

- Manish Reddy Vuyyuru, Chong Jiang, Joana Ijazi, Kazuki Osawa, Celine Smith, Ramya Sree Boppana, Taylan Bilal, Yuma Koizumi, Ying Xu, Yasemin Altun, Nir Shabat, Ben Bariach, Alex Korchemniy, Kiam Choo, Olaf Ronneberger, Chimezie Iwuanyanwu, Shubin Zhao, David Soergel, Cho-Jui Hsieh, Irene Cai, Shariq Iqbal, Martin Sundermeyer, Zhe Chen, Elie Bursztein, Chaitanya Malaviya, Fadi Biadisy, Prakash Shroff, Inderjit Dhillon, Tejas Latkar, Chris Dyer, Hannah Forbes, Massimo Nicosia, Vitaly Nikolaev, Somer Greene, Marin Georgiev, Pidong Wang, Nina Martin, Hanie Sedghi, John Zhang, Praseem Banzal, Doug Fritz, Vikram Rao, Xuezhi Wang, Jiageng Zhang, Viorica Patraucean, Dayou Du, Igor Mordatch, Ivan Jurin, Lewis Liu, Ayush Dubey, Abhi Mohan, Janek Nowakowski, Vlad-Doru Ion, Nan Wei, Reiko Tojo, Maria Abi Raad, Drew A. Hudson, Vaishakh Keshava, Shubham Agrawal, Kevin Ramirez, Zhichun Wu, Hoang Nguyen, Ji Liu, Madhavi Sewak, Bryce Petrini, DongHyun Choi, Ivan Philips, Ziyue Wang, Ioana Bica, Ankush Garg, Jarek Wilkiewicz, Priyanka Agrawal, Xiaowei Li, Danhao Guo, Emily Xue, Naseer Shaik, Andrew Leach, Sadh MNM Khan, Julia Wiesinger, Sammy Jerome, Abhishek Chakladar, Alek Wenjiao Wang, Tina Ornduff, Folake Abu, Alireza Ghaffarkhah, Marcus Wainwright, Mario Cortes, Frederick Liu, Joshua Maynez, Andreas Terzis, Pouya Samangouei, Riham Mansour, Tomasz Kępa, François-Xavier Aubet, Anton Algymr, Dan Banica, Agoston Weisz, Andras Orban, Alexandre Senges, Ewa Andrejczuk, Mark Geller, Niccolo Dal Santo, Valentin Anklin, Majd Al Merey, Martin Baeuml, Trevor Strohmaier, Junwen Bai, Slav Petrov, Yonghui Wu, Demis Hassabis, Koray Kavukcuoglu, Jeff Dean, and Oriol Vinyals. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024.
- [27] Albert S. Yue, Lovish Madaan, Ted Moskowitz, DJ Strouse, and Aaditya K. Singh. Harp: A challenging human-annotated math reasoning benchmark, 2024.
- [28] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.
- [29] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023.
- [30] Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. Inside: LLMs’ internal states retain the power of hallucination detection, 2024.
- [31] Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. Llm-check: Investigating detection of hallucinations in large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 34188–34216. Curran Associates, Inc., 2024.
- [32] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- [33] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, page 785–794. ACM, August 2016.
- [34] Yury Gorishniy, Akim Kotelnikov, and Artem Babenko. TabM: Advancing Tabular Deep Learning With Parameter-Efficient Ensembling. In *ICLR*, 2025.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [\[Yes\]](#)

Justification: The abstract and introduction clearly articulate the scope and core contributions of the paper, accurately reflecting the proposed framework and findings.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The limitations are explicitly discussed in the Appendix A, along with directions for future work.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All theoretical assumptions are clearly stated in Section 3, and their validity is examined in Section 5 and supported by empirical tables in Appendix B.

Guidelines:

- The answer NA means that the paper does not include theoretical results.

- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: Detailed descriptions of the prompts, the mathematical dataset, and the experimental procedures are provided in Section 2 and 3

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [\[Yes\]](#)

Justification: Both the dataset and source code are included in Supplementary Material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Training and evaluation details including data preprocessing, hyperparameter configurations, and optimization choices are described in Section 2 and Appendix B.2

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The statistical significance of the experiment is presented in Section 4.3, where appropriate metrics are reported. Additional statistical analysis and supporting details are provided in the Appendix C.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Implementation details including computing infrastructure, and memory specifications are reported in the Appendix B. Furthermore, the average runtime required to compute features for each of the six methods is summarized in Table 8, offering a clear comparison of computational costs across methods.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The research adheres to the NeurIPS Code of Ethics by employing publicly accessible, non-personal mathematical datasets, excluding human subjects or surveillance data, and ensuring transparency and reproducibility through a structured code release process.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses potential positive and negative societal impacts in both the section 1 (Introduction) and section 7 (Conclusion).

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper does not introduce any new language models or datasets that could pose a risk of misuse. It uses only existing publicly available reasoning models and math datasets to evaluate a creativity detection method.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper uses publicly available datasets (CreativeMath, HARP) and existing open-source reasoning LLMs.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

1001 13. New assets

1002 Question: Are new assets introduced in the paper well documented and is the documentation
1003 provided alongside the assets?

1004 Answer: [NA]

1005 Justification: The paper does not introduce any new assets such as datasets, models. It
1006 proposes a detection method evaluated using existing public datasets and pretrained models.

1007 Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

1016 14. Crowdsourcing and research with human subjects

1017 Question: For crowdsourcing experiments and research with human subjects, does the paper
1018 include the full text of instructions given to participants and screenshots, if applicable, as
1019 well as details about compensation (if any)?

1020 Answer: [No]

1021 Justification: The paper does not involve crowdsourcing or human subject research, as all
1022 evaluations are conducted using LLM-based assessments.

1023 Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

1032 15. Institutional review board (IRB) approvals or equivalent for research with human 1033 subjects

1034 Question: Does the paper describe potential risks incurred by study participants, whether
1035 such risks were disclosed to the subjects, and whether Institutional Review Board (IRB)
1036 approvals (or an equivalent approval/review based on the requirements of your country or
1037 institution) were obtained?

1038 Answer: [NA]

1039 Justification: The paper does not involve any human subjects or participant-based studies,
1040 therefore IRB approval is not applicable.

1041 Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The paper makes essential and original use of LLMs both as solution generators and evaluators, and explicitly describes their usage in Section 2.2.1

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

A Limitations

We have successfully experimented with many RLMS, but have not been able to experiment with general LLMs of similar size because they do not have sufficient creative solution-generating capabilities. Due to the nature of the White-Box approach, using large-sized models(over 20B) with sufficient performance requires a lot of resources. Lastly, we defined ‘Creativity’ based on solving mathematical problems, and expanding it to a various tasks will be our future work.

B Implementation Details

B.1 LLM Evaluator Details

We use an LLM-based evaluator E to classify each generated response R into three categories — Hallucinated Solution, Typical Solution, or Creative Solution — following the evaluation protocol introduced in [11]. As described in Section 2.2.2 and Figure 1, the evaluation process consists of two stages. First, we assess whether the generated response R is mathematically correct. For example, As shown in Figure 5, the evaluator is given two reference solutions and asked to determine whether R is a valid solution to the given problem.

If both evaluators agree that R is correct, the response proceeds to the second stage, where it is further classified as either a Typical Solution or a Creative Solution. As shown in Figure 6, this decision is made by comparing R against the reference solutions S provided in the original input prompt X , and determining whether it satisfies the criteria outlined in the guideline G .

B.2 Hyperparameter Setting

For all LLM Generators, the maximum input token length was set to 2000, and the maximum output token length was limited to 1023. Top- p was fixed at 1.0, and Top- k was fixed at 50 across all models. Temperature values were adjusted for each model to encourage the generation of Creative Solutions, and the final settings used for dataset construction are as follows:

- **DeepSeek-Math-7B (RLHF)** (deepseek-ai/deepseek-math-7b-r1): 0.7
- **Mathstral-7B** (mistralai/Mathstral-7b-v0.1): 0.25

Given the following mathematical problem:
What is the largest power of 2 that is a divisor of $(13^4 - 11^4)$?

