第十届北大—清华统计论坛

The 10th Peking–Tsinghua JointStatistics Colloquium (2026)

Prompt Perturbation for Reliable LLM Evaluation over Comparison Graphs

Presented by **Dong Huang**[1]

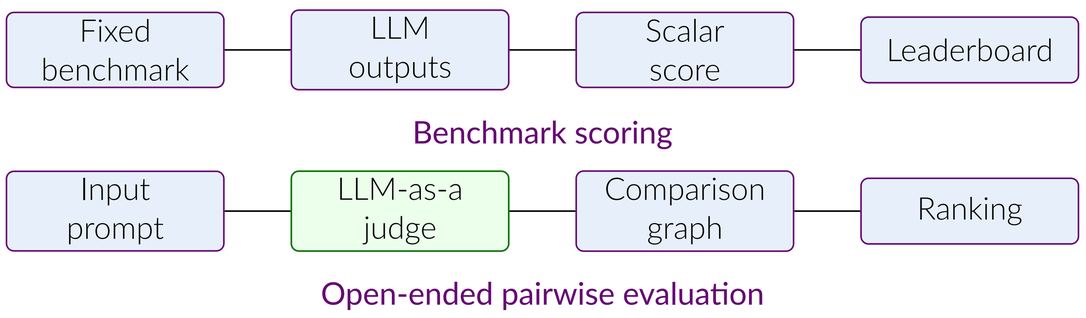
Joint work with Jianbo Sun[1] and Pengkun Yang[1]

Email: hd23@mails.tsinghua.edu.cn

Department of Statistics and Data Science, Tsinghua University[1]

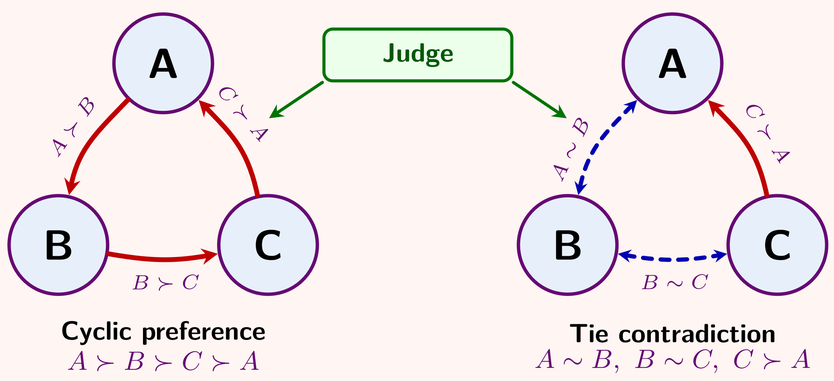
Background and Motivation

- Benchmark scoring: fixed tasks → scalar scores; useful but limited for **open-ended** tasks [1].
- Pairwise evaluation: compare two responses with an LLM judge, then aggregate preferences [2].
- The key object is a comparison graph before it becomes a final ranking.



