

Overconfidence in LLM-as-a-Judge: Diagnosis and Confidence-Driven Solution

Zailong Tian^{1*}, Zhuoheng Han^{2*}, Yanzhe Chen⁴, Haozhe Xu², Xi Yang³,
Richeng Xuan^{3†}, Houfeng Wang^{2‡}, and Lizi Liao^{1§}

¹ School of Computing and Information Systems, Singapore Management University
zltian@alumni.pku.edu.cn, lzliao@smu.edu.sg

² State Key Laboratory for Multimedia Information Processing, Peking University
2501112032@stu.pku.edu.cn, wanghf@pku.edu.cn

³ Beijing Academy of Artificial Intelligence
rcxuan@baai.ac.cn

⁴ School of Computing, National University of Singapore

*Equal contribution. ^{†,‡,§}Corresponding authors.

Abstract. Large Language Models (LLMs) are increasingly used as automatic evaluators of generative AI outputs, a paradigm often referred to as "LLM-as-a-judge," where the confidence reported alongside each judgment is widely used as a practical signal for downstream decisions. However, whether such confidence is actually trustworthy remains largely unexamined. We first present a systematic diagnosis that reveals a pervasive **Overconfidence Phenomenon** in current LLM judges: across 9 leading models and three confidence elicitation methods, predicted confidence consistently overstates true accuracy on JudgeBench. Our further analysis reveals that confidence often remains informative as a relative ranking signal, although its discriminative strength varies across models. Motivated by this gap between calibration and informativeness, we propose **LLM-as-a-Fuser**, an ensemble framework that transforms LLMs into more reliable, risk-aware evaluators. Extensive experiments and analyses show that strong Fuser variants improve over full majority voting in both accuracy and calibration, offering a practical recipe for trustworthy LLM-as-a-Judge deployment.

Keywords: LLM-as-a-Judge · Confidence Calibration · Ensemble Evaluation.

1 Introduction

LLMs [3,1,12,8,9] have rapidly become the default tool for automatically evaluating generative AI outputs, a paradigm widely known as "LLM-as-a-judge" [17]. Compared to human annotation, LLM judges are scalable, low-cost, and capable of providing nuanced comparisons across a wide range of tasks, from instruction following and reasoning to code generation and open-ended writing [11,15]. Beyond producing a discrete preference, LLM judges can be prompted to report a

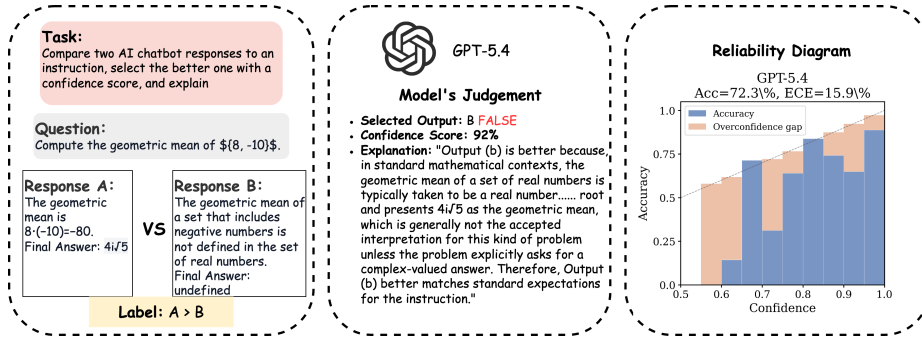


Fig. 1: Overconfidence phenomenon of GPT-5.4 as a Judge on JudgeBench, illustrated with a specific example (content omitted for demonstration purposes) and a Reliability Diagram. The Reliability Diagram plots the model’s confidence scores against actual accuracy, revealing calibration gaps where overconfidence occurs.

confidence score reflecting their certainty about each judgment, following broader work on eliciting verbalized uncertainty and self-evaluation from LLMs [5,4,13]. This confidence signal is widely used in downstream pipelines: high-confidence outputs are automatically accepted to reduce human review, low-confidence samples are flagged for inspection, and confidence values are routinely fed into data filtering, reward modeling, and synthetic data curation workflows [19]. In effect, LLM-as-a-judge pipelines increasingly rely not only on the judgments themselves, but also on the confidence behind them.

