Can LLMs Reflect on Their Reasoning? A Probabilistic VC-Theoretic Perspective

(/pdf?id=ThtG0OvoNM)

Jae Oh Woo (/profile?id=~Jae_Oh_Woo1),
Mengdie Flora Wang (/profile?id=~Mengdie_Flora_Wang1),
Rahul Ghosh (/profile?id=~Rahul_Ghosh1),
Baishali Chaudhury (/profile?id=~Baishali_Chaudhury1),
Mun Young Kim (/profile?id=~Mun_Young_Kim1)

03 May 2025 (modified: 18 Sept 2025) Submitted to NeurIPS 2025 Conference, Senior Area Chairs, Area Chairs, Reviewers, Authors Revisions (/revisions?id=ThtG0OvoNM) BibTeX

Keywords: Reasoning, Self-Reflection, Vapnik–Chervonenkis Theory, Evaluation, LLM-as-a-Judge, Calibration

TL;DR: We introduce probabilistic extensions of VC dimension—PVC and C-PVC—to quantify LLMs' self-reflective reasoning capacity, and empirically evaluate their calibration and generalization behavior across mathematical domains.

Abstract:

As large language models (LLMs) continue to improve in complex reasoning tasks, an important question remains: to what extent can they evaluate the quality of their own reasoning? This work develops a unified theoretical and empirical framework to investigate that ability. We extend classical Vapnik-Chervonenkis (VC) dimension theory to probabilistic predictors by introducing two complexity measures—Probabilistic VC (PVC) and Calibration-aware PVC (C-PVC)—which capture a model's ability to generalize and to assign meaningful confidence scores. While these notions are defined for function classes, we propose empirical counterparts that quantify the number of categories a fixed model can confidently and accurately label under arbitrary labelings. These measures serve as practical indicators of a model's expressive capacity and its ability to assess uncertainty. To support this approach, we construct a benchmark of mathematical reasoning problems. Each model is asked to choose the better of two self-generated solutions, with reference judgments provided by a larger model ensemble. The results show consistent differences across models: OpenThinker2-7B achieves a high PVC score but low C-PVC, suggesting strong pattern coverage but limited calibration; JiuZhang3.0-7B exhibits the opposite trend; LLaMa-3.1-8B-Instruct maintains relatively stable performance in both. We observe that models post-trained with reinforcement learning or supervised objectives often produce overconfident judgments, particularly in tasks subject to extensive optimization. These observations are consistent with our theoretical framework, which links differences in PVC and C-PVC to generalization performance and confidence reliability. Nevertheless, the current metrics for reflection and calibration remain limited in granularity and sensitivity.

Developing more precise and robust measurement tools is essential for capturing subtle variations in introspective behavior. Advancing these foundations may enable future models to improve through more reliable self-evaluation—an important step toward building systems capable of monitoring and understanding their own reasoning processes.

Checklist Confirmation: I confirm that I have included a paper checklist in the paper PDF.

Supplementary Material: zip (/attachment?id=ThtG0OvoNM&name=supplementary_material)

Reviewer Nomination: Rahul Ghosh (/profile?id=~Rahul_Ghosh1)

Responsible Reviewing: We acknowledge the responsible reviewing obligations as authors.

Primary Area: Evaluation (e.g., methodology, meta studies, replicability and validity, human-in-the-loop)

LLM Usage: Editing (e.g., grammar, spelling, word choice), Understanding technical concepts, Data processing/filtering, Visualizing results for submission, Facilitating or running experiments, Implementing standard methods

Declaration: I confirm that the above information is accurate.

Submission Number: 6489

Filter by reply type... | Filter by author... | Search keywords... | Sort: Newest First

− = ≡

Everyone | Program Chairs | Submission6489 Authors | Submission6489... | Submission6489 Area... | Submission6489... | Submission6489... | Submission6489... | Submission6489... | Submission6489... | ✖

27 / 27 replies shown

Add: **Withdrawal**

Paper Decision

Decision by Program Chairs 17 Sept 2025, 05:42 (modified: 18 Sept 2025, 06:25) Program Chairs, Authors

Revisions (/revisions?id=M5OsjlVK92)

Decision: Reject

Comment:

This paper proposes new theoretical complexity measures, Probabilistic VC (PVC) and Calibration-aware PVC (C-PVC), along with empirical counterparts, to evaluate LLMs' capacity for generalization and confidence calibration in self-evaluation. The reviewers initially raised some concerns such as novelty of the work and hyperparameter selection. While the authors provide the detailed responses to resolve the concerns, two reviewers are still learning to negative side mainly due to ambiguities in the current draft. Therefore, AC recommend to reject this paper.

Author Final Remarks
by Authors

Author Final Remarks

by Authors (Jae Oh Woo (/profile?id=~Jae_Oh_Woo1), Baishali Chaudhury (/profile?id=~Baishali_Chaudhury1), Mengdie Flora Wang (/profile?id=~Mengdie_Flora_Wang1), Rahul Ghosh (/profile?id=~Rahul_Ghosh1), +1 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission6489/Authors))

11 Aug 2025, 10:42 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Author Final Remarks:

We sincerely thank all reviewers for their thoughtful engagement during the discussion phase and for the score improvements. We deeply appreciate the time, attention, and constructive feedback you have devoted, which has been instrumental in refining the clarity and positioning of our work. Our study focuses on a targeted, solution-level introspection task: given two model-generated solutions to the same problem, can the model identify the more likely correct one and report calibrated confidence? The resulting PVC and C-PVC measures are then aggregated at the category level to capture domain-sensitive introspective ability. In the revised version, we will make the scope of our work explicit and ensure the core strengths are clearly conveyed—reframing self-evaluation as probabilistic binary discrimination (PVC, C-PVC) with theoretical generalization and sample-complexity bounds; providing a model- and dataset-agnostic protocol for diagnosing calibration and overconfidence; and presenting cross-domain evidence (e.g., TruthfulQA, CommonsenseQA) demonstrating that these measures are not limited to math-specific artifacts. Revisions will also deliver improved figure clarity and presentation, transparent hyperparameter reporting, stronger positioning in related work (including PAC-Bayes and self-reflection methods such as Reflexion/Self-Refine), a dedicated Limitations section, and full code/data/scripts release. We are deeply grateful to all reviewers and the AC/SAC for their careful oversight, which has been invaluable in strengthening the clarity, scope, and impact of our work.

Discussion Phase

Official Comment by Area Chair 3sY6 03 Aug 2025, 00:53

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

Dear Authors and Reviewers,

I would like to thank the authors for providing detailed rebuttal messages.

For the reviewers, I would like to encourage you to carefully read all other reviews and the author responses and engage in an open exchange with the authors. Please post your first response as soon as possible within the discussion time window, so there is time for back and forth discussion with the authors. Ideally, all reviewers will respond to the authors, so that the authors know their rebuttal has been read.

Best regards,
AC

Author AC Confidential Comment by Authors

Author AC Confidential Comment

by Authors (Jae Oh Woo (/profile?id=~Jae_Oh_Woo1), Baishali Chaudhury (/profile?id=~Baishali_Chaudhury1), Mengdie Flora Wang (/profile?id=~Mengdie_Flora_Wang1), Rahul Ghosh (/profile?id=~Rahul_Ghosh1), +1 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission6489/Authors))

28 Jul 2025, 14:02 (modified: 30 Jul 2025, 17:42) Program Chairs, Senior Area Chairs, Area Chairs, Authors

Revisions (/revisions?id=zCWoxs9IN5)

Comment:

We are sincerely grateful to the AC and SAC for their thoughtful oversight throughout the review process. We appreciate the diversity of viewpoints expressed and the careful attention given to both the theoretical and empirical components of this work. In this final note, we respectfully ask for your consideration of the broader context and consistent strengths of the submission—particularly in light of one review that appears to rest on a significant misalignment of scope.

Our paper addresses a critical and underexplored question: not whether a language model can reason correctly, but whether it can recognize when it has reasoned correctly. We propose a new perspective that reframes self-evaluation as a **probabilistic binary classification task**, in which the model is asked to distinguish correct from incorrect reasoning outputs with calibrated confidence. This reframing enables the introduction of two novel complexity measures—**Probabilistic VC (PVC)** and **Calibration-aware PVC (C-PVC)**—which extend classical VC theory to capture both expressiveness and introspective reliability. These measures are theoretically grounded, empirically computable, and directly relevant to the safe and trustworthy deployment of LLMs.

This contribution is not only conceptual but operational:

- It yields sample complexity bounds and generalization guarantees;
- It is practically realizable across a wide range of language models—including both open-source and proprietary systems—and is readily applicable to diverse benchmark datasets;
- And it offers meaningful trends in calibration behavior across different training paradigms, such as SFT, RL, and distillation.

This core contribution was consistently recognized across three of the four reviews:

- **Reviewer Gx9q** endorsed the theoretical foundations and practical relevance of our PVC-based analysis;
- **Reviewer jy8h** highlighted the value of our framework in identifying overconfidence and calibration issues in RL-trained models—an increasingly important concern in LLM alignment;
- **Reviewer R5yk** acknowledged the clarity of motivation and the novelty of our framework, while offering constructive suggestions regarding parameter sensitivity and interpretation.

