Global-Local Confidence Fusion for Hallucination Detection in Mathematical Reasoning Task

Bo Zhang[1,5†], Cong Gao[2], Linkang Yang[3], Bingxu Han[4], Minghao Hu[5], Zhunchen Luo[5], Guotong Geng[5], Xiaoying Bai[5], Jun Zhang[5,6*], Wen Yao[6*], Zhong Wang[1*]

[1]PLA Rocket Force University of Engineering
[2]College of Cryptology and Cyber Science, Nankai University
[3]School of Electronics and Information, Xi'an Jiaotong University
[4]School of Mathematics, Shandong University
[5]Center of Information Research, PLA Academy of Military Science
[6]Defense Innovation Institute, PLA Academy of Military Science
mcgrady150318@163.com, dsp863wang@163.com, wendy0782@126.com

Abstract

Large Reasoning Models (LRMs) achieve promising results on complex reasoning tasks but remain susceptible to hallucinations. Existing hallucination detection methods based on Large Language Models (LLMs) often focus solely on final answers, overlooking inconsistencies between the answer and reasoning process. This limitation reduces their ability to detect hallucinations during inference. Moreover, training-free approaches lack mechanisms for confidence estimation, resulting in an unquantified detection output. In contrast, training-based methods can provide fine-grained assessments but often neglect the self-correction capability of LRMs, where earlier errors may be corrected in subsequent steps, leading to inaccurate hallucination detection. To address these challenges, we propose **ConfFuse**, a unified framework that fuses global and local confidence scores for hallucination detection. A Global Hallucination Detection Model (GHDM) is trained using Direct Preference Optimization (DPO) to assess hallucinations at the level of entire reasoning chains, yielding global confidence estimates. Simultaneously, a Process Reward Model (PRM) estimates step-wise confidence scores to capture local logical flaws. A weighted fusion strategy combines the global confidence score with the minimum local score to jointly reflect overall reasoning consistency and local soundness. Experimental evaluations demonstrate that ConfFuse surpasses Qwen3-1.7B and Qwen3-8B by up to 11.86% and 5.46% in F1 score on in-distribution datasets, and achieves average improvements of 4.65% and 2.80% on out-of-distribution datasets. These results verify the effectiveness and generalizability of the proposed framework.

1 Introduction

Large Language Models (LLMs) have achieved remarkable success on complex reasoning tasks, such as mathematical problem-solving (Nguyen et al. 2024; Xu et al. 2025; Kang et al. 2025) and scientific analysis (Lam et al. 2023; Xia et al. 2025b). A notable subclass of LLMs, known as Large Reasoning Models (LRMs)—including DeepSeek-R1 (Guo et al. 2025) and OpenAI's O-series (OpenAI et al. 2025)—typically generates complete reasoning traces before producing final answers, showcasing advanced multi-step reasoning capabilities. However, despite these advances, LRMs remain susceptible to hallucinations (Yao et al. 2025; Chowdhury et al. 2025), often generating logically inconsistent content that undermines the reliability of their reasoning processes.

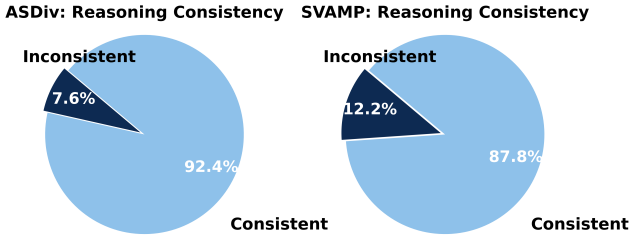

Figure 1: We analyze the consistency between reasoning processes and final answers on mathematical benchmarks (ASDiv and SVAMP), using DeepSeek-R1-Distill-Llama-8B as the generation model and GPT-4o for hallucination annotation. **Our findings reveal that LRMs exhibit inconsistencies between the validity of the reasoning process and the correctness of the final answer.** Specifically, such inconsistencies arise in two forms: cases where hallucination-free reasoning results in incorrect answers, and cases where hallucinated reasoning nonetheless produces correct answers.

Effective hallucination detection in LRMs necessitates rigorous evaluation of whether a model's reasoning process aligns with verifiable ground truth. Mathematical reasoning is particularly well suited for this purpose, owing to its stringent logical dependencies, and deterministic reference solutions that facilitate objective validation.

Hallucination detection methods in mathematical reasoning generally fall into two categories. The first category

*Corresponding authors.
†Work performed while an intern at the Center of Information Research, PLA Academy of Military Science.

comprises training-free methods that employ LLMs to directly detect hallucinations from reasoning traces. Representative approaches include SelfCheckGPT (Manakul, Liusie, and Gales 2023), LM vs LM (Cohen et al. 2023), and CoVe (Dhuliawala et al. 2023). Relying solely on the intrinsic reasoning capabilities of LLMs, these approaches require no supervised training and are thus highly flexible and generalizable. The second category involves training-based methods that use a Process Reward Model (PRM) to detect hallucination in step-level reasoning. Representative methods include FG-PRM (Li, Luo, and Du 2024), PathFinder-PRM-7B (Pala et al. 2025), and ReasonEval-7B (Xia et al. 2025a).

However, most existing hallucination detection methods primarily focus on evaluating the final answers, often overlooking the logical structure and validity of the explicit reasoning processes generated by LRMs. As shown in Figure 1, our empirical analysis reveals that the final answers generated by LRMs exhibit inconsistencies with their corresponding reasoning processes. **Therefore, answer-based detection methods cannot adequately identify hallucinations occurring within the inference process.** Furthermore, training-free methods typically lack explicit confidence estimation mechanisms and therefore struggle to produce quantified judgments for hallucination detection. In contrast, while training-based methods offer fine-grained confidence estimation, most existing PRMs still overlook the self-correction behaviors often observed in LRMs, where errors made in early steps may be corrected later in the reasoning process (Xu et al. 2025).

To address these limitations, we propose ConfFuse—a hallucination detection framework that fuses global and local confidence scores. Specifically, we train a Global Hallucination Detection Model (GHDM) to assess the entire reasoning process and generate a global confidence score reflecting overall reliability. In parallel, we introduce a Process Reward Model (PRM) that identifies local-level reasoning errors and produces local confidence scores. By fusing these two complementary scores via a weighted mechanism, ConfFuse enhances hallucination detection performance. Experimental results demonstrate that ConfFuse consistently outperforms strong baselines on both in-distribution and out-of-distribution datasets.

