Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ29492

Judicial Hallucination: Investigating and Understanding the Unreliability of LLM-as-a-Judge

Zhenzhou Liu

Beijing 21st Century School, Beijing, China
zzlucluu@163.com

Abstract. As the application of Large Language Models in the Artificial Intelligence field broadens, the paradigm "LLM-as-a-Judge", which uses LLM evaluation on other Artificial Intelligence system output, has become an efficient and economical alternative for human evaluation. However, the credibility of this paradigm is challenged due to the hallucination of LLM. This paper will summarize these failure modes in the evaluation system as Judicial Hallucination, including both fact fabrication and cognitive biases. This paper will overview the unreliability of the LLM-as-a-Judge system. First, this paper classifies the Judicial Hallucination into two types: hallucination in content and cognitive and heuristic biases. Next, this paper reviews the key metrics and benchmarks to quantify these failure modes. Then, we analyze several multi-agent collaboration frameworks. In conclusion, ensuring the reliability of LLM-as-a-Judge is a crucial yet unsolved challenge, requiring multi-field methods to mitigate hallucination. This paper provides theoretical frameworks to deal with the unreliability of LLM-as-a-Judge.

Keywords: LLM-as-a-Judge, Hallucination, Cognitive bias, AI security, Multi-Agent Collaboration Framework

1. Introduction

As the LLM develops rapidly, one of the most attractive LLM applications is to use LLM-as-a-Judge to evaluate other AI models' output [1]. With uniquely designed evaluation prompts, LLM-as-a-Judge is one type of LLM that scores and arranges the quality of any content output produced by LLM models, like chatbots, question answering systems, and AI agents [2].

The popularization of this paradigm is due to several reasons. First, compared to expensive and unscalable human evaluation, LLM-as-a-Judge can process vast data outputs rapidly and economically [3]. Second, LLM-as-a-Judge could simulate human evaluation by using its powerful reasoning ability when judging open tasks, like creative writing [2]. Third, LLM-as-a-Judge can generate structured explanatory feedback to promote model iteration, and has been applied in many scenarios, including testing chatting AI, evaluating Retrieval-Augmented Generation (RAG) systems, and rating content quality [1].

Problem and Gaps: However, LLM-as-a-Judge has a severe problem: using a tool with its inherent flaws to evaluate the flaws of another similar tool. All LLMs have a well-known core failure mode: Hallucination, which refers to models generating fluent and seemingly credible

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.TJ29492

incorrect facts, inconsistency with the given source, and fabricated content [4]. When hallucination appears in the LLM evaluators, the result will be more severe.

The unreliability of LLM-as-a-Judge will lead to severe consequences: An inaccurate evaluation score may mislead researchers about the ability of the model; a hallucinated LLM-as-a-Judge might fabricate non-existent cases and learn based on these fabricated cases [5]; a hallucinated LLM-as-a-Judge might ignore models with incorrect medical methods, which harms patients' life security.

Current studies have recognized and researched several failure modes of LLM-as-a-Judge [6]. However, few have studied the integrated problems of all biases and hallucinations. This paper suggests that these biases and hallucinations should be redefined and integrated as specific manifestations of Judicial Hallucination in different dimensions.

Contribution: This paper will conclude the unreliability of LLM-as-a-Judge as "Judicial Hallucination", including hallucination in content and cognitive and heuristic biases. This definition involves all failure modes resulting from groundless evaluations that fail to evaluate based on evaluating content or evaluating standards.

This paper will frame a systematic framework about the hallucination of LLM-as-a-Judge, providing ways to analyze the unreliability of LLM-as-a-Judge. The framework for this paper will be:

• Defining core metrics: by constructing metrics such as hallucination rate and alignment rate, this paper provides systematic metrics for evaluating the reliability of LLM-as-a-Judge on different tasks and contexts, offering a consistent tool for assessing unreliability.

• Constructing unified classification methods: through unifying the classification of hallucination types and bias types, this paper aims to clarify their underlying mechanisms, helping researchers to find the origins of the unreliability in LLM judgements.