Taken together, these reviews affirm that the paper presents a rigorous, applicable, and timely contribution to the study of LLM introspection.

Only **Reviewer 15u2** expresses a dissenting view, primarily on the grounds that our framework does not model the full complexity of dynamic, multi-step reasoning. While we respect this broader aspiration, it reflects a mischaracterization of the paper's clearly defined goal. As stated explicitly throughout, our aim is not to emulate reasoning chains, but to quantify whether a model can evaluate the correctness of its outputs. **Reviewer 15u2** raises no technical objections to our theoretical derivations, empirical methodology, or experimental findings, but instead reinterprets the core research question. We believe such a divergence in scope—absent any substantive critique—should not override three reviews that directly and positively engage with the actual content and contributions of the work.

In closing, we strongly urge that the final evaluation reflect the full scope, rigor, and significance of the paper as it is—not as it was misread to be. The ability for a model to know when it is right is foundational for alignment, trust, and real-world reliability. Our submission offers a principled, verifiable, and actionable framework for quantifying this capacity. We are grateful for the opportunity to present this work to the NeurIPS community, and for your careful and fair consideration.

Official Review of Submission6489 by Reviewer Gx9q

Official Review by Reviewer Gx9q 03 Jul 2025, 12:32 (modified: 18 Sept 2025, 08:42)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer Gx9q

Revisions (/revisions?id=ARUWR9n02e)

Summary:
The paper proposes a new measure of calibration of LLMs. The paper begins by deriving a probabilistic version of the VC-dimension based upon past work on a real-valued version of the VC-dimension. The paper also derives a relaxed version of the probabilistic VC-dimension (PVC) and a calibration-aware VC-dimension (C-PVC), which can then be used for real-world measurement. The paper presents both an empirical test of the measure across a number of task categories and some theoretical results on important model class bounds. The paper finds differing dimensions for different LLMs, which are stated to come from the different training methodologies of the LLMs.

Strengths And Weaknesses:
The paper is of high quality and significance. The paper does a thorough job of both theoretical derivation and empirical testing. I especially appreciate the interpretations of how different model training methods lead to different PVC and C-PVC scores. I also found the derivation of the sample complexity (mirroring what exists in the standard VC-Dimension) to be very useful for evaluating models.

The weaknesses of the paper are few. There was one area I was left wondering, and I think it would increase the validity of the measure, and that is doing a similar assessment of PVC and C-PVC for prompting schemes as was done for the different LLM models. In section 3.2, the paper points out varying prompts is a function class for which there would be different VC-dimensions. I think demonstrating that would shore up the utility of the proposed PVC and C-PVC measures.

Quality: 3: good

Clarity: 3: good
Significance: 3: good
Originality: 3: good
Questions:

1. For the empirical setup, I understand the evaluation of the results' goodness as being done by the Judge Ensemble component. Could this evaluation also work with something like actual labels?

Limitations:
The authors addressed the limitations that I could see within the paper.

Rating: 5: Accept: Technically solid paper, with high impact on at least one sub-area of AI or moderate-to-high impact on more than one area of AI, with good-to-excellent evaluation, resources, reproducibility, and no unaddressed ethical considerations.
Confidence: 2: You are willing to defend your assessment, but it is quite likely that you did not understand the central parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.
Ethical Concerns: NO or VERY MINOR ethics concerns only
Paper Formatting Concerns:
None

Code Of Conduct Acknowledgement: Yes
Responsible Reviewing Acknowledgement: Yes
Final Justification:
After the discussion period and reviewing the other reviews, I will stand by accept rating. I believe there is merit in the work and the authors addressed any issues I had with the work.

Rebuttal by Authors

Rebuttal
by Authors (👁 Jae Oh Woo (/profile?id=~Jae_Oh_Woo1), Baishali Chaudhury (/profile?id=~Baishali_Chaudhury1), Mengdie Flora Wang (/profile?id=~Mengdie_Flora_Wang1), Rahul Ghosh (/profile?id=~Rahul_Ghosh1), +1 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission6489/Authors))

📅 28 Jul 2025, 13:45 (modified: 31 Jul 2025, 11:56)

👁 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=LvFX96nBNx)

Rebuttal:
We sincerely thank **Reviewer Gx9q** for the generous and thoughtful review. We deeply appreciate the recognition of our paper’s theoretical rigor, the practical value of the proposed PVC and C-PVC metrics, and the clarity of the empirical results. We are particularly grateful for your insight regarding the relationship between training paradigms and calibration, as well as for emphasizing the utility of our sample complexity analysis in interpreting model behavior.

Below we address the specific points raised:

Issue 1 – On the Assessment of PVC and C-PVC under Prompt Variation

We fully agree with the reviewer that prompting strategies can be viewed as inducing distinct function classes, each with its own capacity characteristics. As noted in Section 3.2, our framework is compatible with such a view: prompts, as part of the conditioning input, naturally influence the hypothesis space over which PVC and C-PVC are defined.

While the current work focuses on comparing models under a shared prompting scheme to isolate the effects of training, our formulation extends directly to cases with prompt variation. We appreciate the suggestion and will clarify this generality more explicitly in the revised manuscript.

Issue 2 – On the Use of Ensemble Judges versus Ground-Truth Labels

This is an excellent point. In our setting, establishing ground truth for complex reasoning outputs is inherently noisy. To ensure consistency across examples, we used an ensemble of independently queried models as proxy judges—a method akin to majority vote or inter-annotator agreement in human evaluation. This stabilizes noisy judgments and mitigates single-model bias.

That said, our framework is agnostic to the supervision source. If explicit ground-truth labels are available—whether from human experts, symbolic solvers, or curated datasets—they can be directly integrated into the same PVC and C-PVC formulations without modification. This highlights the flexibility of our evaluation protocol: ensemble-based voting is one valid instantiation, not a requirement.

We will revise the paper to make this flexibility clearer and to indicate how other judgment sources can be incorporated.

Once again, we are grateful to **Reviewer Gx9q** for the constructive feedback and for recognizing the contribution and extensibility of our framework. The reviewer's comments demonstrate clear engagement with both the theoretical and empirical components of the work and provide a fair and accurate assessment of the paper's contributions. We sincerely appreciate this thoughtful and rigorous evaluation.

Replying to Rebuttal by Authors

Mandatory Acknowledgement by Reviewer Gx9q

Mandatory Acknowledgement by Reviewer Gx9q 04 Aug 2025, 14:42

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Mandatory Acknowledgement: I have read the author rebuttal and considered all raised points., I have engaged in discussions and responded to authors., I have filled in the "Final Justification" text box and updated "Rating" accordingly (before Aug 13) that will become visible to authors once decisions are released., I understand that Area Chairs will be able to flag up Insufficient Reviews during the Reviewer-AC Discussions and shortly after to catch any irresponsible, insufficient or problematic behavior. Area Chairs will be also able to flag up during Metareview grossly irresponsible reviewers (including but not limited to possibly LLM-generated reviews)., I understand my Review and my conduct are subject to Responsible Reviewing initiative, including the desk rejection of my co-authored papers for grossly irresponsible behaviors. https://blog.neurips.cc/2025/05/02/responsible-reviewing-initiative-for-neurips-2025/ (https://blog.neurips.cc/2025/05/02/responsible-reviewing-initiative-for-neurips-2025/)

Replying to Rebuttal by Authors

Response to Rebuttal

Official Comment by Reviewer Gx9q 05 Aug 2025, 06:34

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

After reading through the Author's response to my rebuttal as well as the points raised in the other reviews, I feel that there is a valuable contribution in the paper. I do find the hyperparameter selection problem raised by R5yk to be an important one and not fully addressed. However, on the whole, I think this is a paper with a contribution and will stay at my rating of Accept.

Replying to Response to Rebuttal

Official Comment by Authors

Official Comment

by Authors (Jae Oh Woo (/profile?id=~Jae_Oh_Woo1), Baishali Chaudhury (/profile?id=~Baishali_Chaudhury1), Mengdie Flora Wang (/profile?id=~Mengdie_Flora_Wang1), Rahul Ghosh (/profile?id=~Rahul_Ghosh1), +1 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission6489/Authors))

05 Aug 2025, 07:55

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

We sincerely appreciate your positive evaluation and your recognition of this work’s contribution. Your thoughtful engagement throughout the review process has been truly encouraging.

We also fully agree that hyperparameter selection is important. In the revised materials, we will provide a detailed walkthrough of how the contrast score was applied to evaluate candidate thresholds and select the final configuration. We hope this added clarity meaningfully improves the reproducibility of our approach.

Thank you again for your generous and constructive feedback—it has directly helped us strengthen the rigor and accessibility of the paper.

Official Review of Submission6489 by Reviewer jy8h

Official Review by Reviewer jy8h 03 Jul 2025, 01:18 (modified: 18 Sept 2025, 08:42)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer jy8h

Revisions (/revisions?id=QsGqyncz4C)

Summary:

This paper investigates whether large language models (LLMs) can evaluate the quality of their own reasoning, presenting a unified theoretical and empirical framework rooted in statistical learning theory. The authors extend classical Vapnik-Chervonenkis (VC) dimension theory to probabilistic predictors, introducing two key measures: Probabilistic VC (PVC) and Calibration-aware PVC (C-PVC). PVC quantifies a model's capacity to make confident predictions across problem categories, while C-PVC additionally requires that confidence scores align with actual correctness probabilities .