In summary, the main contribution of this work is as follows.

- We propose GHDM, the first hallucination detection model specifically designed to assess the entire reasoning process of LRMs. GHDM jointly produces step-level hallucination explanations and global confidence scores that quantify overall reasoning reliability.
- We present ConfFuse, a unified detection framework that integrates GHDM's global confidence with PRM's local-level confidence scores, enabling robust and fine-grained hallucination detection.
- We validate ConfFuse on multiple in-distribution and out-of-distribution mathematical reasoning benchmarks, where it consistently outperforms strong baselines.

2 Related Work

Hallucination Detection. Hallucination detection plays a critical role in ensuring the reliability of LLM-generated content. Recent studies propose detection methods across varying levels of granularity, including token-level (Liu et al. 2022; Zhang et al. 2023), entity-level (Yeh et al. 2025), claim-level (Hu et al. 2024), and response-level (Miao, Teh, and Rainforth 2023). At the token level, HaDes (Liu et al. 2022) introduces a reference-free annotated dataset, whereas HaMI (Niu, Haddadi, and Pang 2025) formulates the task as a multiple-instance learning problem that jointly optimizes token selection and hallucination identification. Entity-level approaches verify the factual correctness of mentioned entities, with HalluEntity (Yeh et al. 2025) evaluating uncertainty-based methods. Claim-level detection focuses on individual statements: FActScore (Wang et al. 2023) decomposes text into atomic facts; Pelican (Sahu, Sikka, and Divakaran 2024) and RefChecker (Hu et al. 2024) further refine the analysis at the claim level. Response-level methods assess overall hallucination risk. SelfCheckGPT and CoVe (Dhuliawala et al. 2023) adopt self-verification, MetaQA (Yang et al. 2025) introduces prompt variation, and FG-PRM (Li, Luo, and Du 2024) focuses on step-level hallucinations in mathematical reasoning.

LLM-as-a-Judge. As LLMs gain stronger evaluation capabilities, the "LLM-as-a-Judge" paradigm becomes a widely adopted approach for hallucination detection (Li et al. 2025; Zhang et al. 2025). It enables automatic assessment of model outputs without human annotation and supports high scalability. Representative works include MT-Bench (Zheng et al. 2023), G-Eval (Liu et al. 2023), and LLM-as-a-Judge (Gu et al. 2025), which show strong consistency and reliability in evaluating dialogue, factual correctness, and safety. This paradigm is also widely applied to the detection of hallucinations. SelfCheckGPT (Manakul, Liusie, and Gales 2023) and ChainPoll (Friel and Sanyal 2023) prompt models to verify the consistency of their output. Reward-model-based variants further extend the LLM-as-a-Judge paradigm. Process Reward Models, such as those introduced in (Lightman et al. 2023), are primarily used to evaluate the soundness of step-wise reasoning, while Outcome Reward Models (Cobbe et al. 2021) focus on the correctness of the final answer. Notably, models like PathFinder-PRM-7B (Pala et al. 2025) and Qwen2.5-Math-PRM-7B (Yang et al. 2024) are specifically designed to assess multi-step reasoning in mathematical tasks.

In this work, we propose a unified LLM-as-a-Judge framework that fuses local-level and global-level hallucination detection, enabling structured analysis of hallucinations in complex multi-step reasoning processes.

3 Methodology

3.1 Global Hallucination Detection Model Training

Reasoning Process Collection. To enhance the global hallucination detection capability of the base model on the reasoning processes of LRMs, we adopt the DPO (Rafailov et al. 2024) training paradigm. The construction of the

Direct Preference Optimization (DPO) training dataset is illustrated in Figure 2. We use GSM8K (Cobbe et al. 2021) and MathQA (Amini et al. 2019) as the original datasets. For each math question, independent reasoning is performed using seven widely used LRMs. The selected models include Marco-o1 (Zhao et al. 2024), five models of the DeepSeek-R1-Distill series: DeepSeek-1.5B/7B/8B/14B/32B (DeepSeek-AI 2025) and Skywork-OR1-7B-Preview (He et al. 2025).

During data construction, we introduce a strong-weak model agreement strategy (Xu et al. 2024), which leverages the performance gap between a strong model and a weak model in hallucination detection to construct preference pairs. Specifically, we use Qwen3-32B as the strong model and Qwen3-0.6B as the weak model. Based on their disagreement in hallucination judgments, we construct a DPO training dataset comprising correct hallucination assessments (chosen) from the strong model and incorrect assessments (rejected) from the weak model, thereby providing supervision signals for preference-based fine-tuning. Details of the prompt are provided in Appendix A.

DPO Chosen. To construct the DPO chosen samples, we utilize hallucination detection outputs generated by a strong model M_s as supervision signals for preference optimization. For each question q_i with its reasoning process r_i and reference solution a_i, the strong model M_s evaluates hallucination by producing an explanatory decision $e_{s,i}$ and a confidence score $c_{s,i} \in [0, 1]$, i.e., $M_s(q_i, r_i, a_i) \mapsto (e_{s,i}, c_{s,i})$. Here, $e_{s,i}$ provides a textual explanation indicating whether hallucinations are present in the reasoning process, while $c_{s,i}$ serves as a global hallucination confidence score, reflecting the model's confidence in the reliability of the entire reasoning process. Lower confidence indicates reduced reliability (i.e., a higher likelihood of hallucination). We adopt a threshold-based hallucination evaluation method (Tao et al. 2024; Chen and Mueller 2024). A sample is retained as a DPO chosen instance if and only if the explanation aligns with the confidence score—specifically: (i) if $e_{s,i}$ indicates `non-hallucinated` and $c_{s,i} \geq 0.5$, or (ii) if $e_{s,i}$ indicates `hallucinated` and $c_{s,i} < 0.5$. This consistency criterion ensures that the selected samples reflect agreement between the model's interpretation and its self-assessed confidence, thereby supporting the construction of high-quality preference data for DPO training.