• Reviewing detection and mitigation mechanisms: by summarizing and analysing different mitigation mechanisms such as multi-agent collaboration and fine-tuning, this paper provides practical ways for reducing Judicial Hallucination in LLM-as-a-Judge evaluations.

Significance: This paper highlights the significance of systematically understanding the unreliability of LLM-as-a-Judge as a whole. Through this framework, this paper provides a comprehensive perspective on the unreliability of the LLM-as-a-Judge system, thereby promoting the development of a more trustworthy, reliable, and robust AI evaluation system.

2. Core metrics

To systematically research the unreliability of LLM-as-a-Judge, this paper will construct a standard for measuring different failure modes of LLM-as-a-Judge.

2.1. Hallucination and factual consistency metrics

• Hallucination Rate: This metric measures the frequency of generating explanatory feedback with factual error, fabricated information, and invalid citation by LLM-as-a-Judge. For example, research shows that GPT-3.5 produced 39.6% of non-existent references when conducting reviews for a common condition such as rotator cuff disease [7]. The formula of the hallucination rate is:

$$Hallucination\ Rate = \frac{N_{Hallucinated\ Judgements}}{N_{Total\ Judgements}} \times 100\% \quad (1)$$

DOI: 10.54254/2755-2721/2026.TJ29492

The number of Hallucinated Judgements means the number of judgements containing hallucinated information. The number of Total Judgements means the total number of judgements.

• Factual Consistency: This metric measures whether the judgement of LLM-as-a-Judge is consistent within the context [8]. A common method is to include outputs of LLM-as-a-Judge as a series of factual claims, and checks the credibility of each claim [8].

2.2. Bias metrics

• Position Bias Score: This metric observes the bias of position preference in the content of LLM-as-a-Judge by exchanging the position of two evaluating answers [9]. Related metrics involve:

Position Consistency: This metric measures whether LLM-as-a-Judge will choose the content with the same position [9]. Lower consistency implies a higher position bias.

• Self-Preference Bias Score: A metric that measures whether LLM-as-a-Judge will score outputs from itself with a higher score [10]. "Equal Opportunity" is a method that compares the discrepancy of evaluation made by LLM-as-a-Judge toward content created by other LLMs and by itself, after controlling the ground-truth labels provided by human evaluators [10]. The formula of Equal Opportunity is:

$$Bias_{self} = P\left(Judge_{pref} \middle| Human_{pref},\ Self\right) - P\left(Judge_{pref} \middle| Human_{pref},\ Other\right) \tag{2}$$

2.3. Metrics related to humans

• Alignment Rate: This metric measures the percentage of LLM-as-a-Judge that will make the same evaluation as a human [11].

• Correlation Coefficients: When evaluation results are consecutive scores, the Pearson Correlation Coefficient can evaluate the linear relationship between scores from LLM-as-a-Judge and human [12].

• Cohen's κ: Compared to raw alignment, chance-corrected measures such as Cohen's κ provide a more robust estimation of reliability for alignment rate between two evaluators (human or LLM) by excluding alignment caused by random chance [13, 14].

To show these core metrics systematically, Table 1 presents these metrics:

Table 1. Core metrics of reliability for LLM-as-a-Judge

Metric	Definition/Formula	Evaluation aspect	Main use cases	Key limitations
Hallucination Rate [7]	Proportion of generated content that contains hallucination	Reliability and veracity of content	Evaluating the frequency of generating explanatory feedback with factual error	Requiring expensive human annotations to identify hallucination
Factual Consistency [8]	Proportion of content consistent with the source document	Consistency and reliability of judgements made in the given context	Evaluating the consistency of LLM-as-a-Judge systems or evaluated models in RAG systems	Claim extraction and verification themselves may not be entirely accurate
Position Consistency [9]	Proportion of unchanged judgments when answer positions are swapped	Robustness to presentation order	Impact of prompt templates on position bias in A/B testing	Unable to distinguish between specific preference for the first or last position

DOI: 10.54254/2755-2721/2026.TJ29492

Table 1. (continued)