The theoretical framework establishes connections between these measures and generalization performance, providing sample complexity bounds analogous to classical VC theory. Empirically, the authors construct a benchmark of mathematical reasoning problems, where models are asked to self-evaluate between two generated solutions, with judgments validated by a larger model ensemble. Experiments on 11 open-source 7–8B parameter models reveal distinct patterns: OpenThinker2-7B shows high PVC but low C-PVC, indicating strong pattern coverage but poor calibration; JiuZhang3.0-7B exhibits the opposite; and LLaMa-3.1-8B-Instruct balances both metrics .

Strengths And Weaknesses:

Considering the following four perspectives: Quality, Clarity, Significance, and Originality, the strengths and weaknesses of the paper are summarized as follows.

Strengths

1. The paper got clear definitions (e.g., PVC, C-PVC), theorems, and experimental protocols. The mathematical formulations are precise, and the comparison of model performance in Table 1 and Figures 2–3 effectively conveys the key findings. The discussion of training regime impacts (e.g., RL vs. SFT) provides practical insights for model development. The authors also acknowledge limitations (e.g., current metrics' granularity), demonstrating scientific rigor.
2. The work addresses a critical challenge in LLM research: the ability to assess reasoning quality and uncertainty. By linking PVC/C-PVC to calibration and generalization, the framework advances tools for building trustworthy systems, particularly for safety-critical applications. The empirical finding that RL-tuned models often exhibit overconfidence (e.g., Qwen2.5-7B-Instruct) highlights practical implications for training methodologies, urging the community to prioritize both expressiveness and calibration.

Weaknesses

However, considering the high quality of papers required by the conference and the degree of completion I believe the papers should have, the weaknesses of the paper are summarized as follows.

1. First of all, there are some minor problem with the writing. In details, Figure 1 is blurry and not pinned to the top. The author may not have much experience in submitting papers, so maybe should pay attention to this in the future. In addition, there are many problems with the figures in the following parts: the font size is too small, the labels overlap, etc. I will explain this in detail in the following questions. But in short, the author can pay attention to this point in the future, which is very important for the rigor of the paper and the reading experience.
2. The benchmark focuses on mathematical reasoning, which may not fully capture the diversity of real-world LLM applications (e.g., natural language understanding, code generation). The paper lacks validation in other domains, leaving open questions about the framework's applicability beyond structured problems. Additionally, the reliance on synthetic problem pairs may not reflect the complexity of unstructured reasoning tasks.
3. The related work section discusses VC theory and calibration but could more deeply contextualize the paper's contributions relative to recent advances in LLM self-reflection (e.g., Reflexion, Self-Refine). The authors mention these heuristics but do not explicitly compare their framework to such methods, leaving readers unclear about how PVC/C-PVC complements or outperforms existing approaches.

Quality: 2: fair

Clarity: 3: good

Significance: 2: fair

Originality: 2: fair

Questions:

My questions are basically consistent with the weaknesses listed above. Please clarify the confusion or solve the limitations.

1. First some questions and suggestions on paper writing. They are as follows: 1.1 Figure 1. The figure is blurry and not placed in [h]. Usually, such figures are exported in PDF format and displayed at the top of the first few pages of the paper to illustrate the methods and models of the paper. The author can pay attention to this. 1.2 Figure 2&3: The annotations in the figures overlap. This may be an oversight by the author, please pay attention to this. At the same time, the font size of the figure should also be paid attention to.
2. My concern is that the benchmark exclusively uses mathematical reasoning tasks. How does this framework generalize to non-mathematical domains (e.g., commonsense reasoning, ethical dilemmas)? Test PVC/C-PVC on diverse tasks (e.g., ARC, TruthfulQA) or provide theoretical justification for cross-domain applicability. Broader validation would strengthen claims about PVC's universality; failure to address this may reduce the score for novelty/impact.

Limitations:

There are some limitations of this work, but these may not be or maybe the weekness or questions for this paper. These may be solved in future work.

1. While acknowledging the "coarse granularity" of current metrics (Abstract), the paper focuses exclusively on mathematical reasoning for evaluating self-reflection. This raises questions about whether the proposed PVC/C-

PVC framework generalizes to other reasoning domains (e.g., ethical or commonsense reasoning). The authors should explicitly discuss this limitation and justify their domain choice.

2. The benchmark relies on larger LLMs as "expert judges" (Sec 5.1). This introduces potential circularity, as the same model families may share systemic biases. The paper should address how judge quality impacts C-PVC reliability, especially for lesser-studied models like JiuZhang3.0-7B.

Rating: 4: Borderline accept: Technically solid paper where reasons to accept outweigh reasons to reject, e.g., limited evaluation. Please use sparingly.

Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

Ethical Concerns: NO or VERY MINOR ethics concerns only

Paper Formatting Concerns:

N/A.

Code Of Conduct Acknowledgement: Yes

Responsible Reviewing Acknowledgement: Yes

Final Justification:

Final Revision: Thank you for your four responses. I have reconsidered the current score and have increased it slightly. I believe it is appropriate.

Considering your responses, A2, A3, and A4 address my previous concerns and provide the primary basis for the current score, which I believe is appropriate. Regarding the formatting and figures in your paper, the author will need to pay close attention in the next revision. Thank you for your response.

Rebuttal by Authors

Rebuttal

by Authors (👁 Jae Oh Woo (/profile?id=~Jae_Oh_Woo1), Baishali Chaudhury (/profile?id=~Baishali_Chaudhury1), Mengdie Flora Wang (/profile?id=~Mengdie_Flora_Wang1), Rahul Ghosh (/profile?id=~Rahul_Ghosh1), +1 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission6489/Authors))

📅 30 Jul 2025, 03:15 (modified: 31 Jul 2025, 11:56)

👁 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=PmBD59zvJi)

Rebuttal:

We sincerely thank **Reviewer jy8h** for their thoughtful and constructive feedback, which has helped us identify important areas for improvement in our paper. We appreciate the reviewer's careful examination of both the theoretical foundations and practical implications of our work, and we address each point comprehensively below.

Issue 1 – On the Presentation Quality

We thank the reviewer for pointing out the issues related to figure clarity and layout. Ensuring that our presentation matches the rigor of our theoretical and empirical contributions is an essential part of effective communication, and we take this responsibility seriously.

To address these concerns, we will revise all figures to meet a high standard of visual clarity and coherence:

- Figure 1 will be re-rendered with improved resolution and appropriately positioned to function as a clear visual overview of our framework.
- Figures 2 and 3 will be reformatted to eliminate label/numbers overlaps, increase font sizes, and enhance overall readability.

We appreciate the reviewer's attention to these details, which ultimately strengthen the accessibility and impact of the paper.

> **Issue 2** – On the Generalization Beyond Mathematical Reasoning

We appreciate the reviewer's thoughtful suggestion to evaluate whether PVC and C-PVC generalize beyond mathematical reasoning tasks. In response, we **extended our empirical analysis to two widely used non-mathematical benchmarks**: **TruthfulQA (domenicrosati/TruthfulQA)**, which assesses factual correctness under adversarial phrasing, and **CommonsenseQA (tau/commonsense_qa)**, which evaluates everyday commonsense inference.

To ensure a balanced evaluation, we grouped each dataset into 10 broad categories and sampled 240 questions per benchmark. We applied the same analytical pipeline as in the main paper, computing PVC, C-PVC, Expected Calibration Error (ECE), sample complexity, and actual error. Thresholds were selected using the same optimization strategy to maximize contrast between PVC and C-PVC across models described in **Issue 2** of **Reviewer R5yk**'s rebuttal: we used $\gamma = 0.56$, $\tau = 0.23$ for TruthfulQA, and $\gamma = 0.58$, $\tau = 0.22$ for CommonsenseQA.