DPO Rejected. To mitigate the impact of linguistic style discrepancies in DPO learning, we utilize smaller Qwen3 models to generate reject samples, ensuring consistency in language expression and formatting with the DPO chosen samples. A weaker model M_w is employed to perform hallucination detection on the reasoning input (q_i, r_i), without access to the reference solution a_i. Specifically, for each sample, M_w outputs an explanatory judgment $e_{w,i}$ and a confidence score $c_{w,i}$, i.e., $M_w(q_i, r_i) \mapsto (e_{w,i}, c_{w,i})$. A DPO reject sample is constructed when the outputs of the weak model and the strong model M_s are contradictory. Typical contradiction cases include: (i) $c_{s,i} \geq 0.5$ while $c_{w,i} < 0.5$; or (ii) $c_{s,i} < 0.5$ while $c_{w,i} \geq 0.5$. These discrepancies indicate a clear divergence in hallucination judgments between the two models, thereby providing effective supervision signals for preference learning.

We then construct a preference pair (y^+, y^-) by combining the positive sample $y_i^+ = (e_{s,i}, c_{s,i})$ and the negative sample $y_i^- = (e_{w,i}, c_{w,i})$, and incorporate it into the DPO training set $\mathcal{D}_{\text{DPO}}$. We construct a total of 55,425 preference pairs to support DPO training. A detailed case study of the constructed training data is provided in Appendix B.

To verify the quality of the DPO training data, we randomly sampled 10% for manual review. Two annotators with mathematical reasoning backgrounds independently labeled each pair. The resulting Cohen's Kappa coefficient of $k = 0.8421$ indicates almost perfect agreement, confirming the reliability of the chosen and rejected labels.

DPO-based Preference Learning for GHDM. DPO enhances the model's probabilistic preference for better responses through pairwise comparisons of preference samples, thus enabling the modeling of preference signals and supporting the calibration of reasoning confidence. We formulate the hallucination detection task as a *preference learning* problem. Given an input x that consists of a math problem and its associated reasoning process, we construct a pair of hallucination detection outputs $\{y^+, y^-\}$. In this setup, y^+ denotes the hallucination detection output produced by a strong model, which serves as the preferred (and presumably accurate) judgment. Conversely, y^- corresponds to the output generated by a weaker model and is treated as the less preferred (and potentially erroneous) alternative. The goal is to train a GHDM π_θ that prefers y^+ over y^-, thereby improving its ability to identify hallucinations in the reasoning process for the same input x.

DPO aims to maximize the relative log-probability margin between the preferred and less preferred outputs. The DPO loss objective is formally defined as:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\log \sigma\Bigg(\beta \cdot \Big[\log \frac{\pi_\theta(y^+ \mid x)}{\pi_\theta(y^- \mid x)} - \log \frac{\pi_{\text{ref}}(y^+ \mid x)}{\pi_{\text{ref}}(y^- \mid x)}\Big]\Bigg) \tag{1}$$

where π_θ and π_{ref} denote the conditional distributions of the target and reference models, respectively. β is a temperature parameter, and $\sigma(\cdot)$ is the sigmoid function. This objective encourages the target model to prefer y^+ over y^- more confidently than the reference model. Detailed training parameters are provided in Appendix C.

3.2 Confidence Fusion Mechanism

To enhance the reliability of hallucination confidence estimation, we propose a fusion mechanism that integrates two complementary scores: the global confidence score predicted by a GHDM and the local-level confidence score assessed by a PRM. Given a problem q_i and its associated reasoning process r_i, the output of the GHDM π_θ can be formally represented as

