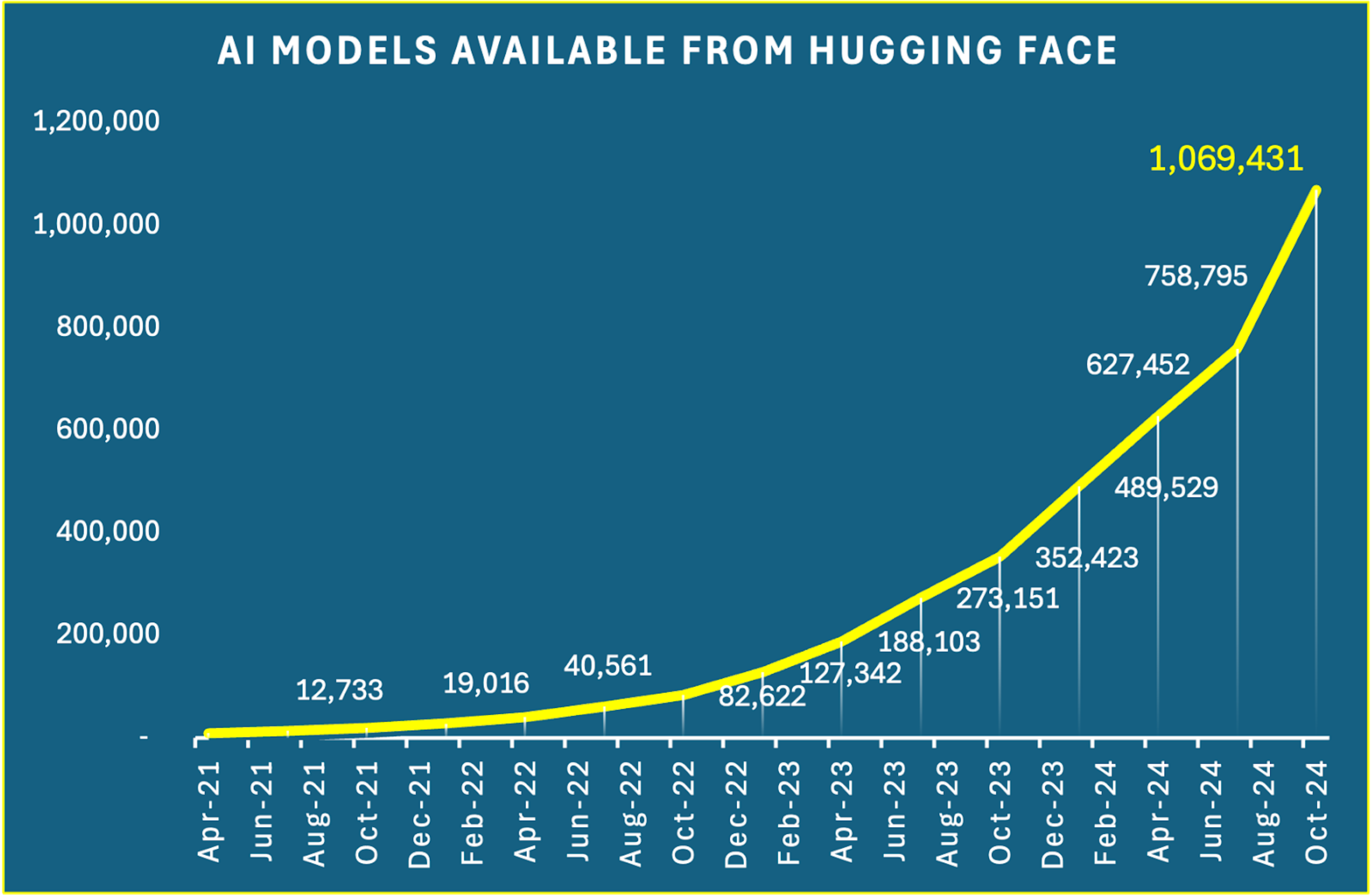
Yuxuan Lu*
yx_lu@pku.edu.cn

Grant Schoenebeck
schoeneb@umich.edu

Yuqing Kong
yuqing.kong@pku.edu.cn

The Thirteenth International Conference on Learning Representations (ICLR2025)