The results are summarized below. For each metric, values are reported as TruthfulQA / CommonsenseQA:

Model	PVC	C-PVC	ECE	Sample Complexity	Actual Error
Qwen2.5-7B (Pretrain)	5 / 4	2 / *2*	0.318 / 0.303	152 / 145	0.435 / 0.429
Qwen2.5-7B-Instruct (SFT+RL)	*7 / 7*	*3* / **3**	**0.247** / 0.290	189 / 207	0.388 / 0.396
Qwen2.5-Math-7B-Instruct (SFT+RL)	2 / 3	1 / *2*	0.388 / 0.329	95 / 124	0.538 / 0.483
Llama-3.1-8B-Instruct (SFT+DPO)	2 / 0	0 / 0	0.462 / 0.408	95 / **62**	0.588 / 0.533
OpenThinker2-7B (SFT)	5 / 4	0 / 0	0.395 / 0.378	152 / 145	0.456 / 0.454
DeepSeek-R1-Distill-Qwen-7B (Distill)	4 / 4	2 / *2*	0.358 / 0.330	133 / 145	0.500 / 0.450
Bespoke-Stratos-7B (SFT)	7 / 6	2 / 1	0.315 / 0.338	189 / 186	0.433 / 0.413
JiuZhang3.0-7B (SFT)	0 / 3	0 / **3**	0.309 / **0.149**	**57** / 124	0.546 / 0.467

Model	PVC	C-PVC	ECE	Sample Complexity	Actual Error
Ministral-8B-Instruct-2410 (SFT+RL)	2 / 7	0 / 0	0.453 / 0.339	95 / 207	0.542 / 0.421
Open-Reasoner-Zero-7B (RL)	4 / 7	1 / 1	0.341 / 0.327	133 / 207	0.454 / 0.413
s1.1-7B (SFT)	**8** / **9**	**4** / **3**	0.256 / 0.267	208 / 248	**0.379** / **0.348**

These results offer several key takeaways:

- **Cross-Domain Generalization**: PVC and C-PVC capture meaningful shifts in model behavior across domains. Models such as s1.1 and Bespoke-Stratos-7B show improved calibration (higher C-PVC) on TruthfulQA and CommonsenseQA compared to math tasks. Conversely, LLaMa-3.1-8B-Instruct, which showed strong calibration on mathematical reasoning, performs worse on the new tasks. This suggests our metrics are not narrowly tailored to math problems, but instead reflect domain-sensitive introspective ability.
- **Training Regime Effects Are Stable**: Models fine-tuned with RL (e.g., Qwen2.5-7B-Instruct, Open-Reasoner-Zero-7B) continue to exhibit high PVC but relatively low C-PVC across domains, indicating persistent overconfidence. On the other hand, SFT-only models like s1.1 demonstrate both strong discrimination and calibration, attaining the highest PVC and C-PVC scores. Distilled models (e.g., DeepSeek-R1) maintain balanced profiles, reinforcing the consistency of training effects observed in the main paper.
- **Theoretical Implications**: The consistency of PVC/C-PVC across diverse benchmarks strengthens our core claim: these measures offer a robust and scalable tool for auditing model introspection, helping to evaluate when models can trust their own judgments—a key requirement for safe and reliable deployment.

We will include this extended cross-domain analysis in the revised manuscript to further support the generalizability and practical utility of our proposed framework.

Issue 3 - On the relation to Reflexion and Self-Refine

We thank the reviewer for raising an important point regarding the relationship between our work and recent self-reflection methods such as Reflexion and Self-Refine. We would like to clarify both the conceptual distinction and the complementary nature of these approaches.

Reflexion and Self-Refine focus on improving task performance through iterative refinement and natural language feedback. These methods assume that a model can recognize when its output is flawed and use that recognition to self-correct in subsequent attempts. However, they do not explicitly measure or quantify this self-recognition ability.

Our work addresses this foundational gap by formalizing and quantifying a model's ability to introspect—i.e., to identify which of two candidate outputs is more likely to be correct, and to do so with calibrated confidence. PVC and C-PVC offer theoretically grounded capacity measures for this introspective competence, which underpins the success of self-reflective techniques.

Importantly, **these approaches are not directly comparable**: Reflexion-type methods operate over multi-step trajectories with language-based revision, whereas our framework evaluates a single-step binary introspection task. Comparing them empirically would be technically invalid and conceptually misleading.

Thus, we view our work not as a competitor to Reflexion or Self-Refine, but as a complementary analytical tool that can assess and potentially explain when and why such methods succeed. We will include this clarification and a more detailed contextualization.

> **Issue 4** - On the Judge quality and bias

We appreciate the reviewer's concern about potential circularity in using LLMs as expert judges. However, we would like to clarify several important aspects of our evaluation setup that help mitigate these concerns.

First, we use an ensemble of three diverse judges (Claude 3.7 Sonnet, Amazon Nova Premier, and DeepSeek-R1) rather than relying on a single judge. As shown in our correlation analysis (Appendix H, Table 3), **these judges exhibit different evaluation patterns and biases, as evidenced by their varying correlation coefficients with the candidate models** (average correlation = 0.18 for Nova Premier, -0.09 for Claude Sonnet 3.7, and 0.01 for DeepSeek-R1). This diversity helps reduce systematic biases that might exist in any single judge.

Second, we carefully selected judge models from different model families than our candidate models to avoid circularity. The models we evaluate (OpenThinker2-7B, JiuZhang3.0-7B, etc.) are based on different architectures and training methodologies than our judge models. This architectural and methodological diversity helps ensure independence between the judges and the evaluated models.

We agree that judge quality is an important consideration, however, our current setup with multiple diverse judges and careful separation of model families provides a robust evaluation framework while minimizing potential circularity concerns.

Replying to Rebuttal by Authors

Reply to the author

Official Comment by Reviewer jy8h 08 Aug 2025, 12:43

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

Final Revision: Thank you for your four responses. I have reconsidered the current score and have increased it slightly. I believe it is appropriate.

Considering your responses, A2, A3, and A4 address my previous concerns and provide the primary basis for the current score, which I believe is appropriate. Regarding the formatting and figures in your paper, the author will need to pay close attention in the next revision. Thank you for your response.

Replying to Rebuttal by Authors

Mandatory Acknowledgement by Reviewer jy8h

Mandatory Acknowledgement by Reviewer jy8h 08 Aug 2025, 12:43

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Mandatory Acknowledgement: I have read the author rebuttal and considered all raised points., I have engaged in discussions and responded to authors., I have filled in the "Final Justification" text box and updated "Rating" accordingly (before Aug 13) that will become visible to authors once decisions are released., I understand that Area Chairs will be able to flag up Insufficient Reviews during the Reviewer-AC Discussions and shortly after to catch any irresponsible, insufficient or problematic behavior. Area Chairs will be also able to flag up during Metareview grossly irresponsible reviewers (including but not limited to possibly LLM-generated reviews)., I understand my Review and my conduct are subject to Responsible Reviewing initiative, including the desk rejection of my co-authored papers for grossly irresponsible behaviors. https://blog.neurips.cc/2025/05/02/responsible-reviewing-initiative-for-neurips-2025/ (https://blog.neurips.cc/2025/05/02/responsible-reviewing-initiative-for-neurips-2025/)

Replying to Reply to the author

Official Comment by Authors

Official Comment

by Authors (Jae Oh Woo (/profile?id=~Jae_Oh_Woo1), Baishali Chaudhury (/profile?id=~Baishali_Chaudhury1), Mengdie Flora Wang (/profile?id=~Mengdie_Flora_Wang1), Rahul Ghosh (/profile?id=~Rahul_Ghosh1), +1 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission6489/Authors))

08 Aug 2025, 17:28

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

Thank you sincerely for your thoughtful reconsideration and for increasing the score. We appreciate your clear guidance on presentation quality and will give it careful attention in the next revision—improving figure clarity and captions, ensuring consistent notation, and enhancing overall readability so the claims and evidence are communicated cleanly. Your constructive feedback has been genuinely helpful, and we would be glad to incorporate any additional suggestions you may have.

Official Review of Submission6489 by Reviewer R5yk

Official Review by Reviewer R5yk 02 Jul 2025, 10:03 (modified: 18 Sept 2025, 08:42)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer R5yk

Revisions (/revisions?id=3SGygJI8kw)

Summary:

This paper introduces a novel framework for self-evaluation of the reasoning abilities of LLMs, which is an important research direction. The authors propose to do it through a probabilistic extension (PVC) of the Vapnik-Chervonenkis (VC) dimension and also introduce a calibration-aware extension (C-PVC). The authors developed a new test set consisting of 360 unique math problems that cover different topics and levels of difficulty, designed to measure how reliably models can judge the correctness of their own reasoning. They tested several modern models with 7 to 8 billion parameters on this set, measuring not just how accurate their answers were, but also how well their confidence scores matched actual performance (calibration) and how much data is theoretically needed for them to generalize.

Strengths And Weaknesses:

Strengths

The motivation is clear; the theoretical reasoning behind using an extension of classical VC dimension theory is well-grounded and thoughtfully applied.

The method builds on existing VC dimension theory and calibration research, but its application to LLM self-reflection is timely and novel.

The paper is generally well-structured, with thorough explanations of both theoretical and experimental aspects.

Sufficient background is provided for readers unfamiliar with VC dimension theory.

The authors introduce a new benchmark specifically designed to evaluate reasoning self-assessment, complementing existing evaluation methods.

Weaknesses

1. The related work section is insufficient. While it briefly overviews calibration techniques and uncertainty measures and mentions they have been adapted to LLMs, it lacks citations supporting this adaptation to LLMs. PAC-Bayes theory is mentioned in line 150 but there is no discussion of this in the related work. It would have been useful to have had some background on other statistical methods that have been used for this purpose.
2. Some assumptions and parameters (e.g., choice of thresholds like γ, τ, δ in Table 1) appear somewhat heuristic and lack justification. Table 1's caption states:

   > "We assume $C = 1$, $\gamma = 0.6$, $\tau = 0.25$, and $\delta = 0.05$ for simplicity."

An ablation or sensitivity analysis on these parameters would strengthen confidence that the framework and findings are not overly dependent on somewhat arbitrary parameter choices.

3. The paper’s core theoretical framework introduces sample complexity bounds intended to predict model generalization and reasoning reliability. However, the empirical results show discrepancies between these theoretical estimates and actual model performance, with no analysis to explain or reconcile this gap.