$$\pi_\theta(q_i, r_i) \mapsto (\hat{e}_i, \hat{c}_i), \tag{2}$$

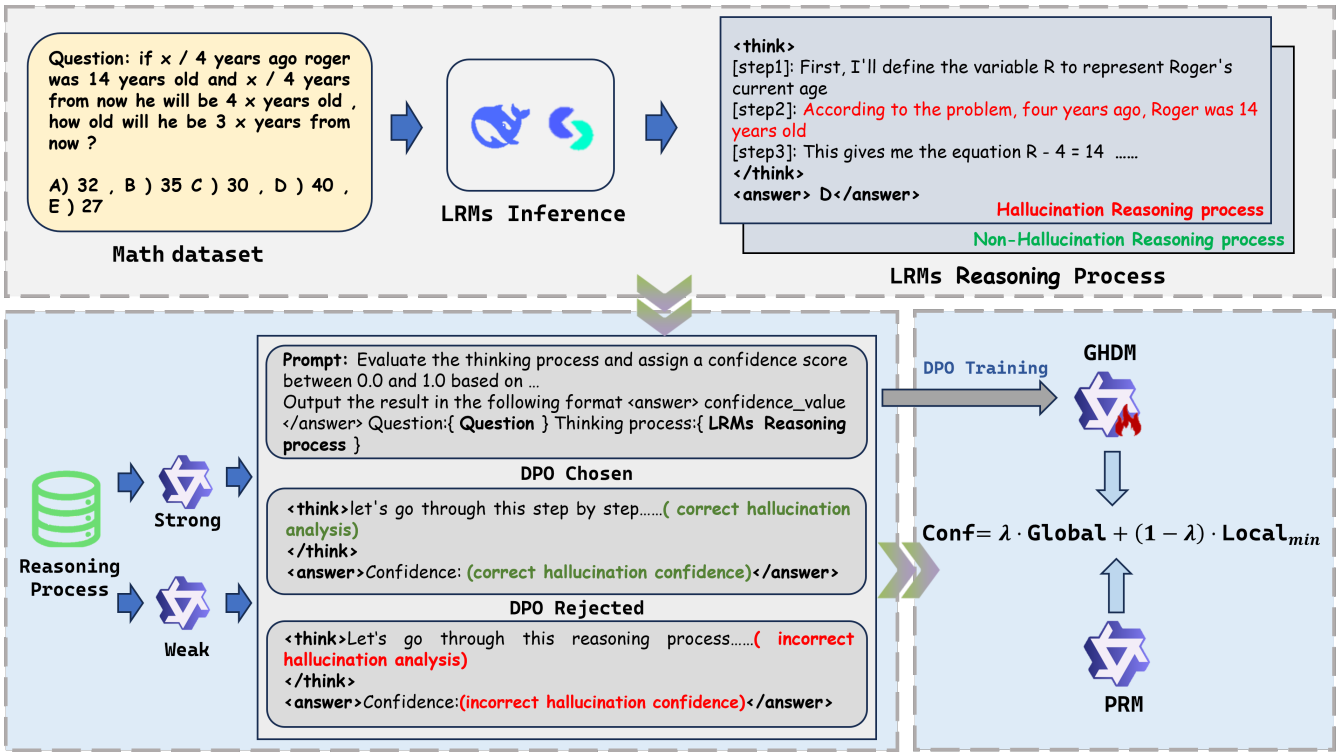

Figure 2: The overall pipeline of ConfFuse involves generating multi-step reasoning processes from large reasoning models, constructing a preference dataset based on strong–weak model agreement, and training a GHDM. It then incorporates local confidence signals from a PRM, fuses global and local confidence within the framework, and finally applies a threshold-based decision strategy to identify hallucinations in the reasoning process.

where $\hat{e}_i$ denotes the global hallucination analysis explanation, and $\hat{c}_i \in [0, 1]$ represents the corresponding global confidence score, indicating the reliability of r_i.

Meanwhile, the PRM evaluates each step $r_i^{(j)}$ in the reasoning process and produces a sequence of local-level scores $\{s_i^{(1)}, s_i^{(2)}, \ldots, s_i^{(n)}\}$, where each $s_i^{(j)} \in [0, 1]$ reflects the model's confidence that step j is logically valid and hallucination-free. To capture the weakest reasoning point that may determine the overall correctness, we take the minimum score across all steps to represent the most vulnerable part of the reasoning chain:

$$s_i^{\min} = \min_{j \in [1,n]} s_i^{(j)}, \tag{3}$$

a higher value of $s_i^{\min}$ indicates that even the weakest step maintains logical validity.

The final fused hallucination confidence score f_i is computed as a convex combination of the two confidence scores:

$$f_i = \lambda \cdot \hat{c}_i + (1 - \lambda) \cdot s_i^{\min}, \tag{4}$$

where $\lambda \in [0, 1]$ is a tunable hyperparameter that balances the importance of the model's overall confidence and the localized reasoning robustness. This fusion mechanism enables a comprehensive assessment of both global consistency and local vulnerabilities, and effectively mitigates the limitations of relying on a single evaluation perspective, particularly in cases where global and local signals diverge. A sample is ultimately classified as non-hallucinated if $f_i \geq \tau$ and hallucinated otherwise, where $\tau \in [0, 1]$ denotes the decision threshold. An information-theoretic analysis (Appendix D) demonstrates that the fused confidence score provides a more informative signal than either the global or local score alone.

4 Experiments

4.1 Experiment Settings

Datasets. We evaluate the proposed method under both in-distribution (ID) and out-of-distribution (OOD) settings. For the ID evaluation, we merge the test datasets of GSM8K (Cobbe et al. 2021) and MathQA (Amini et al. 2019)—denoted as (G+M) to align with the DPO training setup. For the OOD evaluation, we adopt two mathematical reasoning benchmarks, ASDiv (Miao, Liang, and Su 2021) and SVAMP (Patel, Bhattamishra, and Goyal 2021), to assess the model's ability to detect hallucinations under distribution shift. We use LRMs to automatically generate step-by-step reasoning processes on the test sets, which are then annotated for hallucinations to construct evaluation data. Further dataset details are provided in Appendix E.

Models. We adopt Qwen3-1.7B/8B (Team 2025) as base models for DPO training. The reasoning processes in the OOD test datasets are automatically generated by

DeepSeek-R1-Distill-Llama-8B (DeepSeek-AI 2025). Hallucination labels are annotated using GPT-4o (OpenAI et al. 2024). Based on the evaluation of PRM's hallucination detection capability in mathematical reasoning presented in (Pala et al. 2025), we adopt Qwen2.5-Math-PRM-7B as the local-level hallucination detector. In addition, we employ Marco-o1 (Zhao et al. 2024) as an alternative LLM architecture to evaluate the generalizability of ConfFuse.

Metrics. The following evaluation metrics are employed to evaluate hallucination detection performance:

Accuracy (ACC). Measures the model's ability to correctly distinguish hallucinated from non-hallucinated instances.

F1 Score (F1). The harmonic mean of precision and recall, providing a balanced evaluation when false positives and false negatives matter equally.

Area Under the Curve (AUC). AUC measures the ability of the hallucination detection model to distinguish between hallucinated and non-hallucinated samples across different decision thresholds.

Expected Calibration Error (ECE). ECE reflects the reliability of the model's confidence expression, where a lower value indicates that the model's confidence aligns more closely with its true performance, demonstrating better calibration capability.

Pearson Correlation Coefficient (PCC). PCC quantifies the correlation between global hallucination confidence and local-level PRM scores. A lower PCC indicates greater divergence between global and local assessments, suggesting that the two provide complementary perspectives for hallucination detection.

Baselines. We compare the proposed ConfFuse method against four representative categories of hallucination detection approaches: (1) *Hallucination detection frameworks*, which directly assess the consistency or factuality of LLM outputs using the model itself, such as SelfCheckGPT (Manakul, Liusie, and Gales 2023); (2) *LLM-as-a-judge paradigms*, which rely on a strong language model to judge whether a reasoning process contains hallucinations, such as DeepSeek-R1 (DeepSeek-AI 2025); (3) *Process reward models*, which provide fine-grained local-level hallucination scores to evaluate logical correctness within reasoning chains, e.g., PathFinder-PRM-7B (Pala et al. 2025); (4) *Global hallucination detection models*, which are fine-tuned via DPO to distinguish between hallucinated and faithful reasoning traces. A representative model in this category is GHDM-8B, proposed in this paper. Implementation details of the baselines and the proposed ConfFuse framework are provided in Appendix A.