Benchmarking LLMs



Source: https://www.appsoc.com/blog/hugging-face-has-become-a-malware-magnet

Benchmarking LLMs

Open LLM Leaderboard

Comparing Large Language Models in ar

Rank	Type	Model	Ave...	IFE...	BBH	MA...	GP...	MU...	MM...	CO_2...
1		MaziyarPanahi/calme-3.2-instruct-78b	52.08 %	80.63 %	62.61 %	40.33 %	20.36 %	38.53 %	70.03 %	66.01 kg
2		MaziyarPanahi/calme-3.1-instruct-78b	51.29 %	81.36 %	62.41 %	39.27 %	19.46 %	36.50 %	68.72 %	64.44 kg
3		dfurman/CalmeRys-78B-Orpo-v0.1	51.23 %	81.63 %	61.92 %	40.63 %	20.02 %	36.37 %	66.80 %	25.99 kg
4		MaziyarPanahi/calme-2.4-rys-78b	50.77 %	80.11 %	62.16 %	40.71 %	20.36 %	34.57 %	66.69 %	25.95 kg
5		huihui-ai/Qwen2.5-72B-Instruct-abliterated	48.11 %	85.93 %	60.49 %	60.12 %	19.35 %	12.34 %	50.41 %	76.77 kg
6		Qwen/Qwen2.5-72B-Instruct	47.98 %	86.38 %	61.87 %	59.82 %	16.67 %	11.74 %	51.40 %	47.65 kg
7		MaziyarPanahi/calme-2.1-qwen2.5-72b	47.86 %	86.62 %	61.66 %	59.14 %	15.10 %	13.30 %	51.32 %	29.50 kg
8		newsbang/Homer-v1.0-Qwen2.5-72B	47.46 %	76.28 %	62.27 %	49.02 %	22.15 %	17.90 %	57.17 %	29.55 kg
9		ehristoforu/qwen2.5-test-32b-it	47.37 %	78.89 %	58.28 %	59.74 %	15.21 %	19.13 %	52.95 %	29.54 kg
10		Saxo/Linkbricks-Horizon-AI-Avengers-V1-32B	47.34 %	79.72 %	57.63 %	60.27 %	14.99 %	18.16 %	53.25 %	7.95 kg

Benchmarks with Gold-standard References

E.g. Measuring Massive Multitask Language Understanding (MMLU) [Hendrycks et al, 2020]

College Mathematics

In the complex z-plane, the set of points satisfying the equation $z^2 = |z|^2$ is a

(A) pair of points ✗
(B) circle ✗
(C) half-line ✗
(D) line ✓

Benchmarks with Gold-standard References

E.g. Measuring Massive Multitask Language Understanding (MMLU) [Hendrycks et al, 2020]

College Mathematics

In the complex z-plane, the set of points satisfying the equation $z^2 = |z|^2$ is a