Self-Preference Bias Score [10]	Difference in conditional probabilities under equal opportunity	Degree of preference for self-generated content	Evaluating self-reinforcement risk in Reinforcement Learning from Human Feedback reward model	Requiring high cost for comparison with human evaluations
Alignment Rate [11]	$\frac{N_{\text{agree}}}{N_{\text{total}}} \times 100\%$	Alignment with human judgments	Quickly measuring alignment with human preferences	May be overestimated due to chance when the number of options is small
Pearson Correlation Coefficients [12]	$\frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$	Linear relationship with human scores	Checking whether scores given by LLM-as-a-Judge follow the trend of human scores	Limited to linear relationships and sensitive to outliers
Cohen's κ [13, 14]	$\frac{P_o - P_e}{1 - P_e}$	Alignment with human beyond chance	Evaluating the consistency between evaluators (human or LLM) in classification tasks	Limited to two evaluators (human or LLM) and sensitive to class imbalance

3. Judicial hallucination classifications

This paper divides Judicial Hallucination into two types: hallucination in content and cognitive and heuristic biases.

3.1. Hallucination in content

This type of hallucination concentrates on the credibility of generated content by LLM-as-a-Judge.

3.1.1. Extrinsic hallucination

The extrinsic hallucination is content that LLM-as-a-Judge generated training data that uses facts or citations that cannot be verified in the available sources. It is a dangerous hallucination, since it usually presents itself as a credible and reliable way, while its basis is completely fake. This behaviour leads LLM-as-a-Judge to produce unfounded evaluations that appear reasonable, posing a significant risk of misleading developers [7].

3.1.2. Intrinsic hallucination

The intrinsic hallucination is that the evaluation output of LLM-as-a-Judge is inconsistent with the original source material provided. It can be divided into three types:

• Instruction inconsistency, which means LLM-as-a-Judge does not follow the restrictions in the instruction [15].

• Context inconsistency, which means LLM-as-a-Judge outputs information that contradicts the source material [15].

• Logical inconsistency, which means LLM-as-a-Judge provides a logical contradiction when it is reasoning about a task [15].

DOI: 10.54254/2755-2721/2026.TJ29492

3.2. Cognitive and heuristic biases

This type of failure mode originates from the unsound reasoning that LLM-as-a-Judge employs, relying on false heuristics. This type of bias is the cause of false reasoning without a factual basis.

3.2.1. Presentation-related biases

This type of bias is that LLM-as-a-Judge is affected by misleading information.

• Position bias: As mentioned in 2.2, Position bias means that LLM-as-a-Judge might judge the answer by the presentation sequence, instead of judging based on the quality of the content [9]. This bias appeared more when the quality between two choices is similar [9].

• Verbosity bias: LLM-as-a-Judge tends to give a high score for the longer and more detailed answer, even when the quality of this answer is comparable to the more concise one [16].

• Beauty bias: LLM-as-a-Judge tends to give a high score for answers that are more pleasing and organized [17].

3.2.2. Informative and authoritative biases

This type of bias is that LLM-as-a-Judge is affected by misleading information.

• Misinformation Oversight bias: LLM-as-a-Judge cannot distinguish the false facts in the evaluating content [17]. This bias usually occurs when LLM-as-a-Judge does not have knowledge in related fields, or relies on the fluency of the content and fails to do fact-checking [17].

• Authority bias: LLM-as-a-Judge tends to give a high score on content using citations or expertise words, even if these citations and words are fabricated or not directly related to the content [17].

3.2.3. Self-Preference bias

This type of bias, as mentioned in 2.2, means LLM-as-a-Judge will give a high score for models similar to the judge itself, as LLM-as-a-Judge might be similar to content generated by itself, giving a high-quality signal to this content [10].