Despite this growing reliance, whether such confidence is actually *trustworthy* in the LLM-as-a-judge setting has received limited systematic attention. Prior work on LLM-as-a-judge has primarily focused on judgment quality, agreement with human preferences, and evaluation biases [17,14,20,11], whereas studies of confidence and calibration have largely considered general language-model settings rather than LLM-based evaluation specifically [7,18,13,4,10]. To address this gap, we conduct a large-scale diagnostic study across 9 leading open- and closed-source LLMs and three confidence elicitation methods—self-reported, multi-sample, and logit-based—on the JudgeBench benchmark. Our results reveal a pervasive **Overconfidence Phenomenon**: predicted confidence consistently and substantially overstates true accuracy. Even state-of-the-art judges concentrate their predictions in the 70–100% confidence range while they may achieve lower empirical accuracy, with the gap widening sharply for weaker models. This miscalibration directly undermines the reliability of any downstream decision that treats confidence as a probabilistic estimate.

A closer inspection, however, reveals that this is not the end of the story. In statistical learning, *calibration* (matching absolute probabilities) and *discrimination* (ranking correct versus incorrect predictions) are independent properties: a predictor can fail at the former while excelling at the latter. Our analysis shows

that LLM judges often exhibit this pattern, though the strength varies by model. Although their confidence scores are poorly calibrated in absolute terms, they frequently remain informative as a *relative* signal: high-confidence predictions are more accurate than low-confidence ones, with AUC values ranging from 0.642 to 0.877 across the models we examine, and the highest-confidence half of predictions outperforming the lowest-confidence half by a substantial margin. This dichotomy suggests that confidence in LLM judges can be used as a relative ranking signal.

Building on this insight, we propose **LLM-as-a-Fuser**⁵, an ensemble framework that uses confidence as a relative evidence signal for judging which model outputs deserve more trust. Instead of taking any single model’s confidence at face value or aggregating decisions through uniform voting, LLM-as-a-Fuser employs a dedicated *fuser* LLM that ingests the judgments, critiques, and confidence scores from an ensemble of judges, and synthesizes them into a final evidence-aware decision. By grounding the final judgment in cross-model evidence rather than any single model’s miscalibrated confidence, LLM-as-a-Fuser converts the relative informativeness of confidence into substantial, measurable reliability gains. Our contributions are threefold:

1. **Systematic diagnosis.** We identify and characterize the Overconfidence Phenomenon in LLM-as-a-judge across 9 models and three confidence elicitation methods.
2. **Calibration–informativeness dichotomy.** We uncover a fundamental distinction between absolute miscalibration and relative informativeness in LLM judge confidence, showing that confidence remains a useful ranking signal despite poor calibration.
3. **LLM-as-a-Fuser.** We propose an ensemble framework that exploits this dichotomy and improves over full majority voting in both accuracy and calibration on JudgeBench.

2 Related Work

2.1 LLM-as-a-Judge

LLMs are increasingly used as automated evaluators for text quality. MT-Bench showed GPT-4 aligns with human judgments in pairwise comparisons, but proprietary APIs limit reproducibility. PandaLM [15] introduced a 7B-parameter local evaluator with 94% agreement with ChatGPT, supporting offline use. JudgeLM [20] fine-tunes LLMs at different scales as efficient and scalable judges, and studies several biases inherent to LLM-based evaluation. However, alignment between model confidence and accuracy is often ignored, causing inconsistent judgments. Meta’s self-rewarding models [16] generate and evaluate outputs iteratively, but calibration needs further study.

⁵ Dataset: <https://huggingface.co/datasets/Han03430/CalibrationBench>; code: <https://github.com/hzh030430/CalibrationBench>.

As LLM-based judges gain traction for evaluating and enhancing LLMs, various benchmarks have emerged to gauge their effectiveness. Prior works like LLMEval [6], MTBench, and FairEval [14] primarily assess how well LLM-based judges align with subjective human preferences, often emphasizing stylistic differences over factual and logical accuracy. JudgeBench offers a novel benchmark specifically designed to test LLM-based judges’ reasoning capabilities. It features 350 carefully curated challenging response pairs across knowledge, reasoning, math, and coding domains, each containing one objectively correct and one subtly incorrect response, prioritizing factual and logical correctness over subjective or stylistic factors.