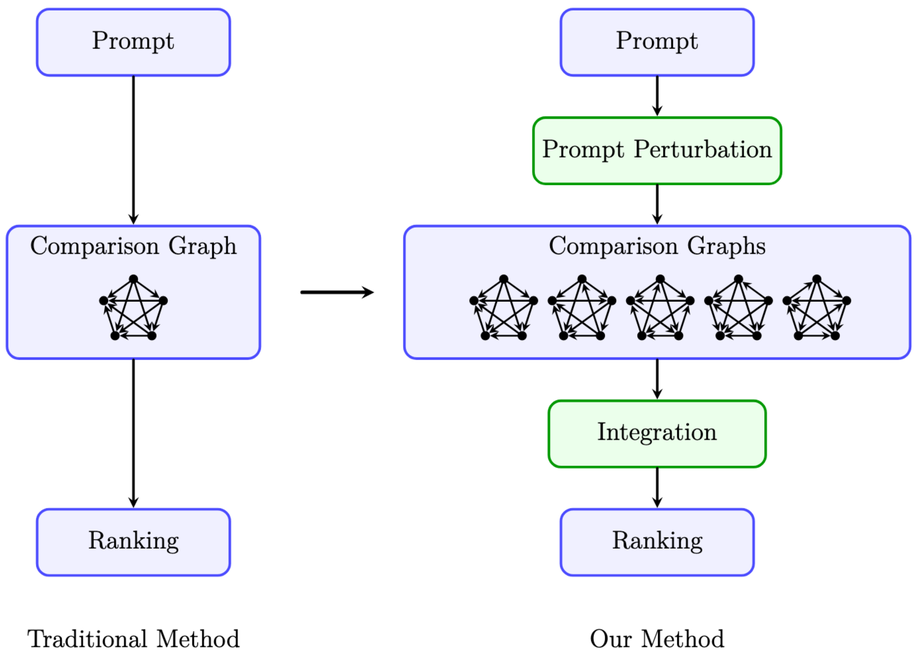
Challenge: Structural Inconsistency

- $A \succ B$: model A performs better than model B under the judge model.
- $A \sim B$: the judge model cannot distinguish between models A and B.
- Cycle: $A \succ B \succ C \succ A$.
- Tie contradiction: $A \sim B,\ B \sim C, C \succ A$.



*How can we design a pairwise evaluation pipeline that **improves the structural consistency** of the induced comparison graph?*

Prompt Perturbation and Integration

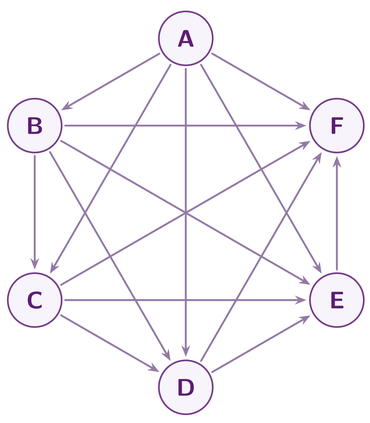


Prompt perturbation

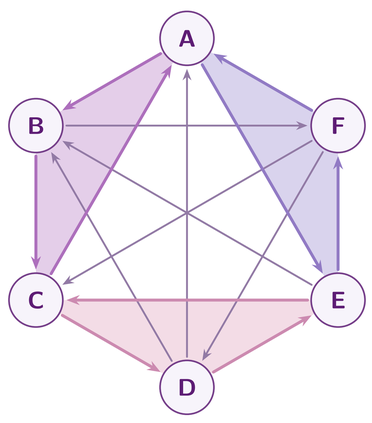
- For each base prompt x_i, generate semantic-preserving rewrites $x_{i,1}, \ldots, x_{i,m}$ that keep the task intent, constraints, and entities unchanged.
- Judge the same model pairs under the original prompt and each rewrite, giving aligned comparison graphs $G_{i,0}, G_{i,1}, \ldots, G_{i,m}$ on the same model set.

Integration

- Treat each graph as one graph-valued proposal and score its local inconsistency by **counting number of cycles**.
- Keep the **top-K** graphs with smallest number of cycles, and fit Bradley–Terry ranking models [3].