In Table 1's caption, it is stated:

> *“Sample complexity represents the theoretical minimum data required for effective generalization.”*

And line 258 states:

> "While OpenThinker2-7B requires the highest number of samples (144), it achieves the lowest actual error (0.351), indicating strong reasoning performance but less reliable confidence estimation."

Even though OpenThinker2-7B has the highest theoretical sample complexity (meaning it should need more data to generalize well), it achieves the lowest empirical error among models tested. So, empirically it performs very well, despite the theory suggesting it needs more data. This highlights a limitation: while the PVC/C-PVC theory is elegant and provides guarantees, it may not fully explain or predict real model behavior.

4. The paper does not include a clear discussion addressing its own limitations.
5. In line 265, the authors write:

> “Models trained via supervised fine-tuning (SFT) — such as OpenThinker2, Bespoke-Stratos, JiuZhang3.0, Open-Reasoner-Zero, and s1.1 — tend to exhibit diverse behaviors.”

And in line 271:

> “Reinforcement learning–optimized models such as Qwen2.5 variants display higher PVCs but poor C-PVC, indicating a tendency toward overconfident reasoning.”

These statements describe observed correlations between training regimes (SFT, RL, distillation) and metrics like PVC and calibration (C-PVC). While the paper finds that certain training regimes tend to correspond with particular patterns of calibration and reasoning quality, it remains unclear whether the training methods are truly the causal factor or if other confounding variables play a role (model architecture differences, data, hyperparameters, etc.). Correlations do not imply causation.

The authors state their findings as trends or hypotheses, but do not provide experimental evidence from controlled interventions or additional analysis to confirm causation. This limits the strength of the conclusions about the impact of training methods.

Minor: Some sections of the paper appear to be written or heavily rephrased using LLMs. While this is not inherently problematic, it can detract from the natural flow and originality of the reading experience.
Quality: 2: fair
Clarity: 2: fair
Significance: 3: good
Originality: 3: good

Questions:

1. Could you clarify the rationale behind the chosen thresholds (γ, τ, δ)? I could not find it in the paper; if it is mentioned, please point me to it.
2. How do you explain the discrepancy between theoretical sample complexity estimates and actual empirical model performance? (see Weakness 3)
3. Could you elaborate on potential confounding factors in the observed correlations between training regimes (SFT, RL) and calibration metrics? (see Weakness 5)
4. Are there any limitations or failure cases of the PVC and C-PVC framework that you observed?

Limitations:
No, limitations have not been addressed. I recommend explicitly discussing parameter choices and potential mismatch between theory and empirical results.

Rating: 3: Borderline reject: Technically solid paper where reasons to reject, e.g., limited evaluation, outweigh reasons to accept, e.g., good evaluation. Please use sparingly.
Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.
Ethical Concerns: NO or VERY MINOR ethics concerns only
Paper Formatting Concerns:
No formatting issues.

Code Of Conduct Acknowledgement: Yes
Responsible Reviewing Acknowledgement: Yes
Final Justification:
The authors have responded to most concerns raised and have promised to make improvements in the camera-ready version (including a limitations section and a guide on threshold selection in the Appendix). However, the question of why certain hypermeters were chosen specifically in the paper still remain unaddressed. After careful thought, and further reading the comments by other reviewers, I have decided to go back to my original score. I believe the paper has good potential, but in its current form, there are several ambiguities in the writing that can be improved in a proper revision.

−
=
≡

Rebuttal by Authors

Rebuttal

by Authors (Jae Oh Woo (/profile?id=~Jae_Oh_Woo1), Baishali Chaudhury (/profile?id=~Baishali_Chaudhury1), Mengdie Flora Wang (/profile?id=~Mengdie_Flora_Wang1), Rahul Ghosh (/profile?id=~Rahul_Ghosh1), +1 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission6489/Authors))

29 Jul 2025, 23:32 (modified: 31 Jul 2025, 11:56)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=JWTmcJZZUG)

Rebuttal:
We sincerely thank **Reviewer R5yk** for the comprehensive, insightful, and constructive evaluation of our work. The review has pinpointed valuable areas for clarification and improvement.

Issue 1 - On the Related Work and PAC-Bayes Theory

We appreciate the reviewer’s suggestion to clarify the theoretical positioning of our work with respect to PAC-Bayes theory. While our current framework is grounded in statistical learning theory via probabilistic extensions of VC dimension, we agree that contrasting it with PAC-Bayes formulations would enrich the context and improve theoretical transparency.

PAC-Bayes theory offers generalization guarantees by averaging over posterior distributions on hypothesis spaces and has been influential in analyzing stochastic learning systems. However, such bounds typically require complex assumptions about prior/posterior distributions and can be computationally intensive to evaluate in practice. More importantly, PAC-Bayes is most naturally suited to stochastic or Bayesian models that involve posterior distributions over hypotheses, whereas our setting centers on single-point predictions made by large language models, where confidence and calibration are derived from fixed model parameters rather than sampled posteriors.

We will revise the manuscript to include a brief comparison with PAC-Bayes theory, highlighting both methodological differences and potential complementarities. Representative citations will be added to guide interested readers toward deeper connections between these theoretical paradigms.

Issue 2 - On the Parameter Sensitivity and Threshold Choices $(\gamma, \tau, C, \delta, \epsilon)$

Issue 2 - On the Parameter Sensitivity and Threshold Choices

We appreciate the reviewer's request for further clarification on parameter selection, particularly for γ and τ. As shown in Proposition 1, PVC increases monotonically as γ decreases, and C-PVC converges to PVC as τ increases. This monotonicity creates a challenge: **we could trivially maximize scores by selecting extreme threshold values, but doing so would severely reduce discriminative power between models**.

To address this, we developed a **contrast score** that maximizes differentiation between models. This scoring function incorporates three key factors:

- **Spatial uniformity**: ensuring models are well-distributed across the measurement space
- **Non-duplication**: minimizing redundant clustered measurements
- **Grid balance**: maintaining interpretable spacing between evaluation points

We conducted a systematic sensitivity analysis across various γ and τ values, with results summarized in the table below:

γ	τ	**contrast_score**
0.60	0.25	0.83 (our choice)

γ	τ	**contrast_score**
0.59	0.25	0.83
0.60	0.27	0.81
0.59	0.27	0.81
0.59	0.24	0.80
0.60	0.24	0.80
0.59	0.26	0.80
0.60	0.26	0.80
0.58	0.24	0.76
0.58	0.27	0.76

Our analysis shows that threshold values converge to a range ($\gamma \approx 0.59 - 0.60$, $\tau \approx 0.24 - 0.27$) where model differentiation is maximized (contrast_score = $0.80 - 0.83$). Moving outside this range (e.g., $\gamma = 0.58$) causes a noticeable drop in discriminative power (contrast_score = 0.76). We will provide the complete details of this optimization process in the revised Appendix.

For C, δ, and ϵ in our sample complexity calculations, our selections are directly grounded in Theorem 1. The constant C governs sample complexity scaling, while δ and ϵ reflect generalization confidence and tolerance bounds. We set $\epsilon = \tau$ to align theoretical tolerance with empirical calibration margin. While C and δ could in principle be model-dependent, we intentionally fix $C = 1$ and $\delta = 0.05$ across all models to isolate model-specific effects in PVC/C-PVC measurements, ensuring theoretical comparisons remain fair and interpretable.

Issue 3 - On the Discrepancy between Theoretical Sample Complexity and Empirical Error

We appreciate the reviewer’s insightful question on the apparent discrepancy between theoretical sample complexity and empirical performance. To address this clearly, we must carefully distinguish **sample complexity** (the minimal number of samples theoretically sufficient for generalization, **given a model’s hypothesis space**) from **empirical generalization error** (the model's actual performance on realistic data).

- Theoretical sample complexity, derived via PVC and C-PVC, explicitly quantifies the minimum number of samples required to guarantee generalization across the entire hypothesis space, under the worst-case assumption. Each model defines its own unique hypothesis class; thus, differences in sample complexity reflect differences in structural complexity and expressibility of these hypothesis spaces.
- The observation that OpenThinker2-7B empirically achieves low error despite high theoretical sample complexity reflects the fact that its hypothesis space is richer and more expressive. A larger and more complex hypothesis space naturally entails higher sample complexity in worst-case theoretical analysis.

However, in practice, real-world data distributions are far from adversarial and often exhibit structure that can be effectively exploited. So, OpenThinker2-7B leverages beneficial inductive biases from its pretraining and architecture, achieving lower empirical error than the theoretical bound would suggest.

To clarify through contrast, consider linear regression:

- Linear regression possesses low theoretical sample complexity due to its small VC dimension. This low complexity arises from its inherently simple hypothesis class (linear functions).
- However, low complexity also severely restricts expressibility: linear regression fundamentally cannot represent complex or nonlinear functions adequately. Therefore, even if the theoretical sample complexity requirement is minimal, linear regression may exhibit large empirical errors when confronted with structurally complex, nonlinear data.
- Thus, interpreting low theoretical complexity as implying universally good generalization performance is not true in general.