2.2 Calibration in LLMs

Accurate calibration, which aligns a model’s confidence with its empirical accuracy, is crucial for reliable LLM applications. Classical post-hoc methods such as temperature scaling [2] recalibrate confidence with a single scalar, while recent LLM-specific methods such as Thermometer [10] train auxiliary models to improve uncertainty estimates across tasks. However, most existing calibration methods focus on single-model predictions, whereas LLM-as-a-judge pipelines often involve multiple judges and downstream decisions such as filtering, aggregation, and data curation.

3 Overconfidence in LLM-as-a-Judge

3.1 How to Measure Confidence in LLMs?

We employed three methods for calculating confidence: Self-Confidence (SC), Multiple-Prompting (MP) confidence, and Logprob (LogP) confidence.

SC setting: We prompt the model to output both the result and its confidence. Model’s temperature is set to 0 to ensure the reproducibility of the setting.

MP setting: We sample the same query 10 times at temperature 0.7. The final response is determined by majority voting, and confidence is defined as the vote share of the selected response.

LogP Setting. In this setting, confidence scores are derived from softmax-normalized logits for the final output tokens (e.g., ‘A’ or ‘B’). For a binary choice task with options A and B , and corresponding logits l_A and l_B , we first compute the softmax probabilities:

$$p(A) = \frac{e^{l_A}}{e^{l_A} + e^{l_B}}, \quad p(B) = \frac{e^{l_B}}{e^{l_A} + e^{l_B}}.$$

Confidence is then defined as the maximum probability:

$$\text{Confidence}_{\text{logp}} = \max(p(A), p(B)).$$

Temperature is set to 0 to ensure deterministic outputs.

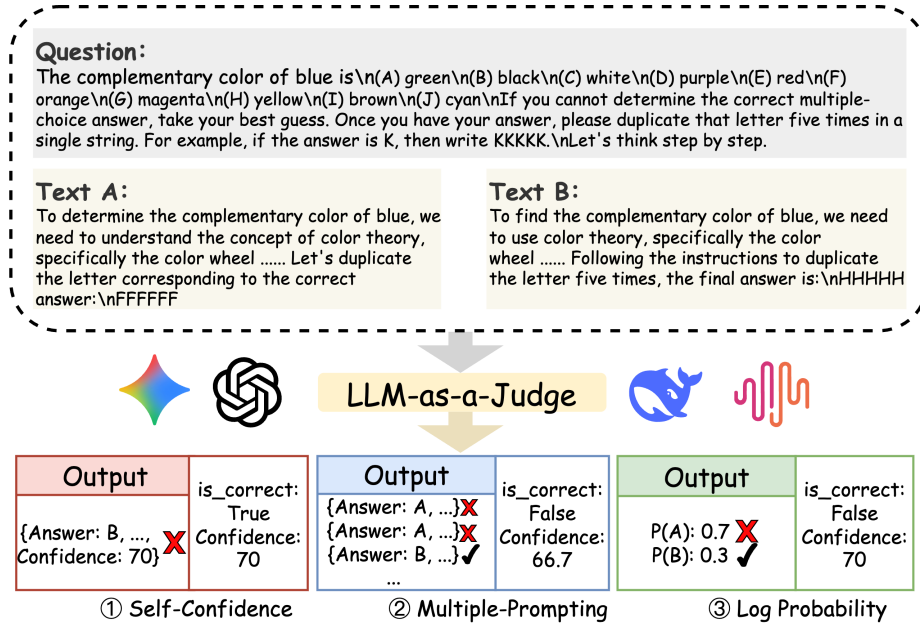


Fig. 2: Visualization of the three confidence calculation settings: Self-Confidence (SC), Multiple-Prompting (MP), and Logprob (Logp), using data with ID 122 from JudgeBench as an example.

3.2 Overall Calibration Results

We systematically evaluate 9 cutting-edge models on the JudgeBench benchmark under the Self-Confidence (SC) and Multiple-Prompting (MP) settings, with complete results presented in Table 1. These include open-source models (DeepSeek-R1, DeepSeek-v4-Flash, DeepSeek-v4-Pro, MiniMax-M2.7, Llama-3.3-70B) and proprietary models (Gemini-3.1-Pro, Gemini-3-Flash, GPT-5.4, GPT-5.4-mini). For models whose APIs expose usable token-level probabilities for the binary output options, we additionally evaluate the LogP setting. Specifically, GPT-5.4, GPT-5.4-mini, DeepSeek-R1, and Llama-3.3-70B achieve accuracies of 71.71%, 54.57%, 78.29%, and 43.43%, with ECEs of 19.62%, 12.97%, 10.82%, and 55.05%, respectively. These results show that substantial miscalibration persists under LogP confidence, particularly for Llama-3.3-70B.