4.2 Main Results

Performance on the in-distribution test dataset. As shown in Table 1, we train GHDM at two different parameter scales: GHDM-1.7B and GHDM-8B. Compared to their respective base models (Qwen3-1.7B and Qwen3-8B), these models achieved AUC improvements of 3.9% and 4.65% on the ID dataset, respectively, demonstrating that DPO training enhances the hallucination detection capability of the models. For the smaller GHDM-1.7B model, the F1 score also increases significantly by 8.64%, indicating that even with a smaller parameter size, appropriate optimization strategies can substantially boost hallucination detection performance. Compared to the current best-performing PRM, Qwen2.5-Math-PRM-7B, GHDM-8B achieves an F1 score that is 6.26% higher, demonstrating a stronger capability for hallucination detection in LRM reasoning scenarios. Moreover, compared with the QWQ-32B, GHDM-8B achieves comparable performance in terms of F1 and accuracy, while further improving AUC by 4.55%. This suggests that our DPO training strategy achieves overall performance comparable to that of much larger models and demonstrates superior hallucination detection capability. The ConfFuse-8B framework, which integrates hallucination confidence scores from both GHDM-8B and Qwen2.5-Math-PRM-7B, achieves the best performance on the ID dataset. Specifically, its F1 score improves by 2.73% compared to GHDM-8B and by 8.99% compared to Qwen2.5-Math-PRM-7B.

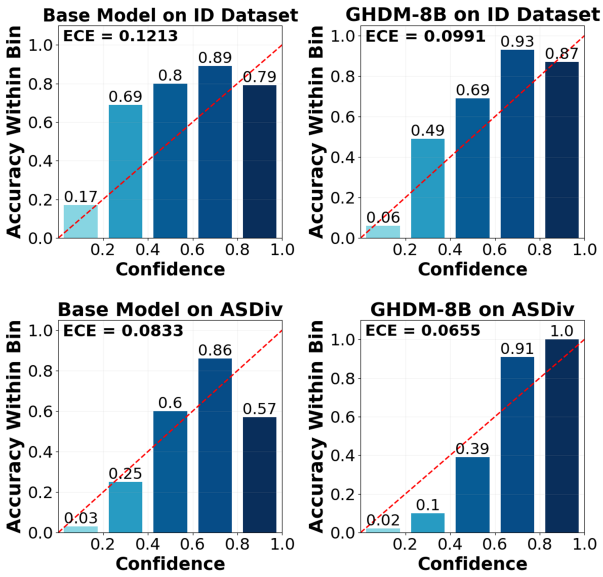

Figure 3: Confidence calibration results of GHDM-8B on two datasets. The top plot shows results on the ID dataset, and the bottom on OOD dataset. Each curve depicts the relationship between predicted confidence and observed accuracy, partitioned into confidence bins. The dashed diagonal line in the figure represents perfect calibration, where the predicted confidence exactly matches the observed accuracy.

Performance on Out-of-Distribution Datasets. We evaluate the generalization ability of various hallucination detection models on two OOD datasets: ASDiv and SVAMP. Both GHDM-1.7B and GHDM-8B exhibit consistent improvements over their respective base models. On SVAMP, GHDM-1.7B and GHDM-8B achieve AUC gains of 2.69% and 2.26%, respectively, confirming that DPO training not only enhances in-distribution performance but also improves overall discriminative ability under distribution shift. More-

Method Type	Model	G+M (ID)			ASDiv (OOD)			SVAMP (OOD)		
		AUC	F1	ACC	AUC	F1	ACC	AUC	F1	ACC
HD Framework	SelfCheckGPT	0.7300	0.6522	0.7243	0.6975	0.7876	0.8569	0.7590	0.6626	0.8250
	ChainPoll	0.8506	0.8077	0.8319	0.9433	0.8552	0.9380	0.9604	0.8947	0.9233
LLM-as-Judge	JudgeLRM-7B	0.7232	0.5997	0.6737	0.7939	0.5021	0.8324	0.7890	0.6239	0.7739
	QWQ-32B	0.8789	0.8220	0.8379	0.9435	0.8071	0.9121	0.9507	0.8703	0.9125
	DeepSeek-R1	**0.9421**	0.8538	0.8733	0.9571	0.8675	0.9351	0.9701	0.8831	0.9353
PRM	ReasonEval-7B	0.7164	0.5848	0.6555	0.8067	0.4349	0.8350	0.7850	0.5402	0.8296
	PathFinder-PRM-7B	0.8146	0.7350	0.7240	0.9171	0.7694	0.9001	0.9097	0.8260	0.8885
	Qwen2.5-Math-PRM-7B	0.8802	0.7694	0.7994	0.9363	0.8023	0.9169	0.9405	0.8465	0.9052
GHDM	Qwen3-1.7B-base	0.8032	0.6822	0.7558	0.8662	0.7493	0.8978	0.8531	0.7955	0.8554
	Qwen3-8B-base	0.8779	0.8047	0.8296	0.9338	0.8363	0.9209	0.9468	0.8764	0.9333
	GHDM-1.7B	0.8422	0.7686	0.8068	0.8585	0.7790	0.9169	0.8800	0.8216	0.8969
	GHDM-8B	0.9244	0.8320	0.8491	0.9606	0.8550	0.9208	0.9694	0.8864	0.9343
ConfFuse	ConfFuse-1.7B	0.8983	0.8008	0.8272	0.9335	0.7911	0.9188	0.9380	0.8467	0.9083
	ConfFuse-8B	0.9360	**0.8593**	**0.8739**	**0.9665**	**0.8709**	**0.9458**	**0.9777**	**0.8978**	**0.9365**

Table 1: Performance of hallucination detection methods on ID and OOD datasets.

over, the ConfFuse framework, which integrates global and local confidence scores, achieves further performance gains across all evaluation metrics. On ASDiv, ConfFuse-8B surpasses GHDM-8B by 1.59% in F1 (0.8709 vs. 0.8550) and 2.50% in accuracy (0.9458 vs. 0.9208). Compared to the strong baseline DeepSeek-R1, ConfFuse-8B achieves comparable performance, with similar F1 scores (0.8709 vs. 0.8675) and accuracy (0.9458 vs. 0.9351). These results highlight the robustness and effectiveness of ConfFuse in hallucination detection for LRM reasoning traces, demonstrating its strong generalization capability across both ID and OOD settings.