(A) pair of points ✗

(B) circle ✗

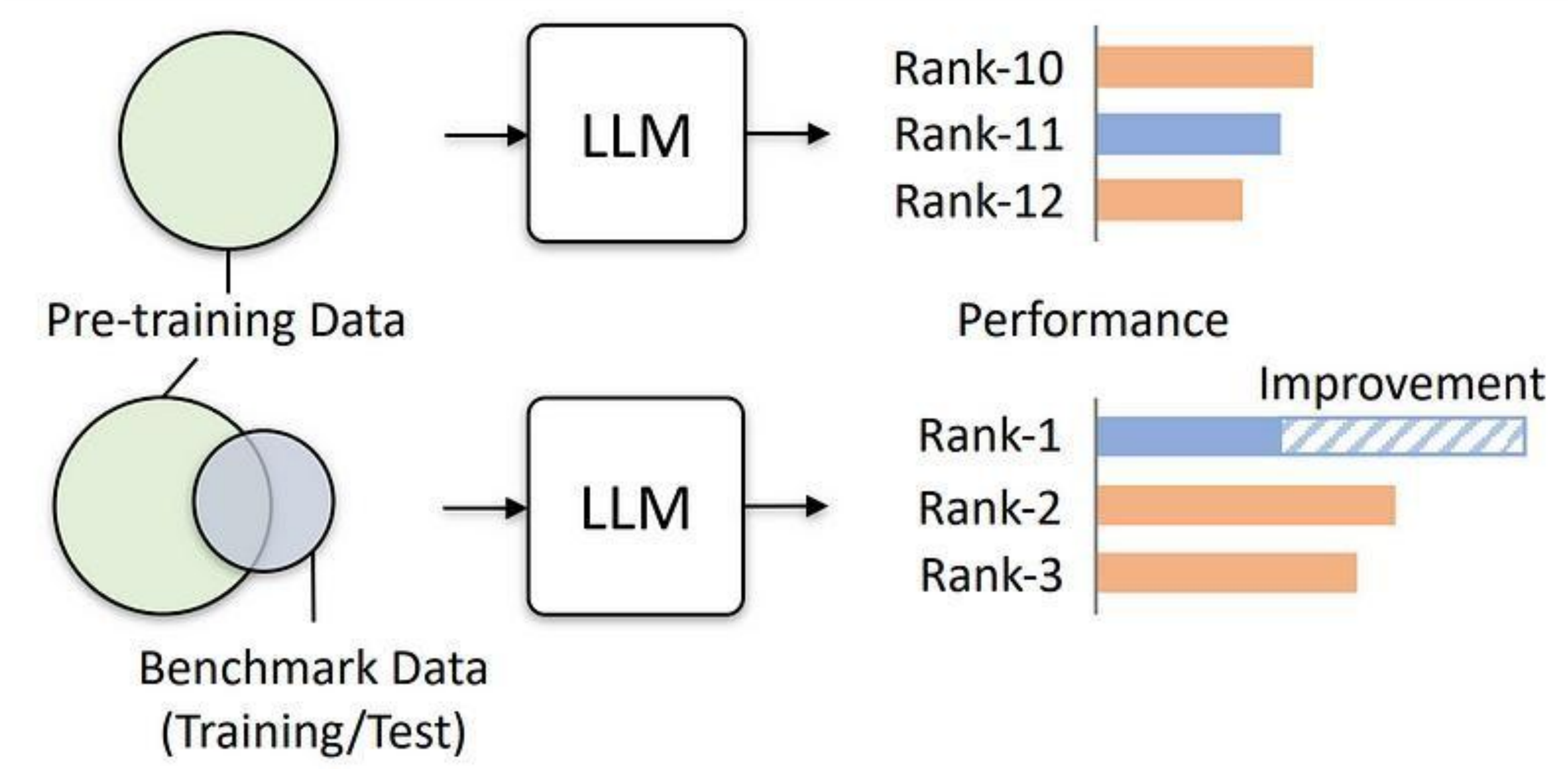
(C) half-line ✗

(D) line ✓

👍 Easy to verify LLMs' outputs

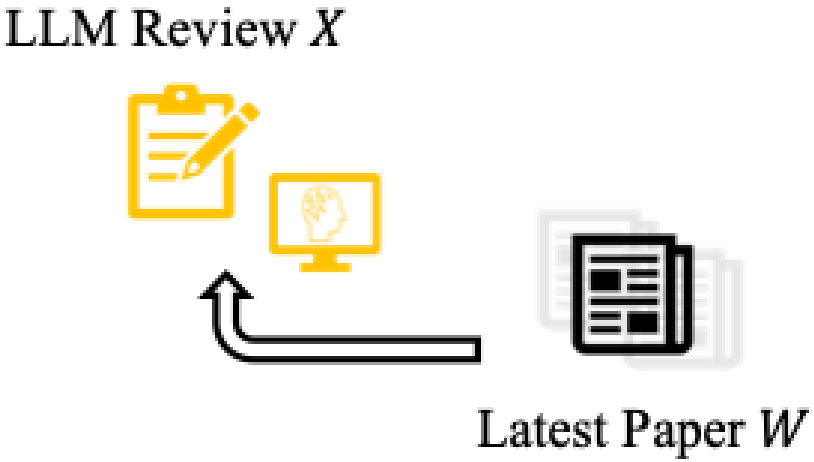
👎 Lack of Subjective Reasoning

Data Contamination



Source: https://arxiv.org/pdf/2311.01964 [Zhou et al., 2023]