3.3 Empirical Evidence of Overconfidence

Table 1 reveals a systematic overconfidence phenomenon across all evaluated models. The most alarming cases occur among weak models: DeepSeek-v4-Flash achieves only 49.43% accuracy under SC yet reports an average confidence of 86.7%, yielding an ECE of 37.23. Similarly, MiniMax-M2.7 attains 55.71% accuracy with 81.8% average confidence (ECE 26.08), and DeepSeek-v4-Pro reaches

Table 1: Performance of nine models under Self-Confidence (SC) and Multiple-Prompting (MP) settings on JudgeBench. Arrows indicate optimization direction: $\uparrow$ higher is better for accuracy, $\downarrow$ lower is better for ECE. Best result per metric is bolded.

Model	Acc. (SC) ($\uparrow$)	ECE (SC) ($\downarrow$)	Acc. (MP) ($\uparrow$)	ECE (MP) ($\downarrow$)	Avg. Conf.
Proprietary Models					
Gemini-3.1-Pro	90.57	6.29	91.14	6.91	96.9
Gemini-3-Flash	75.14	15.32	81.71	10.94	90.5
GPT-5.4	72.29	15.92	73.43	18.71	88.2
GPT-5.4-mini	55.43	9.26	54.57	14.11	64.4
Open Source Models					
DeepSeek-R1	76.86	9.68	85.43	6.40	87.6
DeepSeek-v4-Pro	64.29	23.00	64.29	18.03	87.3
MiniMax-M2.7	55.71	26.08	84.29	6.56	81.8
DeepSeek-v4-Flash	49.43	37.23	50.86	32.29	86.7
Llama-3.3-70B	42.00	45.62	42.86	54.31	88.7

64.29% accuracy with 87.3% confidence (ECE 23.00). These models operate near random-guess levels while expressing strong certainty, making them particularly dangerous for automated judgment pipelines that rely on confidence thresholds for quality control.

Macro-level metrics, however, only capture average behavior. A more insidious problem lies at the sample level: even when models are wrong, they often express high confidence. Among incorrect predictions under the SC setting, Llama-3.3-70B assigns confidence $\geq 80\%$ to 90.1% of its errors and $\geq 90\%$ to 62.1%; Gemini-3.1-Pro does so for 78.8% and 54.5%, respectively. DeepSeek-v4-Flash, despite only 49.43% SC accuracy, still assigns confidence $\geq 80\%$ to 76.8% of its errors and $\geq 90\%$ to 25.4%. This pervasive high-confidence error pattern means that confidence cannot serve as a reliable hard threshold for filtering decisions: a model declaring “I am 90% certain” still has a substantial probability of being wrong.

4 Confidence Is Miscalibrated but Still Informative

The preceding section diagnosed a pervasive Overconfidence Phenomenon: predicted confidence systematically overstates actual accuracy, with high-confidence errors pervading even the strongest models. At first glance, this appears to render confidence signals unreliable for practical decision-making. However, a closer inspection reveals a more nuanced picture. In statistical learning, *calibration* and *discrimination* are two independent concepts: a predictor can be poorly calibrated yet possess strong discriminative power—that is, it cannot estimate