Confidence Calibration Analysis. As shown in Figure 3, the base model exhibits a certain degree of miscalibration, particularly in high-confidence regions where a clear gap remains between predicted confidence and actual accuracy. For example, on the ASDiv dataset, the base model achieves only 57% accuracy in the highest confidence bin, while GHDM-8B improves this to 100%; correspondingly, the ECE decreases from 0.0833 to 0.0655. On the ID dataset, ConfFuse raises the high-confidence accuracy from 79% to 87% and reduces the ECE from 0.1213 to 0.0991. These results indicate that GHDM-8B further enhance confidence calibration in high-confidence intervals.

Framework Generalization Analysis. We provide an extended discussion on the generalization capability of the proposed framework in appendix F.

4.3 Parameter Sensitivity Analysis

We perform a parameter sensitivity analysis to assess the impact of varying the global score weight λ (from GHDM) and its complementary local-level weight $1-\lambda$ (from PRM) in ConfFuse-8B, as illustrated in Figure 4. We report AUC scores on the ID dataset and the OOD dataset SVAMP. On the ID dataset, performance improves monotonically with increasing λ, reaching a peak AUC of 0.936 at $\lambda = 0.75$. A similar trend is observed on SVAMP, where the highest AUC of 0.978 also occurs at $\lambda = 0.75$, suggesting that

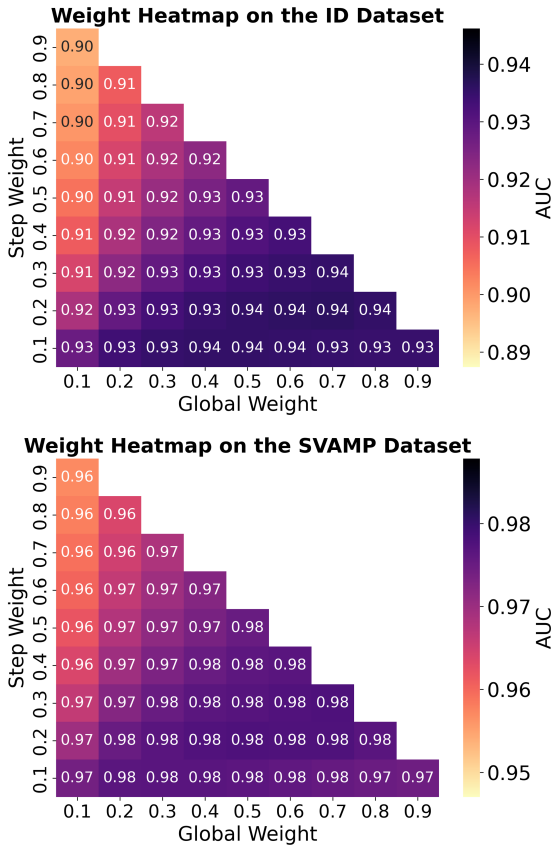

Figure 4: AUC heatmaps under varying weights of global and local-level confidence. The upper figure shows results on ID dataset, and the lower on SVAMP. Each cell indicates the AUC score for a specific weight combination, with darker colors representing better performance.

local-level robustness continues to contribute under distribution shift. As both datasets favor a strong—but not exclusive—reliance on global confidence, we adopt $\lambda = 0.75$ in all subsequent experiments. **This configuration achieves near-optimal results and underscores the importance of combining global and local-level confidence for robust hallucination detection.** We provide a detailed threshold analysis in Appendix G, confirming 0.5 as the optimal and consistent choice with training.

4.4 Ablation Studies

Ablation Study on Reasoning Process in GHDM. We investigate the impact of explicitly generating step-by-step reasoning in GHDM on hallucination detection performance. Table 2 compares a GHDM under two inference settings. On the ID dataset, enabling the reasoning process for hallucination analysis in GHDM-1.7B improves the F1 score from 0.5904 to 0.7686 (+17.82%) and ACC from 0.6981 to 0.8068 (+10.87%). The gains are even more pronounced on OOD datasets: on ASDiv, F1 increases by 24.55%, and on SVAMP by 27.20%. The GHDM-8B model exhibits a similar trend across all benchmarks. **These results demonstrate that, compared to relying solely on confidence scores, explicitly generating the reasoning process within GHDM substantially improves hallucination detection capability and enhances generalization performance in OOD scenarios.**

Model	G+M		ASDiv		SVAMP	
	F1	ACC	F1	ACC	F1	ACC
w/o R-1.7B	0.5904	0.6981	0.5335	0.8063	0.5496	0.7964
w/ R-1.7B	0.7686	0.8068	0.7790	0.9169	0.8216	0.8969
w/o R-8B	0.7947	0.8263	0.7193	0.8977	0.8098	0.8958
w/ R-8B	0.8320	0.8491	0.8550	0.9208	0.8864	0.9343

Table 2: We conduct an ablation study to evaluate the impact of hallucination analysis generation on the hallucination detection performance of GHDM. "w/o R" denotes models that output only a global confidence score, while "w/ R" denotes models that generate both hallucination analysis reasoning and confidence scores.

Ablation Study of Confidence Fusion Mechanism. As shown in Table 3, the F1 scores of different models are presented on the ID dataset and OOD datasets (ASDiv), comparing the results before and after applying the ConfFuse confidence fusion mechanism. Integrating global and local-level confidence scores leads to greater performance changes in GHDM. For example, the F1 score of GHDM-8B increases from 0.8320 to 0.8593 on the ID dataset and from 0.8550 to 0.8709 on ASDiv; GHDM-1.7B exhibits gains of 3.2% and 1.2%, respectively. In contrast, the changes observed in the Base models without DPO fine-tuning are relatively limited. The F1 score of 1.7B-Base improves by 0.7 points on ASDiv, while 8B-Base shows minimal change on the ID dataset (0.8047 to 0.8050) and a slight decrease on ASDiv (0.8363 to 0.8346). **These results indicate that the fusion mechanism has a more pronounced impact on models with established global discriminative capacity, particularly under distribution shift, while its impact on base models lacking hallucination-specific fine-tuning is minimal.**

Model	G+M	ASDiv
Qwen3-1.7B / ConfFuse	$0.6822 \rightarrow 0.7234$	$0.7493 \rightarrow 0.7565$
GHDM-1.7B / ConfFuse	$0.7686 \rightarrow 0.8008$	$0.7790 \rightarrow 0.7911$
Qwen3-8B / ConfFuse	$0.8047 \rightarrow 0.8050$	$0.8363 \rightarrow 0.8346$
GHDM-8B / ConfFuse	$0.8320 \rightarrow 0.8593$	$0.8550 \rightarrow 0.8709$

Table 3: Effect of ConfFuse confidence fusion on F1 scores for G+M and ASDiv datasets.