So, **sample complexity and empirical error measure fundamentally distinct aspects**:

- Sample complexity measures structural simplicity of the hypothesis space and minimal data sufficiency under worst-case conditions.
- Empirical error reflects how effectively a given model, leveraging its inductive biases, approximates real-world data distributions.

Therefore, **the observed discrepancy is not a flaw or inconsistency but rather a natural consequence of interpreting generalization bounds correctly.** We will explicitly clarify this important distinction in the revised manuscript to help prevent misinterpretation of generalization bounds as direct predictors of empirical performance.

Issue 4 - On the Causal Interpretations of Training Regimes (SFT, RL, Distillation)

We appreciate the reviewer raising this important point regarding causal interpretation. Establishing causality requires carefully controlled experiments that isolate individual factors. As recognized, our paper does not make any claims of causal effect. Rather, we present descriptive correlations between training paradigms and the observed introspective behaviors as captured by PVC and C-PVC.

That said, **we acknowledge the importance of transparency in how these correlations are presented and interpreted.** In the revised version, we will:

- Explicitly state that all observed correlations are hypothesis-generating and not indicative of causal relationships;
- Clarify that our findings should be understood as empirical signals, meant to motivate more targeted and controlled future investigations;
- Enumerate specific potential confounding factors that must be disentangled in any future causal analysis.

We take full responsibility for making these correlations explicit, with the hope that they will inform and inspire future research aimed at testing, refining, and extending this line of inquiry.

Issue 5 - On the Explicit Limitations Discussion

We fully agree that a clear and well-structured articulation of limitations is essential for responsible scholarship. While Section 7 outlines several important limitations—including metric granularity, reliance on ensemble judges, benchmark domain scope, and sensitivity to threshold parameters—we recognize that these points would benefit from being consolidated and made more prominent in a dedicated "Limitations" section. We will implement this structural improvement in the revised manuscript to enhance transparency and accessibility.

In direct response to the reviewer's question regarding **observed limitations or failure cases of the PVC and C-PVC framework**, we emphasize the following:

- **Sampling and Resource Constraints**: Estimating introspective capacity via PVC and C-PVC requires sufficient statistical density—both across task categories and within-category instances—to ensure stable and interpretable measurements. The practical challenge of meeting this requirement, due to constraints on both high-quality problem pools and computational resources, emerged as one of the most significant limitations observed during this work. As we noted in the paper, the present study captures only a narrow slice of the broader vision for this framework. Our empirical findings, while promising, remain an early step. We hope this initial effort can serve as a foundation for more comprehensive evaluations at scale—potentially involving community-driven initiatives to extend the benchmark coverage and resolution needed for robust capacity diagnostics.

Replying to Rebuttal by Authors

Mandatory Acknowledgement by Reviewer R5yk

Mandatory Acknowledgement by Reviewer R5yk 03 Aug 2025, 08:31

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Mandatory Acknowledgement: I have read the author rebuttal and considered all raised points., I have engaged in discussions and responded to authors., I have filled in the "Final Justification" text box and updated "Rating" accordingly (before Aug 13) that will become visible to authors once decisions are released., I understand that Area Chairs will be able to flag up Insufficient Reviews during the Reviewer-AC Discussions and shortly after to catch any irresponsible, insufficient or problematic behavior. Area Chairs will be also able to flag up during Metareview grossly irresponsible reviewers (including but not limited to possibly LLM-generated reviews)., I understand my Review and my conduct are subject to Responsible Reviewing initiative, including the desk rejection of my co-authored papers for grossly irresponsible behaviors. https://blog.neurips.cc/2025/05/02/responsible-reviewing-initiative-for-neurips-2025/ (https://blog.neurips.cc/2025/05/02/responsible-reviewing-initiative-for-neurips-2025/)

Replying to Rebuttal by Authors

Official Comment by Reviewer R5yk

Official Comment by Reviewer R5yk 03 Aug 2025, 09:03

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

Thank you to the authors for the thoughtful and detailed responses. I appreciate the clarifications for all concerns raised.

However, regarding the threshold selection, I remain unconvinced by the current justification. The proposed contrast score appears insightful and could certainly help guide threshold choices, but it does not fully explain or justify the hyperparameter values that were ultimately selected in the paper. Additionally, it would significantly improve clarity and reproducibility to include an explicit formulation or pseudocode for how the contrast score is actually computed in the Appendix. This would allow readers to fully understand and reproduce the selection mechanism.

I also agree with Reviewer jy8h's comment regarding the clarity of the figures. The caption for Figure 3 should clearly explain the two plots separately, as they represent distinct aspects of the analysis. Additionally, the font size in the plots is too small to be readable, especially in the current format. While I understand the space constraints in conference papers, improving the readability of these figures needs a bit of attention.

Official Comment by Authors

Official Comment

by Authors (Jae Oh Woo (/profile?id=~Jae_Oh_Woo1), Baishali Chaudhury (/profile?id=~Baishali_Chaudhury1), Mengdie Flora Wang (/profile?id=~Mengdie_Flora_Wang1), Rahul Ghosh (/profile?id=~Rahul_Ghosh1), +1 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission6489/Authors))

03 Aug 2025, 13:29 (modified: 03 Aug 2025, 13:34)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=HKI3i7dko3)

Comment:

We sincerely appreciate your thoughtful feedback, which has directly improved the clarity and transparency of our work.

Regarding threshold selection, we agree that a more detailed explanation is warranted. In the revised version, we will:

- Include pseudocode in the Appendix,
- Provide the full implementation of the `calculate_contrast_score` module (see below),
- Explain how it was used to evaluate candidate thresholds and select the final configuration.

On Figures 2 & 3, we will:

- Separate the subplots with clear labels,
- Revise the caption to describe each plot distinctly,
- Improve visual clarity (e.g., font size, layout).

All relevant code, including threshold scoring and figure generation scripts, will be included in the supplementary materials. As strong advocates for open and reproducible research, we are committed to sharing all key components in full to support transparency and community use.

Below is the actual implementation of the `calculate_contrast_score` function used to systematically score candidate thresholds and guide the final selection:

```python
def calculate_contrast_score(x_coords, y_coords, max_points_per_line=4,
                                   duplicate_threshold=0.005, weight_uniformity=0.33,
                                   weight_grid=0.33, weight_non_duplicate=0.33):
    """
    Calculate a contrast score for points on a grid, considering:
    1. Overall spatial uniformity
    2. Even distribution per grid line
    3. Minimization of duplicate/nearby points

    Args:
        x_coords (array-like): List of x coordinates (pvc)
        y_coords (array-like): List of y coordinates (c-pvc)
        max_points_per_line (int): Maximum allowed points on any x or y line (default 4)
        duplicate_threshold (float): Distance threshold for nearby points
        weight_uniformity, weight_grid, weight_non_duplicate (float): Weights for each component

    Returns:
        float: Total quality score (0-1)
    """
    # Convert to numpy arrays
    x = np.array(x_coords)
    y = np.array(y_coords)
    n_points = len(x)

    if n_points < 2:
        return 0

    # 1. Overall spatial uniformity score
    points = np.column_stack((x, y))
    tree = cKDTree(points)
    distances, _ = tree.query(points, k=2)
    nearest_distances = distances[:, 1]

    cv = np.std(nearest_distances) / np.mean(nearest_distances) if np.mean(nearest_distances) > 0 else float('inf')
    uniformity_score = 1.0 / (1.0 + cv)

    # 2. Score for grid line distribution with maximum constraint
    x_counts = Counter(x)
    y_counts = Counter(y)

    # Check if any line exceeds the maximum allowed points
    x_exceeds = any(count > max_points_per_line for count in x_counts.values())
    y_exceeds = any(count > max_points_per_line for count in y_counts.values())

    # Calculate penalty for exceeding the limit
    if x_exceeds or y_exceeds:
        # Count how many lines exceed the limit and by how much
        x_excess = sum(max(0, count - max_points_per_line) for count in x_counts.values())
        y_excess = sum(max(0, count - max_points_per_line) for count in y_counts.values())
        total_excess = x_excess + y_excess

        # Apply a severe penalty based on the excess
        penalty_factor = min(1.0, total_excess / (n_points * 0.2))  # Normalize penalty
        grid_score = max(0.0, 1.0 - penalty_factor)
    else:
        # If no lines exceed, calculate evenness of distribution
        x_values = list(x_counts.values())
        y_values = list(y_counts.values())

        if len(x_values) == 0 or len(y_values) == 0:
            grid_score = 1.0
        else:
            # Calculate CV for x-line and y-line point counts
            x_cv = np.std(x_values) / np.mean(x_values) if np.mean(x_values) > 0 else floa
t('inf')
            y_cv = np.std(y_values) / np.mean(y_values) if np.mean(y_values) > 0 else floa
t('inf')

            # Convert to scores (lower CV = higher score)
            x_score = 1.0 / (1.0 + x_cv)
            y_score = 1.0 / (1.0 + y_cv)

            # Average the x and y scores
            grid_score = (x_score + y_score) / 2

    # 3. Non-duplication score
    x_range = np.max(x) - np.min(x) if len(x) > 0 else 0
    y_range = np.max(y) - np.min(y) if len(y) > 0 else 0
    domain_scale = max(x_range, y_range, 1)  # Minimum scale of 1 for integer grid

    scaled_threshold = duplicate_threshold * domain_scale
    pairs = tree.query_pairs(scaled_threshold)
    nearby_pairs_count = len(pairs)

    max_pairs = n_points * (n_points - 1) / 2
    nearby_pairs_ratio = nearby_pairs_count / max_pairs if max_pairs > 0 else 0
    non_duplicate_score = 1.0 - nearby_pairs_ratio

    # Calculate weighted total score
    total_score = (weight_uniformity * uniformity_score +
                   weight_grid * grid_score +
                   weight_non_duplicate * non_duplicate_score)

    return total_score
```

Replying to Official Comment by Authors

Official Comment by Reviewer R5yk

Official Comment by Reviewer R5yk 04 Aug 2025, 10:16

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

Thank you, most of my concerns have been addressed and I am happy to adjust my score accordingly.