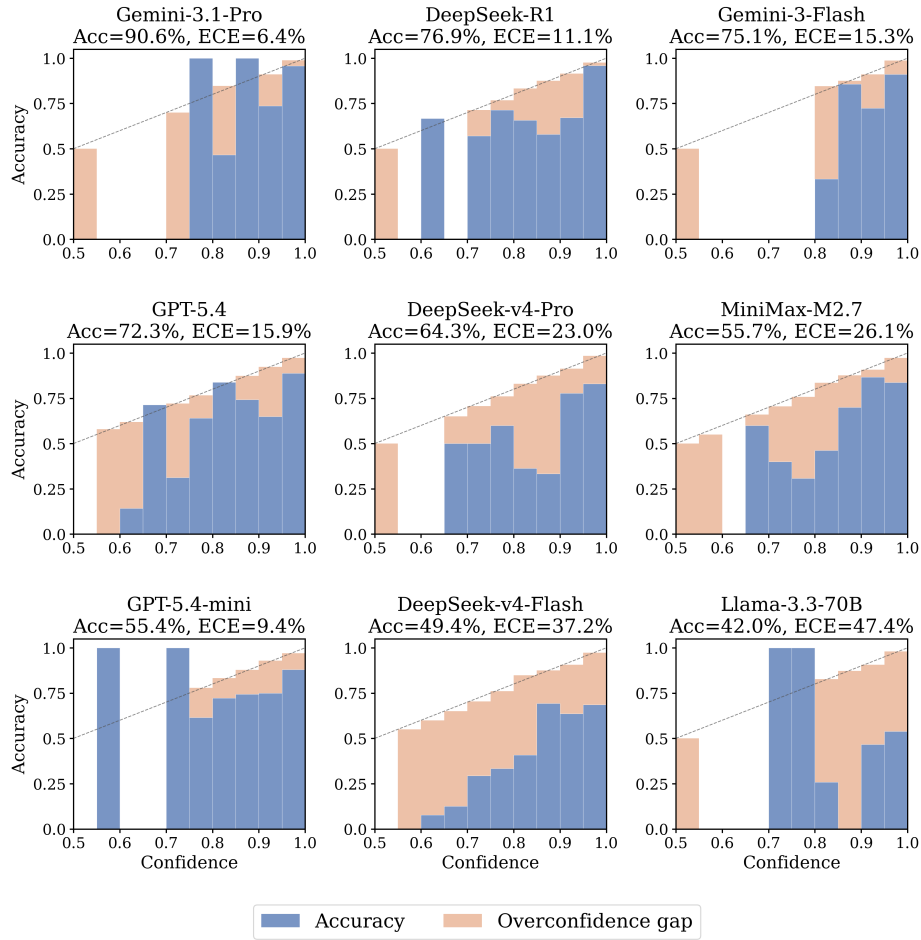


Fig. 3: Reliability diagrams for all nine models under the Self-Confidence (SC) setting on JudgeBench. Blue bars show empirical accuracy within each confidence bin; orange regions indicate the overconfidence gap. The dashed diagonal represents perfect calibration.

absolute probabilities accurately, but can still reliably distinguish high-quality from low-quality samples. This section characterizes this critical distinction and demonstrates why confidence should not be discarded but repurposed as a relative ranking signal.

4.1 Global Discriminative Power

To quantify the discriminative power of confidence, we compute the Area Under the ROC Curve (AUC) for each model’s confidence scores against correctness labels. Table 2 reports the AUC and correlation statistics for all nine models under

the Single-Confidence (SC) setting. The results reveal substantial heterogeneity: the strongest models—led by MiniMax-M2.7 (AUC 0.877)—cluster above 0.85, while the weakest, Llama-3.3-70B (0.642), falls below 0.65. Pearson correlations follow the same pattern, spanning from 0.251 to 0.744. In other words, confidence often retains useful discriminative capacity despite its absolute calibration flaws, but this property is far from universal across models.

Table 2: Confidence-accuracy discriminative power under the SC setting. AUC = Area Under ROC Curve; Pearson r = correlation between confidence and correctness.

Model	AUC	Pearson r	Spearman r
MiniMax-M2.7	0.877	0.679	0.654
DeepSeek-v4-Pro	0.875	0.684	0.642
Gemini-3-Flash	0.873	0.744	0.595
DeepSeek-R1	0.865	0.562	0.539
GPT-5.4-mini	0.858	0.704	0.624
Gemini-3.1-Pro	0.850	0.558	0.406
GPT-5.4	0.753	0.405	0.393
DeepSeek-v4-Flash	0.675	0.316	0.316
Llama-3.3-70B	0.642	0.251	0.246

4.2 Bin-Level Accuracy Analysis

Global metrics such as AUC validate overall ranking quality, but they do not reveal how confidence translates into accuracy at the granular level. To examine this, we partition predictions into confidence bins and compute empirical accuracy within each bin. Figure 4 presents this analysis for all nine models. The broad pattern is upward, though not perfectly monotonic in low-count bins. For instance, Gemini-3.1-Pro climbs from roughly 53% accuracy in the $[0.8, 0.9]$ bin to over 94% in the $[0.9, 1.0]$ bin, while even weaker models such as DeepSeek-v4-Flash more than double their accuracy across the top three bins. These trends reinforce that confidence is more useful as a relative ranking signal than as an absolute probability estimate.