Benchmarking with Open-ended Questions



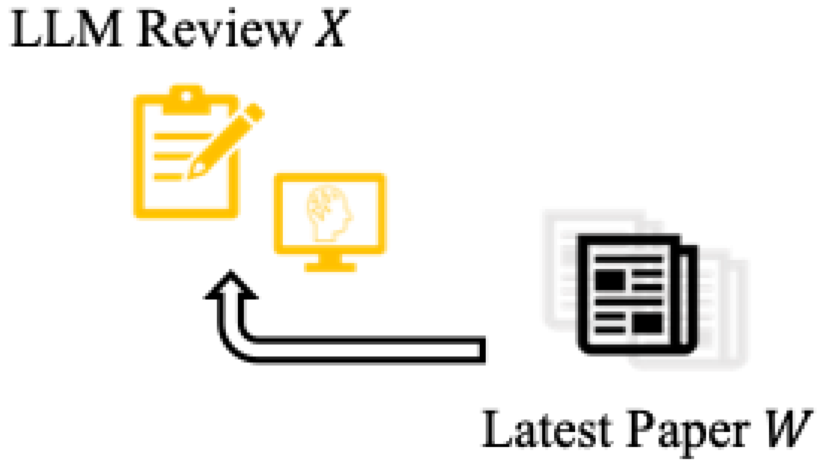
arXiv

Open Review .net

ResearchGate

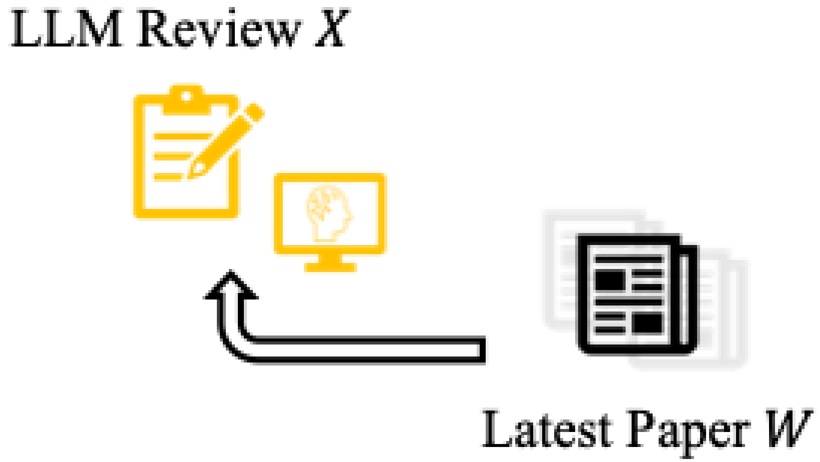
Benchmarking with Open-ended Questions

- 👍 Evaluate both objective and subjective reasoning
- 👍 Circumvents data contamination