Replying to Official Comment by Reviewer R5yk

Official Comment by Authors

Official Comment

by Authors (Jae Oh Woo (/profile?id=~Jae_Oh_Woo1), Baishali Chaudhury (/profile?id=~Baishali_Chaudhury1), Mengdie Flora Wang (/profile?id=~Mengdie_Flora_Wang1), Rahul Ghosh (/profile?id=~Rahul_Ghosh1), +1 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission6489/Authors))

04 Aug 2025, 10:27

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

We are truly grateful for your thoughtful engagement and kind follow-up. Thank you so much for taking the time to revisit our responses and for your willingness to re-evaluate your score.

Your constructive feedback has been invaluable in helping us improve both the clarity and rigor of our work, and we deeply appreciate your support throughout this process.

Please do not hesitate to let us know if any additional questions or suggestions arise—we'd be more than happy to incorporate them in the revised version.

Official Review of Submission6489 by Reviewer 15u2

Official Review by Reviewer 15u2 01 Jul 2025, 08:51 (modified: 18 Sept 2025, 08:42)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer 15u2

Revisions (/revisions?id=0axU1PZEL6)

Summary:

This paper attempts to address the high level question of whether LLMs can reflect on their own reasoning. It does so by taking an approach based on some simple generalizations of VC dimension to the probabilistic case, taking into account confidence and calibration. It proves some simple relationships between VC dimension and probabilistic VC dimensions as well as some generalization bounds based on probabilistic VC dimension. It then measures probabilistic VC dimension for a variety of language models, in a very specific self-reflection scenario on a very specific novel dataset of 360 mathematical questions. It finds a diversity of behaviors for how (calibration aware) probabilistic VC dimension estimates and measures of calibration and error covary across 11 models.

Strengths And Weaknesses:

Strengths:

1. The paper makes simple definitions of probabilistic VC dimension and calibration aware probabilistic VC dimension.
2. It proves some simple inequalities relating them.
3. It creates a novel but small math dataset.
4. It makes several measurements of its measures on different models.

Weaknesses:

1. The new definitions are not well motivated from the perspective of probing reasoning and are simple extensions of fat-shattering dimension. Reasoning involves taking multiple deductive steps, backtracking and restarting if necessary - resulting in a complex dynamic process. This paper, through its measure of probabilistic VC dimension,

measures the diversity of next-token distributions subject to confidence constraints generated by an LLM over a diversity of prompts, in a specific self-reflection scenario. It is not at all clear how these measures relate to the quality of a dynamic reasoning process.

2. The new math benchmark is small, but potentially useful. However, nothing is said about the benchmark in detail.
3. The paper is not very well written. For most of the entire paper it was not clear why definitions were being made in the context of analyzing reasoning.
4. Many numerical experiments were done in Table 1 but the probabilistic VC dimension measures and the error and calibration outcomes were all over the place. There are no clear take home messages or clear conceptual insights that came out of these experiments. The paper tries to construct a story of how the varied outcomes correlate with training procedures, but there is no logic for why they see the patterns of covariation they do. Overall it seems like an attempt to construct a story posthoc from a series of ambiguous experiments, using measures that are not well designed to address the very question they seek to ask. The strongest statement I can extract is: "models with high probabilistic VC dimension often suffer from poor calibration, while better-calibrated models tend to be more conservative." If this is the main take home message of this paper, it does not rise in my mind to the level of a NeurIPS paper. The phrases "often" and "tend to" indicate they have not found strong regularities.

Quality: 2: fair
Clarity: 1: poor
Significance: 1: poor
Originality: 2: fair
Questions:
I would appreciate a clear set of take home messages and conceptual insights about reasoning that I am supposed to take away from this paper, stated in a manner that is independent of math.

Moreover, I would like a clear logic for why (if at all) the measures you are proposing address the fidelity of the complex dynamic process of reasoning.

Limitations:
Not at all. They need an entire section devoted to what their measures do not capture about reasoning.

Rating: 3: Borderline reject: Technically solid paper where reasons to reject, e.g., limited evaluation, outweigh reasons to accept, e.g., good evaluation. Please use sparingly.
Confidence: 3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.
Ethical Concerns: NO or VERY MINOR ethics concerns only
Paper Formatting Concerns:
None.

Code Of Conduct Acknowledgement: Yes
Responsible Reviewing Acknowledgement: Yes
Final Justification:
I feel this paper oversells its results in the introduction, and does not have clear take home messages, and could benefit from significant revisions in writing for clarity.

Rebuttal by Authors

Rebuttal

by Authors (👁 Jae Oh Woo (/profile?id=~Jae_Oh_Woo1), Baishali Chaudhury (/profile?id=~Baishali_Chaudhury1), Mengdie Flora Wang (/profile?id=~Mengdie_Flora_Wang1), Rahul Ghosh (/profile?id=~Rahul_Ghosh1), +1 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission6489/Authors))

📅 29 Jul 2025, 02:24 (modified: 31 Jul 2025, 11:56)

👁 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

📄 Revisions (/revisions?id=b0R64Q4sW2)

Rebuttal:

We respectfully but firmly disagree with several central assessments in this review, which appear to rest on a mischaracterization of the paper's goals and contributions. The review evaluates our work against criteria we explicitly do not claim to satisfy—namely, the modeling of multi-step, generative reasoning chains—while overlooking the primary objective stated repeatedly throughout the paper: to rigorously assess whether language models can recognize when their own reasoning is correct.

Our work introduces a novel and operational formulation of this question, grounded in statistical learning theory. We ask: **Can a model distinguish between correct and incorrect solutions with calibrated confidence?** This formulation leads to Probabilistic VC (PVC) and Calibration-aware PVC (C-PVC)—capacity measures that quantify introspective reliability, not expressive power alone. These are theoretically sound, empirically computable, and practically meaningful for alignment and trustworthiness.

We now address the key issues raised:

Issue1 - On the Motivation and Novelty of PVC and C-PVC

The reviewer claims our definitions are "simple extensions" of fat-shattering dimension and are "not well motivated." This fundamentally misrepresents our contribution. The novelty lies not in isolated mathematical tweaks, but in **reframing self-evaluation as a probabilistic binary classification task**—a conceptual shift that enables rigorous generalization analysis of confidence-aware correctness recognition. PVC captures the model's ability to separate correct and incorrect outputs under confidence constraints, while C-PVC measures this ability under calibration constraints—both essential components of introspective reliability. These definitions are not superficial extensions. They are carefully constructed and theoretically grounded.

Issue 2 - On the Relationship to Dynamic Reasoning

The reviewer critiques our framework for not modeling multi-step deductive reasoning. But this is precisely not our claim. Our framework does not attempt to simulate dynamic reasoning trajectories. Instead, it evaluates the outcome of reasoning: can a model assess the correctness of what it has produced?

This is a fundamentally different question, and one with **immediate practical implications**. A model that can reason correctly but cannot recognize when it is wrong is unsafe. Our framework measures this critical capacity with mathematical precision. Dismissing it for not doing something else entirely—namely, modeling the full internal reasoning process—is a categorical error in review scope.

Issue 3 - On the Benchmark Design and Description

The reviewer refers to our math benchmark as “small” and “not described in detail.” This is inaccurate. The benchmark consists of 360 carefully curated problems across arithmetic, algebra, geometry, and logic. We describe its construction methodology, difficulty balancing, and domain coverage in Section 5 and the Appendix I.3.

More importantly, the benchmark is not the central contribution—it is an instrument for evaluating the proposed capacity measures. We also observe similar behavior in Math-500 benchmark in Appendix G. What matters is not volume, but the signal quality it enables—and our empirical results consistently differentiate models along interpretable axes of calibration and capacity.

Issue 4 - On the Interpretation of Experimental Results

We respectfully disagree with the reviewer’s characterization that the paper “lack take-home messages.” In fact, the reviewer themselves highlights one of the paper’s central insights: **models with high probabilistic VC dimension often suffer from poor calibration, while better-calibrated models tend to be more conservative.** This succinctly captures one of our main contributions.