5 LLM-as-a-Fuser

Section 4 left us with a mixed signal: confidence is miscalibrated, yet it still carries some useful information about which judgments to trust. A natural reaction would be to design a calibration procedure or a principled confidence-weighted aggregation rule around this observation. We try the simplest thing instead: hand the judges’ decisions, confidences, and rationales to another LLM and ask

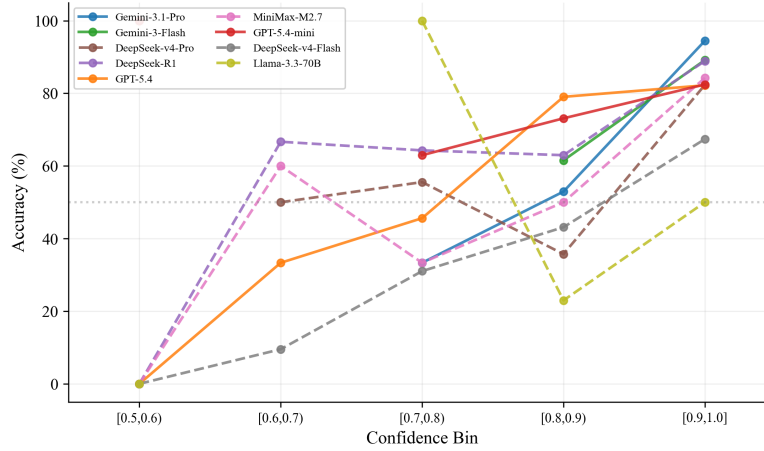


Fig. 4: Binned accuracy analysis for all nine models under the SC setting. Predictions are partitioned into five confidence bins, and empirical accuracy is computed within each bin. Proprietary models are shown with solid lines and open-source models with dashed lines.

it to produce a verdict. We call this setup **LLM-as-a-Fuser**. The framework is deliberately minimal—it is essentially a prompt—and the point of this section is not that the design is novel, but that this light-touch use of confidence, together with the rationales each judge already produces, is enough to improve substantially over majority voting on both accuracy and calibration.

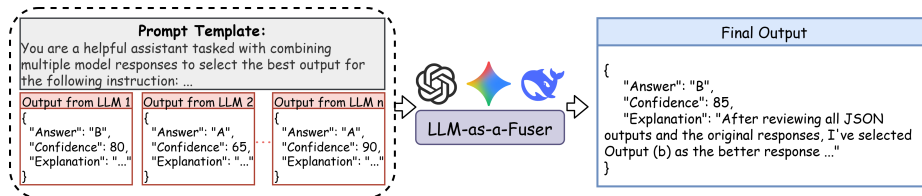


Fig. 5: Illustration of the LLM-as-a-Fuser framework, where the fuser model aggregates decisions and their corresponding critiques from an ensemble of models.

5.1 Experimental Results

Table 4 reports Fuser performance on JudgeBench. The best variant (Gemini-3.1-Pro) reaches 95.67% accuracy at 2.25% ECE, and every Fuser variant stays below 7.53% ECE—well within the range one would want for downstream use, and below the SC numbers in Table 1 (which span 6.29% to 45.62%). Majority

Table 3: Comparison with calibration and aggregation baselines on JudgeBench. Best results are bolded.

Method	Acc. ($\uparrow$)	ECE ($\downarrow$)
Temperature Scaling	64.64	9.25
Confidence-Weighted Voting	89.71	5.34
Logistic Stacking	93.43	1.13
LLM-as-a-Fuser	95.67	2.25

voting over the same nine judges reaches 88.29%, and over the top three judges 92.86%, so the gains are not just attributable to access to better inputs.

The biggest improvements show up when a weak judge plays the role of fuser. DeepSeek-v4-Flash goes from 49.43% as an individual judge to 94.72% as a Fuser, and several other weak-to-mid judges show double-digit gains. Even Gemini-3.1-Pro, already the strongest individual judge at 90.57% SC, improves to 95.67% when fusing.