Benchmarking with Open-ended Questions

- 👍 Evaluate both objective and subjective reasoning
- 👍 Circumvents data contamination



arxiv

Open Review .net

ResearchGate

Challenge: **No gold-standard** quality response to compare with

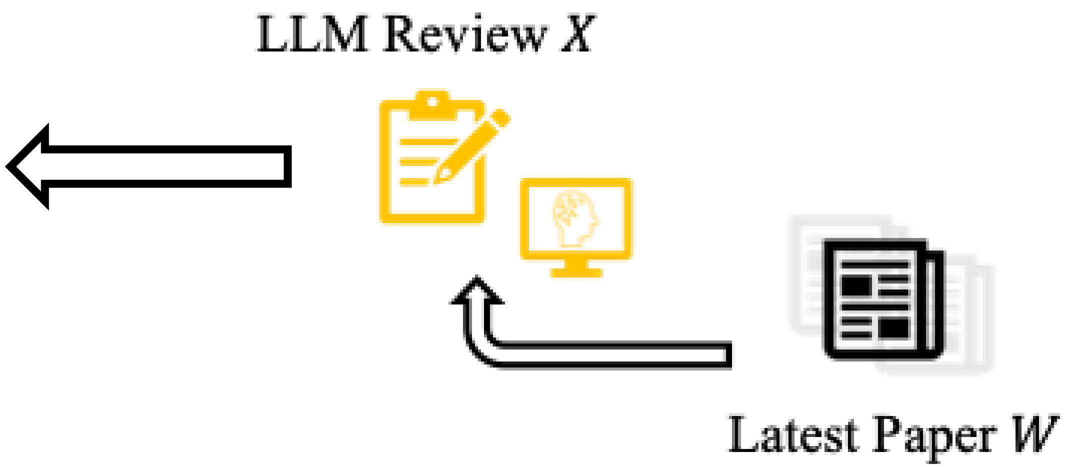
Research Questions

- Can we develop **accurate**, **manipulation-resistant**, and **automated** evaluation metrics for textual responses
- with **no gold standard** reference to compare with?

LLM as an Oracle Examiner[Bai et al, 2023]

User Prompt: You are an expert tasked with **evaluating the quality of a review** for a Machine Learning paper. Your goal is to assess how well the review critiques the paper and provides valuable feedback to the authors, according to the following criteria: **understanding, coverage, substantiation, constructiveness** (Review Quality Indicators [Goldberg et al., 2019, Rooyen et al., 1999])

Output of GPT-4o Examiner



LLM Examiner is not Manipulation-resistant

Meaningless Elongation: adding the same fixed sentences significantly increases the score given by GPT-4o LM examiner.

This section provides an overview of the key contributions, methodologies, and findings presented in the paper. It summarizes the main arguments and highlights the scope and significance of the research conducted. Specifically,

Summary of the Paper:

[Original Text]

In this section, I will discuss the strengths and weaknesses of the paper, focusing on its contributions to the field, methodological rigor, and areas where improvements could be made. The aim is to provide constructive feedback that can help enhance the quality of the work.

This paper has several notable strengths, including [strength summary]. However, there are some areas that could be improved, such as [weakness summary]. Specifically, I discuss the strengths and weaknesses of the paper as follows.

Strength and Weaknesses:

[Original Text]

I will evaluate the clarity of the paper's presentation, the quality of the research, the novelty of the findings, and the reproducibility of the experiments. This assessment will address whether the research is presented in a clear and coherent manner, offers new insights, and can be replicated based on the provided information.

Clarity, Quality, Novelty and Reproducibility:

[Original Text]

This summary encapsulates the main points of the review, reflecting on the paper's contributions, strengths, and areas for improvement. It provides a balanced overview of the paper's impact and offers final recommendations for the authors.