To provide an analytical framework for understanding the unreliability of LLM-as-a-Judge, Table 2 concludes these classifications:

Table 2. Classifications of unreliability for LLM-as-a-Judge

Type	Class	Definition	Example	Key Findings
Hallucination in content	Extrinsic hallucination [7]	Generated content that is invalid in source material	Citing an invalid paper to support its statement	Hallucination rate of the model Bard in medical reviews can reach 91.4%
	Intrinsic Hallucination [15]	Generated content that contradicts contexts, instructions, and logic in the source material	Evaluating the creativity of the content, while the instruction is to evaluate the grammar of the content	Including three main types: Instruction inconsistency, Context inconsistency, and Logical inconsistency
Cognitive and heuristic biases	Position bias [9]	Judgments affected by the presenting sequence of the answers, instead of answers quality	Choosing A when giving answers with a sequence “A and B”; choosing B when given answers in the sequence “B and A”	Position bias increases notably when the quality difference between two answers is small

DOI: 10.54254/2755-2721/2026.TJ29492

Table 2. (continued)

Verbosity bias [16]	Giving a high score to the longer and more detailed answer	Choosing the answer with more details, while both answers have the same accuracy	Higher human alignment occurs when humans prefer a longer answer
Beauty bias [17]	Giving a high score for the tidier and clearer answer	Choosing the answer using Markdown and symbols, while answers have similar quality	Beauty bias will be more influential when the quality of answers is similar
Misinformation Oversight bias [17]	Unable to recognize false facts and invalid information in content	Evaluating an answer with obvious scientific error as "accurate and flawless"	This bias can be reduced by providing more effective knowledge to LLM-as-a-Judge systems
Authority bias [17]	Giving a high score to answers that cite sources or use an authoritative tone, without checking its validity	An answer with fabricated citations is rated highly due to its "extensive references"	Human evaluators and nearly all LLM-as-a-Judge systems perform worse than random baseline under reference perturbation
Self-Preference bias [10]	Giving a high score to answers generated by itself or models in the same series	A GPT-4 judge will evaluate outputs from GPT-4 with a higher score than outputs from other models	The main cause might be the model's preference for low-perplexity, which means more familiar outputs

4. Detections and mitigation mechanisms

4.1. Benchmark and evaluation framework

4.1.1. Judicial evaluation benchmarks

These benchmarks aim to test the ability of LLM-as-a-Judge to evaluate challenging scenarios.

• JudgeBench: The core of this benchmark is to evaluate factual and logical correctness by using response pairs with one answer objectively correct and the other one subtly flawed [18].

• Prometheus: An open-source LLM-as-a-Judge, adapted from a large-scale and high-quality dataset, that can evaluate based on standards provided by users [19].

4.1.2. Hallucination detection benchmarks and methods

These benchmarks aim to test the hallucination on the content of LLM-as-a-Judge.

• HalluLens: This kind of benchmark provides a clear classification and high-quality dataset, enabling researchers to train and evaluate the hallucination detection models [20].

• Technical detection signals: By using the perplexity, semantic similarity, and response length to detect hallucination of models [6].

4.2. Simple mitigation mechanisms

4.2.1. Prompt engineering

Prompt engineering is the lowest cost and the easiest mitigation mechanism. Its core is to make reliable judgments based on clear instructions. Two ways of prompt engineering are given as

DOI: 10.54254/2755-2721/2026.TJ29492

examples: First, giving rubrics and structured outputs can promote the consistency of evaluation [21]; second, generating chain-of-thought (CoT) can increase the stability and consistency with humans [21]. However, on simple tasks, overly complex CoT may lower the alignment with humans [21].

4.2.2. Framework and data intervening

• Retrieval-augmented generation: By providing another external, complete, credible database, LLM-as-a-Judge can evaluate based on the external information, mitigating extrinsic hallucination [22]. However, it is limited by the completeness of the database.

• Fine-Tuning: It can help LLM-as-a-Judge to adapt to new data, but it might also damage the original trained database, causing overfitting and catastrophic forgetting [23]. Prometheus is an example of an expert model resulting from targeted fine-tuning [19].

4.3. Advanced mitigation mechanisms through a multi-agent collaboration framework

Multi-agent collaboration frameworks aim to use several LLMs from different perspectives to evaluate collaboratively, achieving fairer, more robust, and more comprehensive conclusions [24].

• Multi-Agent-as-Judge Evaluation: this framework involves three steps, and it frames different perspectives related to crucial benefits [24]. First, evaluating individually; second, discussing in groups; third, aggregating the results [24].

• Other methods: Involving CourtEval, which mocks a court [25]; Agent-as-a-Judge, which evaluates both results and thinking process during the task [26].