4.5 Correlation Analysis Between Global and Local Confidence Scores

Table 4 presents the results of a correlation analysis. As the parameter scale of GHDM increases from 1.7B to 8B, the PCC between the global and local confidence scores rises from 0.6250 to 0.6911, indicating that improved reasoning ability enhances their correlation. Conversely, after DPO training, the PCC values of GHDM decrease across all three datasets compared to the base models, suggesting reduced alignment but increased complementarity between the global and local confidence scores. **These findings support ConfFuse's design goal of integrating both perspectives to improve hallucination detection performance.** We provide case studies in Appendix H that further illustrate the complementary nature of PRM and GHDM in hallucination detection.

Model	G+M	ASDiv	SVAMP
Qwen3-1.7B	0.6331	0.7341	0.6958
GHDM-1.7B	0.6250	0.7175	0.6594
Qwen3-8B	0.7429	0.8217	0.7975
GHDM-8B	0.6911	0.7780	0.7694

Table 4: Pearson correlation coefficient of GHDM and base models on the ID dataset and OOD datasets.

5 Conclusion

This paper proposes ConfFuse, a unified hallucination detection framework that integrates global confidence scores from a trained GHDM with local-level assessments from a PRM. The GHDM is explicitly trained to evaluate hallucination risk across entire reasoning chains, enabling calibrated and holistic detection. ConfFuse facilitates the accurate identification of hallucinations within the reasoning processes of LRMs on mathematical tasks. Our analysis underscores the importance of evaluating intermediate reasoning steps, which are often overlooked by existing response-level methods. Experimental results across multiple mathematical reasoning benchmarks demonstrate that ConfFuse consistently outperforms strong baselines in hallucination detection performance.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62476283).

References

Amini, A.; Gabriel, S.; Lin, P.; Koncel-Kedziorski, R.; Choi, Y.; and Hajishirzi, H. 2019. MathQA: Towards Interpretable Math Word Problem Solving with Operation-Based Formalisms. arXiv:1905.13319.

Chen, J.; and Mueller, J. 2024. Automated Data Curation for Robust Language Model Fine-Tuning. arXiv:2403.12776.

Chowdhury, N.; Johnson, D.; Huang, V.; Steinhardt, J.; and Schwettmann, S. 2025. Investigating truthfulness in a pre-release o3 model.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; Hesse, C.; and Schulman, J. 2021. Training Verifiers to Solve Math Word Problems. arXiv:2110.14168.

Cohen, R.; Hamri, M.; Geva, M.; and Globerson, A. 2023. LM vs LM: Detecting Factual Errors via Cross Examination. arXiv:2305.13281.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Dhuliawala, S.; Komeili, M.; Xu, J.; Raileanu, R.; Li, X.; Celikyilmaz, A.; and Weston, J. 2023. Chain-of-Verification Reduces Hallucination in Large Language Models. arXiv:2309.11495.

Friel, R.; and Sanyal, A. 2023. Chainpoll: A high efficacy method for LLM hallucination detection. arXiv:2310.18344.

Gu, J.; Jiang, X.; Shi, Z.; Tan, H.; Zhai, X.; Xu, C.; Li, W.; Shen, Y.; Ma, S.; Liu, H.; Wang, S.; Zhang, K.; Wang, Y.; Gao, W.; Ni, L.; and Guo, J. 2025. A Survey on LLM-as-a-Judge. arXiv:2411.15594.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

He, J.; Liu, J.; Liu, C. Y.; Yan, R.; Wang, C.; Cheng, P.; Zhang, X.; Zhang, F.; Xu, J.; Shen, W.; Li, S.; Zeng, L.; Wei, T.; Cheng, C.; Liu, Y.; and Zhou, Y. 2025. Skywork Open Reasoner Series. https://capricious-hydrogen-41c.notion.site/Skywork-Open-Reaonser. Notion Blog.

Hu, X.; Ru, D.; Qiu, L.; Guo, Q.; Zhang, T.; Xu, Y.; Luo, Y.; Liu, P.; Zhang, Y.; and Zhang, Z. 2024. RefChecker: Reference-based Fine-grained Hallucination Checker and Benchmark for Large Language Models. arXiv:2405.14486.

Kang, L.; Deng, Y.; Xiao, Y.; Mo, Z.; Lee, W. S.; and Bing, L. 2025. First Try Matters: Revisiting the Role of Reflection in Reasoning Models. *arXiv preprint arXiv:2510.08308*.

Lam, R.; Sanchez-Gonzalez, A.; Willson, M.; Wirnsberger, P.; Fortunato, M.; Alet, F.; Ravuri, S.; Ewalds, T.; Eaton-Rosen, Z.; Hu, W.; et al. 2023. Learning skillful medium-range global weather forecasting. *Science*, 382(6677): 1416–1421.

Li, D.; Jiang, B.; Huang, L.; Beigi, A.; Zhao, C.; Tan, Z.; Bhattacharjee, A.; Jiang, Y.; Chen, C.; Wu, T.; Shu, K.; Cheng, L.; and Liu, H. 2025. From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge. arXiv:2411.16594.

Li, R.; Luo, Z.; and Du, X. 2024. FG-PRM: Fine-grained Hallucination Detection and Mitigation in Language Model Mathematical Reasoning. arXiv:2410.06304.

Lightman, H.; Kosaraju, V.; Burda, Y.; Edwards, H.; Baker, B.; Lee, T.; Leike, J.; Schulman, J.; Sutskever, I.; and Cobbe, K. 2023. Let's Verify Step by Step. arXiv:2305.20050.

Liu, T.; Zhang, Y.; Brockett, C.; Mao, Y.; Sui, Z.; Chen, W.; and Dolan, B. 2022. A Token-level Reference-free Hallucination Detection Benchmark for Free-form Text Generation. arXiv:2104.08704.

Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; and Zhu, C. 2023. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. arXiv:2303.16634.

Manakul, P.; Liusie, A.; and Gales, M. J. F. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. arXiv:2303.08896.

Miao, N.; Teh, Y. W.; and Rainforth, T. 2023. SelfCheck: Using LLMs to Zero-Shot Check Their Own Step-by-Step Reasoning. arXiv:2308.00436.

Miao, S.-Y.; Liang, C.-C.; and Su, K.-Y. 2021. A Diverse Corpus for Evaluating and Developing English Math Word Problem Solvers. arXiv:2106.15772.

Nguyen, X.-P.; Aljunied, M.; Joty, S.; and Bing, L. 2024. Democratizing LLMs for low-resource languages by leveraging their English dominant abilities with linguistically-diverse prompts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3501–3516.

Niu, M.; Haddadi, H.; and Pang, G. 2025. Robust Hallucination Detection in LLMs via Adaptive Token Selection. arXiv:2504.07863.

OpenAI; :; El-Kishky, A.; Wei, A.; Saraiva, A.; Minaiev, B.; Selsam, D.; Dohan, D.; Song, F.; Lightman, H.; Clavera, I.; Pachocki, J.; Tworek, J.; Kuhn, L.; Kaiser, L.; Chen, M.; Schwarzer, M.; Rohaninejad, M.; McAleese, N.; o3 contributors; Mürk, O.; Garg, R.; Shu, R.; Sidor, S.; Kosaraju, V.; and Zhou, W. 2025. Competitive Programming with Large Reasoning Models. arXiv:2502.06807.

OpenAI; :; Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; and Ramesh, A. 2024. GPT-4o System Card. arXiv:2410.21276.

Pala, T. D.; Sharma, P.; Zadeh, A.; Li, C.; and Poria, S. 2025. Error Typing for Smarter Rewards: Improving Process Reward Models with Error-Aware Hierarchical Supervision. arXiv:2505.19706.

Patel, A.; Bhattamishra, S.; and Goyal, N. 2021. Are NLP Models really able to Solve Simple Math Word Problems? arXiv:2103.07191.

Rafailov, R.; Sharma, A.; Mitchell, E.; Ermon, S.; Manning, C. D.; and Finn, C. 2024. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. arXiv:2305.18290.

Sahu, P.; Sikka, K.; and Divakaran, A. 2024. Pelican: Correcting Hallucination in Vision-LLMs via Claim Decomposition and Program of Thought Verification. *arXiv preprint arXiv:2407.02352*.

Tao, S.; Yao, L.; Ding, H.; Xie, Y.; Cao, Q.; Sun, F.; Gao, J.; Shen, H.; and Ding, B. 2024. When to Trust LLMs: Aligning Confidence with Response Quality. arXiv:2404.17287.

Team, Q. 2025. Qwen3 Technical Report. arXiv:2505.09388.

Wang, Y.; Reddy, R. G.; Mujahid, Z. M.; Arora, A.; Rubashevskii, A.; Geng, J.; Afzal, O. M.; Pan, L.; Borenstein, N.; Pillai, A.; et al. 2023. Factcheck-bench: Fine-grained evaluation benchmark for automatic fact-checkers. *arXiv preprint arXiv:2311.09000*.

Xia, S.; Li, X.; Liu, Y.; Wu, T.; and Liu, P. 2025a. Evaluating Mathematical Reasoning Beyond Accuracy. arXiv:2404.05692.

Xia, Y.; Jin, P.; Xie, S.; He, L.; Cao, C.; Luo, R.; Liu, G.; Wang, Y.; Liu, Z.; Chen, Y.-J.; et al. 2025b. Nature Language Model: Deciphering the Language of Nature for Scientific Discovery. *arXiv preprint arXiv:2502.07527*.

Xu, C.; Rosset, C.; Chau, E. C.; Corro, L. D.; Mahajan, S.; McAuley, J.; Neville, J.; Awadallah, A. H.; and Rao, N. 2024. Automatic Pair Construction for Contrastive Posttraining. arXiv:2310.02263.

Xu, F.; Hao, Q.; Zong, Z.; Wang, J.; Zhang, Y.; Wang, J.; Lan, X.; Gong, J.; Ouyang, T.; Meng, F.; et al. 2025. Towards large reasoning models: A survey of reinforced reasoning with large language models. *arXiv preprint arXiv:2501.09686*.

Yang, A.; Zhang, B.; Hui, B.; Gao, B.; Yu, B.; Li, C.; Liu, D.; Tu, J.; Zhou, J.; Lin, J.; Lu, K.; Xue, M.; Lin, R.; Liu, T.; Ren, X.; and Zhang, Z. 2024. Qwen2.5-Math Technical Report: Toward Mathematical Expert Model via Self-Improvement. arXiv:2409.12122.

Yang, B.; Mamun, M. A. A.; Zhang, J. M.; and Uddin, G. 2025. Hallucination Detection in Large Language Models with Metamorphic Relations. *arXiv preprint arXiv:2502.15844*.

Yao, Z.; Liu, Y.; Chen, Y.; Chen, J.; Fang, J.; Hou, L.; Li, J.; and Chua, T.-S. 2025. Are Reasoning Models More Prone to Hallucination? *arXiv preprint arXiv:2505.23646*.

Yeh, M.-H.; Kamachee, M.; Park, S.; and Li, Y. 2025. Can Your Uncertainty Scores Detect Hallucinated Entity? *arXiv preprint arXiv:2502.11948*.

Zhang, B.; Gao, C.; Yang, L.; Han, B.; Hu, M.; Luo, Z.; Geng, G.; Bai, X.; Zhang, J.; Yao, W.; et al. 2025. SafeConf: A Confidence-Calibrated Safety Self-Evaluation Method for Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 3483–3495.

Zhang, J.; Yao, W.; Chen, X.; and Feng, L. 2023. Transferable post-hoc calibration on pretrained transformers in noisy text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 13940–13948.

Zhao, Y.; Yin, H.; Zeng, B.; Wang, H.; Shi, T.; Lyu, C.; Wang, L.; Luo, W.; and Zhang, K. 2024. Marco-o1: Towards Open Reasoning Models for Open-Ended Solutions. arXiv:2411.14405.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685.