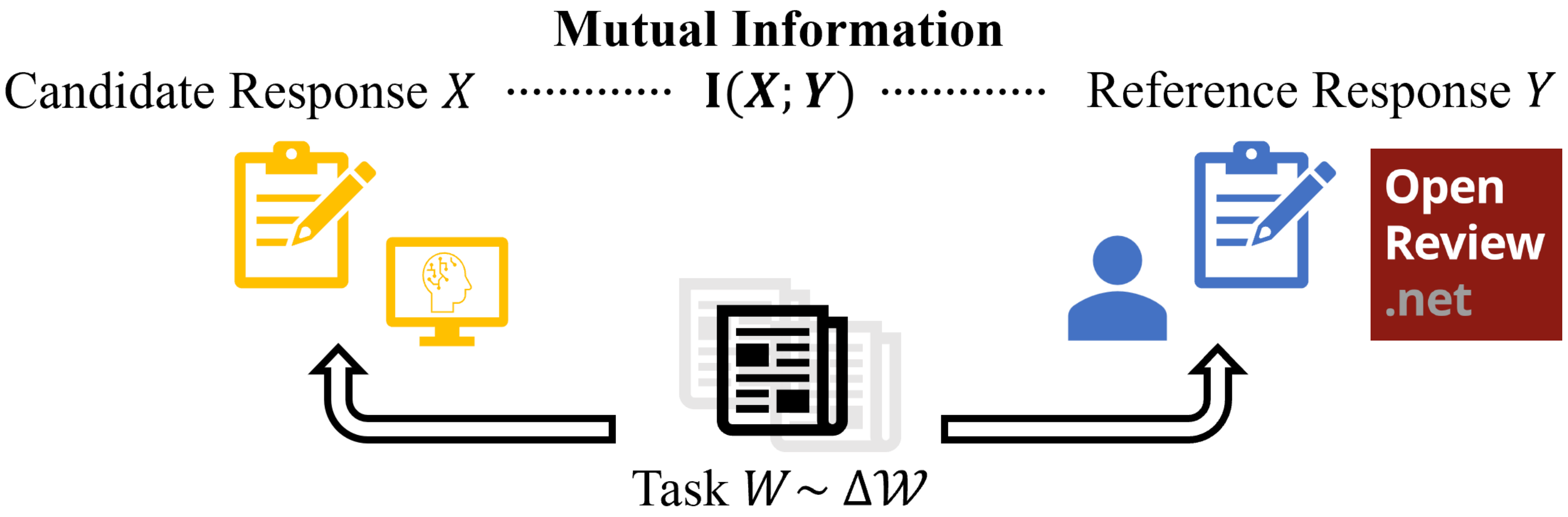
Summary of the Review:

[Original Text]

Inspired by human-subject experiment by [Goldberg et al, 2023]

Our Method: Mutual Information

- **Accurate:** Measuring informativeness
- **Manipulation-resistant:** Data processing inequality



Related Work: Generative Peer Prediction

[Lu, Xu, Zhang, Kong, and Schoenebeck, EC'24]
"Eliciting Informative Text Evaluations with Large Language Models "

- Agent i's report $\tilde{x}_i \in \Sigma$, agent j's (the peer) report $\tilde{x}_j \in \Sigma$
- Score of agent i $= \log \Pr[\tilde{x}_j \mid \tilde{x}_i]$
  - When applying a log scoring rule

Use LLM to estimate

Main Theorem (Informal):

- When the KL-divergence between the real distribution $\log \Pr[x_j \mid x_i]$ and the LLM estimated $\log \Pr_{\mathrm{LLM}}[x_j \mid x_i]$ can be bounded by ϵ
  - And this distribution is common knowledge for all agents
- **Exerting effort & reporting truthfully is $\alpha\epsilon$-Nash equilibrium**
  - α depends on the cost of effort
  - When ignoring the cost of effort, truthful reporting is ϵ-Nash equilibrium

From Information Elicitation to Natural Language Generation (NLG) Evaluation

- Generative Estimator for Mutual Information (GEM)

$$\mathrm{PMI}\left(\tilde{x}_i; \tilde{x}_j\right) = \log \Pr[\tilde{x}_j \mid \tilde{x}_i] - \log \Pr[\tilde{x}_j]$$