To provide a systematic comparison between different strategies, Table 3 concludes these mitigation mechanisms:

Table 3. Comparison of judicial hallucination mitigation techniques

Technique	Main Mechanism	Main Objective	Effectiveness Evaluation	Key limitations
Prompt engineering [21]	Improving the evaluation process by designing rubrics and structured outputs	Enhancing the stability and interpretability of judgments, and mitigating the prompted biases	Improving the consistency with humans, but with limited effects on complex tasks	Highly reliant on prompting quality, unable to solve the inherent flaws in models
Retrieval-augmented generation [22]	Judging based on an external, complete, credible database	Mitigating extrinsic hallucination; evaluating the consistency of answers in context	Effectively reduces fact fabrication, yet cannot eliminate hallucinations	Effectiveness limited by the retriever quality and the completeness of the knowledge base
Fine-tuned models (Prometheus) [19, 23]	Training LLM-as-a-Judge on a high-quality and annotated evaluation database	Creating expert models for specific evaluation tasks	Prometheus can achieve evaluation capabilities comparable to GPT-4	Risk of catastrophic forgetting, with limited generalization ability
Multi-Agent-as-Judge Evaluation [24]	Simulating a process of multi-perspective human expert panel review	Mitigating bias for a single model, enhancing comprehensiveness and robustness of judgements	High alignment of evaluation results with human experts across several dimensions	The debate process may fail to converge or be dominated by group bias

DOI: 10.54254/2755-2721/2026.TJ29492

5. Conclusions

This paper systematically analyzes and summarizes the unreliability of LLM-as-a-Judge. Judicial Hallucination is a complex and multi-dimensional phenomenon. The effectiveness of every solution is highly dependent on concrete evaluation scenarios, the complexity of tasks, and the requirement of reliability. This paper defines core metrics of hallucination, constructs unified classification methods, and reviews detection and mitigation mechanisms. To build a reliable LLM evaluation system, future research should achieve breakthroughs in the following areas: 1) Models resisting hallucination: Requiring new model frameworks or training methods that resist bias and hallucination. 2) Efficient Generation of evaluation benchmarks: This field needs research about how to effectively generate high-quality and challenging evaluation samples to reduce the cost for benchmarks like JudgeBench. 3) Efficient multi-agent collaboration framework: The current multi-agent collaboration framework is expensive and complex, and future research should focus on increasing the efficiency of these systems. 4) Theoretical research about recognition of LLMs: The study of reasons why LLMs can exhibit biases similar to humans might help researchers to understand these reasons and promote more creative solutions.

Ethical statement

This research promotes the reliability of LLM-as-a-Judge by systematically exploring failure modes in Judicial Hallucination to help defenders and developers understand potential threats and flaws of LLM-as-a-Judge, in order to develop effective countermeasures. It enhances system reliability and ensures safe evaluation from LLM-as-a-Judge, strictly prohibiting malicious uses.

References

[1] Zheng, L., Chiang, W. L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., ... & Stoica, I. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in neural information processing systems, 36, 46595-46623.
[2] Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., ... & Guo, J. (2024). A survey on llm-as-a-judge. arXiv preprint arXiv: 2411.15594.
[3] Li, H., Dong, Q., Chen, J., Su, H., Zhou, Y., Ai, Q., ... & Liu, Y. (2024). Llms-as-judges: a comprehensive survey on llm-based evaluation methods. arXiv preprint arXiv: 2412.05579.
[4] Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., ... & Shi, S. (2025). Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. Computational Linguistics, 1-46.
[5] Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., & Shou, M. Z. (2024). Hallucination of multimodal large language models: A survey. arXiv preprint arXiv: 2404.18930.
[6] Janiak, D., Binkowski, J., Sawczyn, A., Gabrys, B., Schwartz-Ziv, R., & Kajdanowicz, T. (2025). The Illusion of Progress: Re-evaluating Hallucination Detection in LLMs. arXiv preprint arXiv: 2508.08285.
[7] Chelli, M., Descamps, J., Lavoué, V., Trojani, C., Azar, M., Deckert, M., ... & Ruetsch-Chelli, C. (2024). Hallucination rates and reference accuracy of ChatGPT and bard for systematic reviews: comparative analysis. Journal of medical Internet research, 26(1), e53164.
[8] Honovich, O., Aharoni, R., Herzig, J., Taitelbaum, H., Kukliansy, D., Cohen, V., ... & Matias, Y. (2022). TRUE: Re-evaluating factual consistency evaluation. arXiv preprint arXiv: 2204.04991.
[9] Shi, L., Ma, C., Liang, W., Diao, X., Ma, W., & Vosoughi, S. (2024). Judging the judges: A systematic study of position bias in llm-as-a-judge. arXiv preprint arXiv: 2406.07791.
[10] Wataoka, K., Takahashi, T., & Ri, R. (2024). Self-preference bias in llm-as-a-judge. arXiv preprint arXiv: 2410.21819.
[11] Baysan, M. S., Uysal, S., İşlek, İ., Çığ Karaman, Ç., & Güngör, T. (2025). LLM-as-a-Judge: automated evaluation of search query parsing using large language models. Frontiers in Big Data, 8, 1611389.
[12] Pearson, K. (1895). VII. Note on regression and inheritance in the case of two parents. proceedings of the royal society of London, 58(347-352), 240-242.