We also disagree with the claim that our findings are post-hoc or based on isolated observations. We present a cohesive narrative supported by both theoretical and empirical results (See additional experiments in **Issue 2** in the **Reviewer jy8h**):

- **Consistent calibration–capacity tradeoff across benchmarks**: Across diverse tasks—including arithmetic (Math Benchmark), high-school-level math (Math-500), factual reasoning (TruthfulQA), and commonsense inference (CommonsenseQA)—we observe a recurring pattern: models with high PVC scores tend to be poorly calibrated, while better-calibrated models often exhibit conservative confidence. This calibration–capacity tradeoff is not task-specific but reflects a fundamental dimension of LLM behavior.
- **Model-type differentiation with introspective metrics**: Our framework enables comparison across RL-finetuned, SFT, and distilled models on a new axis: introspective reliability. The results reveal structural differences—e.g., RL models exhibit higher expressiveness but poorer confidence calibration—providing insights into the consequences of current training paradigms.
- **Bridging theory and practice**: Our results are not anecdotal. They are supported by formal generalization bounds derived from PVC/C-PVC theory, and their empirical trends match theoretical expectations. This duality reinforces the paper’s central message: that introspective competence is measurable, meaningful, and model-dependent.

These findings together provide a clear takeaway: **confidence without calibration is not introspection.** Our paper introduces a novel lens—PVC and C-PVC—for evaluating model introspection, and our results consistently reinforce this perspective across diverse settings.

> **Issue 5** - On the Discussion of Limitations

The reviewer claims the paper "says nothing" about limitations and demands a "dedicated section." We respectfully disagree.

First, Section 7 clearly outlines limitations regarding metric granularity, reliance on judgment sources (e.g., ensemble judges), benchmark scope, and extensibility to broader reasoning domains. These are candid and specific.

Second, we recognize that a dedicated "Limitations" section can improve clarity and accessibility for readers. We appreciate the reviewer's suggestion and will add a standalone limitations section in the final version to make these points more explicit and easier to reference.

We respectfully ask that **Reviewer 15u2** consider our work through the lens of the actual research question we aim to address. While we value the reviewer's perspective on dynamic reasoning, we kindly suggest that our contribution be evaluated on its intended scope: a principled approach to evaluating introspective reliability, rather than expressive reasoning capacity.

We are grateful for your time and consideration, and we remain open to any further clarification that might help align expectations with the paper's actual scope and goals.

−
=
≡

Author AC Confidential Comment by Authors

Author AC Confidential Comment

by Authors (👁 Jae Oh Woo (/profile?id=~Jae_Oh_Woo1), Baishali Chaudhury (/profile?id=~Baishali_Chaudhury1), Mengdie Flora Wang (/profile?id=~Mengdie_Flora_Wang1), Rahul Ghosh (/profile?id=~Rahul_Ghosh1), +1 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission6489/Authors))

📅 29 Jul 2025, 02:44 (modified: 30 Jul 2025, 17:43)

👁 Program Chairs, Senior Area Chairs, Area Chairs, Authors 📄 Revisions (/revisions?id=BUlE7V99lr)

Comment:

We write to express our serious concern about **Reviewer 15u2**'s evaluation, which we believe is not only misaligned with the paper's clearly defined scope, but also internally inconsistent, technically unengaged, and out of step with the consensus of the other reviewers. Given these issues, we strongly urge that this review be discounted in the final decision-making process.

At the heart of this review is a categorical error: it faults the paper for not modeling dynamic, multi-step reasoning trajectories, despite the fact that the paper **explicitly and repeatedly states that it does not attempt to do so**. Instead, our contribution is to formalize a different, but equally critical question: *can a model recognize when its reasoning is correct?* This is a well-scoped, operationally meaningful, and theoretically grounded problem—central to trustworthy deployment of language models—and our work proposes the first capacity-theoretic framework for addressing it.

Despite this, the reviewer evaluates the paper as if it had promised to model human-like reasoning chains and failed to deliver, thereby assessing it against criteria the paper does not claim to meet. This is not a difference of opinion—it is a **fundamental misreading of the paper's objective**, and allowing it to stand risks setting a precedent where papers are penalized for not solving unrelated problems.

Even more troubling is the review offers **no technical counterarguments**. There is **no engagement** with our generalization bounds, no critique of our sample complexity derivations, no discussion of the empirical methodology, and no alternative explanation for the empirical trends reported. Instead, the review is based almost entirely on personal intuition about what constitutes "reasoning"—a subjective standard that lies outside the scope of this submission and outside the standards of objective peer review.

Furthermore there is internal inconsistency throughout the review. The reviewer simultaneously dismisses our PVC and C-PVC definitions as "simple extensions" and then claims the resulting empirical outcomes are "all over the place." This contradiction undermines the credibility of the critique: if the definitions were indeed trivial or uninformative, they would not produce the nuanced, model-dependent outcomes that we observe—and that the other reviewers recognized as valuable.

In contrast, all three other reviewers (**Gx9q**, **jy8h**, **R5yk**) independently acknowledged the theoretical rigor, practical relevance, and empirical clarity of our work. They engaged with the paper's actual content and provided thoughtful, constructive feedback. This pattern suggests that the review by **15u2** is a clear outlier—one that not only misrepresents the paper, but does so in a way that is inconsistent with the values of fair and informed evaluation.

Given these facts, we believe it would be **deeply inappropriate** for this single, anomalous review—one that lacks both technical substance and interpretive fidelity—to exert decisive influence on the outcome. Doing so would not only be unjust to this submission, but damaging to the integrity of the review process more broadly.

We recognize that the discussion period is still ongoing, and we are committed to fully participating in that process in good faith. We will continue to monitor the dialogue and incorporate any clarifications, updates, or reviewer interactions into a final version of this author comment before the end of the rebuttal/discussion deadline.

Nevertheless, we respectfully request that the AC and SAC evaluate this review in light of its technical shortcomings and divergence from the paper's clearly stated scope, and ensure that the final decision reflects a fair and informed assessment of the submission's actual contributions.

Replying to Rebuttal by Authors

Response to rebuttal

Official Comment by Reviewer 15u2 08 Aug 2025, 08:43

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

Thank you for your response. I still believe the paper oversells its claims in the title, abstract and introduction. Let me quote parts of each:

Title: "Can LLMs reflect on their reasoning."

Abstract: "to what extent can they evaluate the quality of their own reasoning? This work develops a unified theoretical and empirical framework to investigate that ability."

Introduction:

"A key requirement for autonomous reasoning systems is not only the ability to solve complex tasks, but also the ability to evaluate the reliability of their own reasoning *processes*."

These are very highly level ambiguous sentences that suggest a paper which will be able to assess whether every individual step of a reasoning process is correct or not. Instead, the paper does something else, which is evaluate whether the final answer or entire response's confidence is well calibrated or not. There is a mismatch between what the paper advertises and what the paper does. Finally even within the context of what the paper achieves, I do not find the take home messages all that insightful or compelling. Of course confidence means nothing if it is not calibrated. I feel this paper would need to be rewritten for clarity to make it much more understandable and not to oversell the results in the title, abstract and intro. The paper could benefit from a succinct and intuitive summary of what it actually achieves in normal english language there. I suspect this could be possible, but it would be hard to review without seeing a new draft.

Despite all this, I will raise my already low score, to give the readers the benefit of the doubt. But I still feel that for the paper to be valuable to the community it could benefit from an entire round of revision, and search for crisper take home messages, suitable for submission to the next conference.

Replying to Rebuttal by Authors

Mandatory Acknowledgement by Reviewer 15u2

Mandatory Acknowledgement by Reviewer 15u2 08 Aug 2025, 09:11

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Mandatory Acknowledgement: I have read the author rebuttal and considered all raised points., I have engaged in discussions and responded to authors., I have filled in the "Final Justification" text box and updated "Rating" accordingly (before Aug 13) that will become visible to authors once decisions are released., I understand that Area Chairs will be able to flag up Insufficient Reviews during the Reviewer-AC Discussions and shortly after to catch any irresponsible, insufficient or problematic behavior. Area Chairs will be also able to flag up during Metareview grossly irresponsible reviewers (including but not limited to possibly LLM-generated reviews)., I understand my Review and my conduct are subject to Responsible Reviewing initiative, including the desk rejection of my co-authored papers for grossly irresponsible behaviors. https://blog.neurips.cc/2025/05/02/responsible-reviewing-initiative-for-neurips-2025/ (https://blog.neurips.cc/2025/05/02/responsible-reviewing-initiative-for-neurips-2025/)

Replying to Response to rebuttal

Official Comment by Authors

Official Comment

by Authors (Jae Oh Woo (/profile?id=~Jae_Oh_Woo1), Baishali Chaudhury (/profile?id=~Baishali_Chaudhury1), Mengdie Flora Wang (/profile?id=~Mengdie_Flora_Wang1), Rahul Ghosh (/profile?id=~Rahul_Ghosh1), +1 more (/group/edit?id=NeurIPS.cc/2025/Conference/Submission6489/Authors))

08 Aug 2025, 17:26

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

We sincerely appreciate your thoughtful reconsideration and the updated score. We acknowledge the concern about overselling and will revise to communicate the scope plainly and conservatively. Our work does not assess step-wise reasoning; it evaluates a focused, solution-level introspection task—given two model-generated solutions to the same problem, can the model select the more likely correct one and report calibrated confidence? We will tighten the language, align claims strictly with the evidence, and strengthen presentation and reporting to ensure clarity, transparency, and reproducibility.