We further compare LLM-as-a-Fuser with standard calibration and aggregation baselines. As shown in Table 3, temperature scaling improves confidence calibration without changing the underlying decisions, while confidence-weighted voting provides a simple confidence-aware aggregation baseline. Logistic stacking achieves the lowest ECE, whereas the best Fuser obtains the highest observed accuracy without requiring a labeled task-specific meta-classifier.

5.2 Ablation Studies

To test whether confidence is useful for selecting judge evidence, we construct confidence-filtered inputs for the Fuser. **Base** uses all nine judges. For each test instance, **Top3** keeps only the three judges with the highest confidence on that instance, while **Bottom3** keeps only the three with the lowest confidence. We also include **wo3**, which removes the three strongest individual judges, as a control to check whether the Fuser is merely exploiting the outputs of a few high-performing models.

The contrast between Top3 and Bottom3 is the central result in Table 4. Top3 slightly improves the average Fuser accuracy over Base (93.18% vs. 92.64%), and Gemini-3.1-Pro reaches its best result under this setting (96.84%). In other words, keeping only the most confident judges is not worse than feeding all judges, despite discarding two thirds of the inputs. Bottom3 shows the opposite pattern: the average accuracy falls to 85.73%, and the weakest cases lose more than ten points. This gap indicates that judge confidence, although poorly calibrated as an absolute probability, is useful for identifying which model judgments are worth passing to the Fuser.

By removing the three strongest individual judges, the wo3 setting tests whether the Fuser’s gains come simply from copying or exploiting a few high-performing inputs. The average accuracy remains high at 91.59%, only modestly below Base, suggesting that the Fuser still extracts useful evidence from

Table 4: Fuser vs. Majority Voting across ablation configurations. **Bold** indicates the best in each column.

Method	Base	wo3	Top3	Bottom3
Majority Voting	88.29	81.43	92.86	71.71
Best-Model	92.29	87.71	92.29	74.29
Gemini-3.1-Pro	95.67	96.00	96.84	96.05
DeepSeek-v4-Flash	94.72	93.84	92.75	89.57
DeepSeek-v4-Pro	94.69	94.61	94.07	90.03
Gemini-3-Flash	94.39	92.88	94.95	94.64
DeepSeek-R1	93.87	92.82	94.47	89.74
MiniMax-M2.7	92.49	91.27	91.94	83.33
GPT-5.4	91.07	91.91	92.82	80.64
GPT-5.4-mini	88.44	87.46	90.86	73.23
Llama-3.3-70B	88.41	83.48	89.94	74.38
Fuser Avg.	92.64	91.59	93.18	85.73

the remaining mid- and low-tier judges. Together, these ablations support the interpretation from Section 4: confidence should not be treated as a calibrated probability, but it is informative enough to guide judge selection.

6 Conclusion and Limitations

This paper studies the trustworthiness of confidence in LLM-as-a-judge systems. We find that LLM judges are consistently overconfident on JudgeBench: their reported confidence often exceeds their actual accuracy, especially when they make wrong judgments. Nevertheless, confidence still carries useful relative information, helping distinguish more reliable from less reliable predictions. Based on this observation, we propose LLM-as-a-Fuser, which aggregates judge decisions, rationales, and confidence scores. Experiments show that Fuser variants improve over majority voting in both accuracy and calibration, suggesting that confidence should be used as relative evidence rather than literal probability.

Our study is limited to pairwise evaluation on JudgeBench, where each example has an objective ground-truth preference; results may differ for subjective, multi-turn, or domain-specific evaluation. In addition, LogP confidence is only available for models with API support, and LLM-as-a-Fuser introduces extra inference cost. Future work should test broader benchmarks and develop cheaper, more systematic judge-selection strategies.