DOI: 10.54254/2755-2721/2026.TJ29492

[13] Bedemariam, R., Perez, N., Bhaduri, S., Kapoor, S., Gil, A., Conjar, E., ... & Nayyar, N. (2025). Potential and perils of large language models as judges of unstructured textual data. arXiv preprint arXiv: 2501.08167.
[14] Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and psychological measurement, 20(1), 37-46.
[15] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., ... & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems, 43(2), 1-55.
[16] Saito, K., Wachi, A., Wataoka, K., & Akimoto, Y. (2023). Verbosity bias in preference labeling by large language models. arXiv preprint arXiv: 2310.10076.
[17] Chen, G. H., Chen, S., Liu, Z., Jiang, F., & Wang, B. (2024). Humans or llms as the judge? a study on judgement biases. arXiv preprint arXiv: 2402.10669.
[18] Tan, S., Zhuang, S., Montgomery, K., Tang, W. Y., Cuadron, A., Wang, C., ... & Stoica, I. (2024). Judgebench: A benchmark for evaluating llm-based judges. arXiv preprint arXiv: 2410.12784.
[19] Kim, S., Shin, J., Cho, Y., Jang, J., Longpre, S., Lee, H., ... & Seo, M. (2023, October). Prometheus: Inducing fine-grained evaluation capability in language models. In The Twelfth International Conference on Learning Representations.
[20] Bang, Y., Ji, Z., Schelten, A., Hartshorn, A., Fowler, T., Zhang, C., ... & Fung, P. (2025). Hallulens: Llm hallucination benchmark. arXiv preprint arXiv: 2504.17550.
[21] Yamauchi, Y., Yano, T., & Oyamada, M. (2025). An Empirical Study of LLM-as-a-Judge: How Design Choices Impact Evaluation Reliability. arXiv preprint arXiv: 2506.13639.
[22] Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E. (2025). Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. Journal of Empirical Legal Studies, 22(2), 216-242.
[23] Hu, Z., Wang, L., Lan, Y., Xu, W., Lim, E. P., Bing, L., ... & Lee, R. K. W. (2023). Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models. arXiv preprint arXiv: 2304.01933.
[24] Chen, J., Lu, Y., Wang, X., Zeng, H., Huang, J., Gesi, J., ... & Wang, D. (2025). Multi-agent-as-judge: Aligning llm-agent-based automated evaluation with multi-dimensional human evaluation. arXiv preprint arXiv: 2507.21028.
[25] Kumar, S., Nargund, A. A., & Sridhar, V. (2025, July). CourtEval: A courtroom-based multi-agent evaluation framework. In Findings of the Association for Computational Linguistics: ACL 2025 (pp. 25875-25887).
[26] Zhuge, M., Zhao, C., Ashley, D., Wang, W., Khizbullin, D., Xiong, Y., ... & Schmidhuber, J. (2024). Agent-as-a-judge: Evaluate agents with agents. arXiv preprint arXiv: 2410.10934.