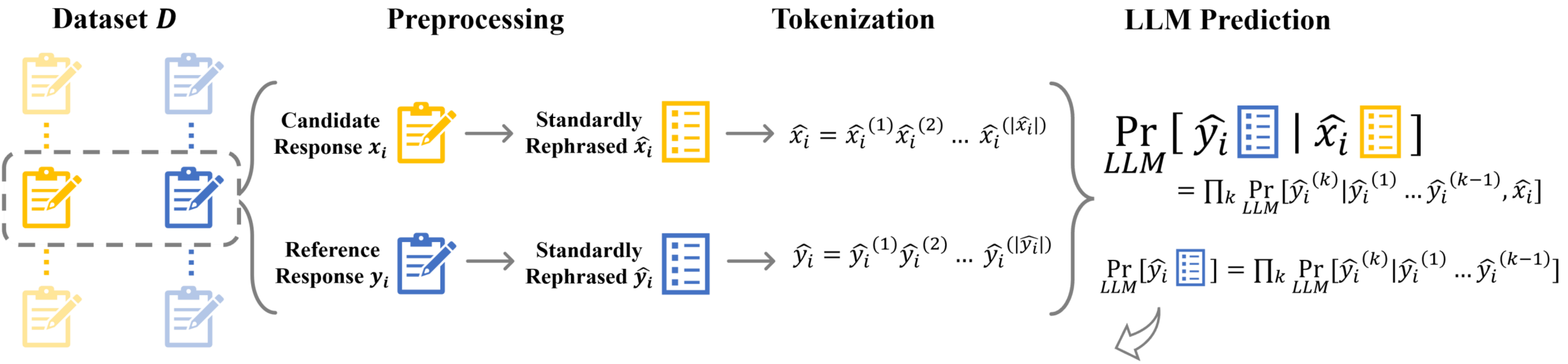
- Generative Estimator for Mutual Information with Synopsis (GEM-S)

$$\mathrm{PMI}\left(\tilde{x}_i; \tilde{x}_j \mid \theta\right) = \log \Pr[\tilde{x}_j \mid \tilde{x}_i, \theta] - \log \Pr[\tilde{x}_j \mid \theta]$$

Implementation of GEM

Following [Lu, Xu, Zhang, Kong, and Schoenebeck, 2024, Yuan, Neubig, and Liu, 2021]

Use the LLM's ability in predicting the next token



Estimated MI: $\mathrm{GEM}(D) = \frac{1}{n}\sum_i \widehat{\mathrm{PMI}}(\hat{x}_i, \hat{y}_i) = \frac{1}{n}\sum_i \log \Pr_{LLM}[\hat{y}_i \mid \hat{x}_i] - \log \Pr_{LLM}[\hat{y}_i]$

Empirical Results: GEM’s Effectiveness

- **Alignment:** Positive correlation with human annotation

peer grading dataset

  - GEM and GEM-S have significant positive Spearman’s correlations with human annotation

- **Sensitivity:** Sensitivity to semantic degradation

ICLR 2023 dataset

  - After semantic degradations, GEM and GEM-S are the **only metrics** that demonstrate significant score decreases.

- **Robustness:** Robustness against manipulation

ICLR 2023 dataset

  - After **manipulations** (e.g. meaningless elongation), GEM and GEM-S are the **only metrics** that demonstrate **no significant score increases**.

GRE-bench (Generating Review Evaluation Benchmark)