Acknowledgments. This research is supported by the National Research Foundation, Singapore under its National Large Language Models Funding Initiative (AISG Award No: AISG-NMLP-2024-002). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not reflect the views of the National Research Foundation or the Ministry of Education, Singapore. This work was also supported by the National Natural Science Foundation of China (62576010).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al.: The llama 3 herd of models. arXiv e-prints pp. arXiv-2407 (2024)
2. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International conference on machine learning. pp. 1321–1330. PMLR (2017)
3. Guo, D., Yang, D., Zhang, H., et al.: Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature* **645**, 633–638 (2025). <https://doi.org/10.1038/s41586-025-09422-z>
4. Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., Ganguli, D., Hernandez, D., Jacobson, J., Kernion, J., Kravec, S., Lovitt, L., Ndousse, K., Olsson, C., Ringer, S., Amodei, D., Brown, T., Clark, J., Joseph, N., Mann, B., McCandlish, S., Olah, C., Kaplan, J.: Language models (mostly) know what they know. arXiv preprint arXiv:2207.05221 (2022). <https://doi.org/10.48550/arXiv.2207.05221>, <https://arxiv.org/abs/2207.05221>
5. Lin, S., Hilton, J., Evans, O.: Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research* (2022), <https://openreview.net/forum?id=8s8K2UZGTZ>
6. Lin, Y.T., Chen, Y.N.: Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In: Proceedings of the 5th Workshop on NLP for Conversational AI. pp. 47–58. Association for Computational Linguistics, Toronto, Canada (2023). <https://doi.org/10.18653/v1/2023.nlp4convai-1.5>, <https://aclanthology.org/2023.nlp4convai-1.5/>
7. Mielke, S.J., Szlam, A., Dinan, E., Boureau, Y.L.: Reducing conversational agents’ overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics* **10**, 857–872 (2022). https://doi.org/10.1162/tacl_a_00494
8. MiniMax: Minimax m2.7: Early echoes of self-evolution. <https://www.minimax.io/news/minimax-m27-en> (Mar 2026), accessed: 2026-05-24
9. OpenAI: Introducing gpt-5.4. <https://openai.com/index/introducing-gpt-5-4/> (Mar 2026), accessed: 2026-05-24
10. Shen, M., Das, S., Greenewald, K., Sattigeri, P., Wornell, G.W., Ghosh, S.: Thermometer: Towards universal calibration for large language models. In: Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 235, pp. 44687–44711. PMLR (2024), <https://proceedings.mlr.press/v235/shen24c.html>
11. Tan, S., Zhuang, S., Montgomery, K., Tang, W.Y., Cuadron, A., Wang, C., Popa, R.A., Stoica, I.: Judgebench: A benchmark for evaluating llm-based judges. In: The Thirteenth International Conference on Learning Representations (2025), <https://openreview.net/forum?id=G0dksFayVq>
12. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)

13. Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., Finn, C., Manning, C.: Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 5433–5442. Association for Computational Linguistics, Singapore (Dec 2023). <https://doi.org/10.18653/v1/2023.emnlp-main.330>, <https://aclanthology.org/2023.emnlp-main.330/>
14. Wang, P., et al.: Large language models are not fair evaluators. In: Proceedings of ACL. pp. 9440–9450 (2024). <https://doi.org/10.18653/v1/2024.acl-long.511>
15. Wang, Y., Yu, Z., Zeng, Z., Yang, L., Wang, C., Chen, H., Jiang, C., Xie, R., Wang, J., Xie, X., et al.: Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. In: The Twelfth International Conference on Learning Representations (2024), <https://openreview.net/forum?id=5Mn2BLV7SB>
16. Wu, T., Yuan, W., Golovneva, O., Xu, J., Tian, Y., Jiao, J., Weston, J.E., Sukhbaatar, S.: Meta-rewarding language models: Self-improving alignment with LLM-as-a-meta-judge. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 11537–11554. Association for Computational Linguistics, Suzhou, China (Nov 2025). <https://doi.org/10.18653/v1/2025.emnlp-main.583>, <https://aclanthology.org/2025.emnlp-main.583/>
17. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al.: Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems* **36**, 46595–46623 (2023)
18. Zhou, K., Jurafsky, D., Hashimoto, T.: Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 5506–5524. Association for Computational Linguistics, Singapore (2023). <https://doi.org/10.18653/v1/2023.emnlp-main.335>, <https://aclanthology.org/2023.emnlp-main.335/>
19. Zhou, X., Kim, K., Zhang, T., Weyssow, M., Gomes, L.F., Yang, G., Liu, K., Xia, X., Lo, D.: An llm-as-judge metric for bridging the gap with human evaluation in se tasks (2025), <https://arxiv.org/abs/2505.20854>
20. Zhu, L., Wang, X., Wang, X.: JudgeLM: Fine-tuned large language models are scalable judges. In: The Thirteenth International Conference on Learning Representations (2025), <https://openreview.net/forum?id=xsELpEPn4A>