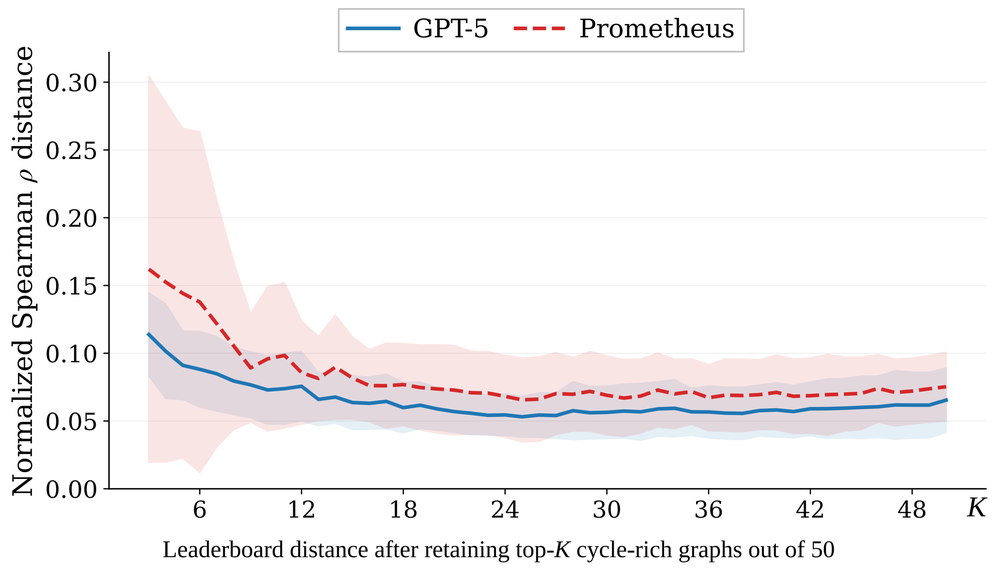
Reliable



Unstable

Experimental Results

- MT-Bench: 8 categories, 80 questions.
- $n = 20$ target LLMs.
- 5 prompt instances per question: original + 4 semantic perturbations.
- LLM-as-a-judge: GPT-5 and Prometheus.
- Metric: normalized Spearman distance.
- Ground truth ranking: arena leaderboard.



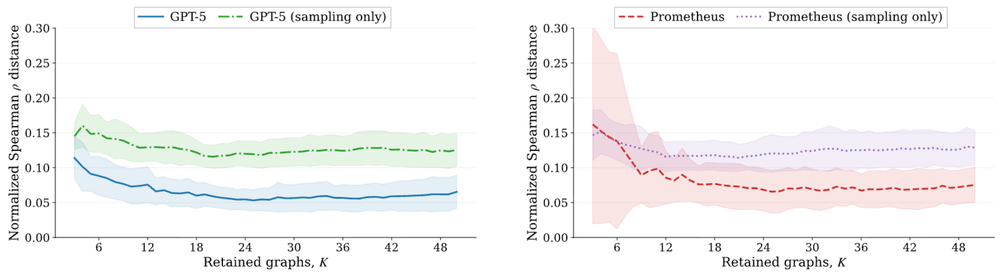
Leaderboard distance after retaining top-K cycle-rich graphs out of 50

Judge	**Trunc25**	NoTrunc	Sample25
GPT-5	**0.056**	0.068	0.120
Prometheus	**0.070**	0.083	0.121

- Best budget: $K \approx 25$.
- Best performance: Trunc25 (our method) wins for both judges.
- Main gain:

  prompt perturbation + cycle truncation ≻ prompt perturbation

  prompt perturbation + cycle truncation ≻ cycle truncation



Theoretical Results

- Latent order: $1 \succ 2 \succ \cdots \succ n$, with transitive graph G^*.
- Each observed graph G_k flips an edge with probability $\frac{1}{2} - p$.
- Integrate t graphs by pairwise majority vote to obtain graph $\tilde{G}$.

Theorem 1: majority voting without truncation

For any constant $\epsilon > 0$,

$$t \geq \frac{(4+\epsilon)\log n}{\log(1/(1-4p^2))} \quad \Longrightarrow \quad \mathbb{P}(\tilde{G} = G^*) = 1 - o(1).$$

Below the matching $(4-\epsilon)$ threshold, $\tilde{G}$ contains a directed triangle with high probability.

Theorem 2: truncation on cycles

Under the conditional law $G \mid \{G \text{ has no directed triangle}\}$ [4],

$$t \geq \frac{(2+\epsilon)\log n}{\log(1/(1-4p^2))} \quad \Longrightarrow \quad \mathbb{P}(\tilde{G} = G^*) = 1 - o(1).$$

Takeaway. Cycle truncation improves the leading constant in this stylized recovery threshold.

Our Solution to the Challenge

Prompt perturbation creates multiple comparison graphs, while cycle-count truncation **removes structurally inconsistent ones**, leading to more reliable rankings for open-ended LLM evaluation.

References

[1] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging LLM-as-a-judge with MT-Bench and Chatbot Arena," in *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[2] W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, H. Zhang, B. Zhu, M. Jordan, J. E. Gonzalez, and I. Stoica, "Chatbot Arena: An open platform for evaluating LLMs by human preference," in *Proceedings of the 41st International Conference on Machine Learning*, ser. *Proceedings of Machine Learning Research*, vol. 235, 2024.

[3] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. The method of paired comparisons," *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.

[4] C. L. Mallows, "Non-null ranking models. I," *Biometrika*, vol. 44, no. 1/2, pp. 114–130, 1957.