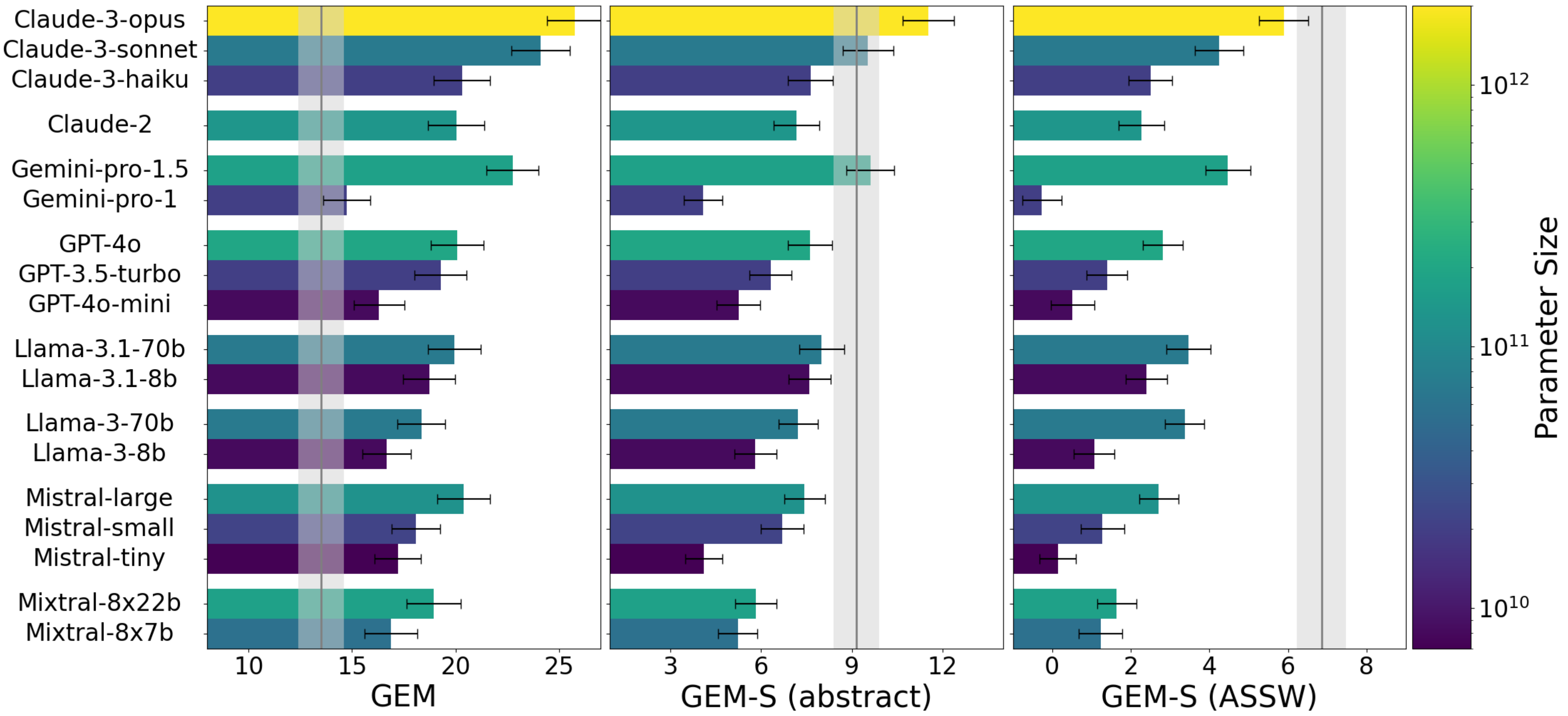
Evaluation Metric + Dataset = Benchmark

- GEM/GEM-S + ICLR Dataset = GRE-bench

Evaluate LLMs’ ability to generate high-quality peer reviews

- Inherit GEM’s accuracy and robustness properties.
- Circumvent data contamination by using the continuous influx of new open-access research papers and peer reviews each year.

GRE-bench on ICLR2023 Dataset



The grey line represents the average human baseline

Conclusion

- Bridge Information Elicitation and NLG evaluation
- Propose GEM/GEM-S for NLG evaluation
    - GEM’s manipulation resistance aligned to GPPM’s incentive compatibility
    - Make necessary changes to be more suitable for the NLG evaluation
    - Validate GEM‘s accuracy and manipulation resistance empirically
- Propose the GRE-bench
    - Inherit GEM’s accuracy and manipulation resistance properties
    - Mitigate data contamination issues

Thank you for your listening!

Paper QR Code

ICLR

International Conference On Learning Representations