Reference solutions:
Solution 1:
The factors of 36 are 1, 2, 3, 4, 6, 9, 12, 18, and 36. The multiples of 4 up to 36 are 4, 8, 12, 16, 20, 24, 28, 32, and 36. Only 4, 12, and 36 appear on both lists, so the answer is 3, which is option **(B)** 3.

Solution 2:
 $36 = 4^1 \cdot 3^2$. All possible factors of 36 will be here, except for ones divisible by 2 and not by 4. $(1+1) \cdot (2+1) = 6$. Subtract factors not divisible by 4, which are $1, 3^1$, and 3^2 . $6-3=3$, which is **(B)** 3.

New solution:
To solve the problem of finding how many positive factors of 36 are also multiples of 4, we can use a more systematic approach by first determining the prime factorization of 36 and then identifying the factors that are multiples of 4.
1. Prime Factorization of 36: $[36 = 2^2 \times 3^2]$
2. Identify the General Form of Factors: Any factor of 36 can be written in the form $(2^a \times 3^b)$, where $(0 \leq a \leq 2)$ and $(0 \leq b \leq 2)$.
3. Determine the Conditions for Multiples of 4: For a factor to be a multiple of 4, it must contain at least (2^2) in its prime factorization. Therefore, (a) must be at least 2.
4. Possible Values for (a) and (b):
- Since (a) must be at least 2, (a) can be either 2.
- (b) can be 0, 1, or 2.
5. List All Possible Factors:
- For (a = 2):
- (b = 0): $(2^2 \times 3^0 = 4)$
- (b = 1): $(2^2 \times 3^1 = 12)$
- (b = 2): $(2^2 \times 3^2 = 36)$
6. Count the Factors: There are 3 factors of 36 that are also multiples of 4: (4, 12,) and (36). Therefore, the number of positive factors of 36 that are also multiples of 4 is **(B)**.

Please output "YES" if the new solution leads to the same results as the reference solutions; otherwise, output "NO". YES or NO?

Figure 5: Example prompt used for correctness evaluation

Criteria for evaluating the difference between two mathematical solutions include:

- If the methods used to arrive at the solutions are fundamentally different, such as algebraic manipulation versus geometric reasoning, they can be considered distinct;
- Even if the final results are the same, if the intermediate steps or processes involved in reaching those solutions vary significantly, the solutions can be considered different;
- If two solutions rely on different assumptions or conditions, they are likely to be distinct;
- A solution might generalize to a broader class of problems, while another solution might be specific to certain conditions. In such cases, they are considered distinct;
- If one solution is significantly simpler or more complex than the other, they can be regarded as essentially different, even if they lead to the same result.

Given the following mathematical problem:
What is the largest power of 2 that is a divisor of $(13^4 - 11^4)$?

Reference solutions:
Solution 1:
The factors of 36 are 1, 2, 3, 4, 6, 9, 12, 18, and 36. The multiples of 4 up to 36 are 4, 8, 12, 16, 20, 24, 28, 32, and 36. Only 4, 12, and 36 appear on both lists, so the answer is 3, which is option **(B)** 3.

New solution:
To solve the problem of finding how many positive factors of 36 are also multiples of 4, we can use a more systematic approach by first determining the prime factorization of 36 and then identifying the factors that are multiples of 4.
1. Prime Factorization of 36: $[36 = 2^2 \times 3^2]$
2. Identify the General Form of Factors: Any factor of 36 can be written in the form $(2^a \times 3^b)$, where $(0 \leq a \leq 2)$ and $(0 \leq b \leq 2)$.
3. Determine the Conditions for Multiples of 4: For a factor to be a multiple of 4, it must contain at least (2^2) in its prime factorization. Therefore, (a) must be at least 2.
4. Possible Values for (a) and (b):
- Since (a) must be at least 2, (a) can be either 2.
- (b) can be 0, 1, or 2.
5. List All Possible Factors:
- For (a = 2):
- (b = 0): $(2^2 \times 3^0 = 4)$
- (b = 1): $(2^2 \times 3^1 = 12)$
- (b = 2): $(2^2 \times 3^2 = 36)$
6. Count the Factors: There are 3 factors of 36 that are also multiples of 4: (4, 12,) and (36). Therefore, the number of positive factors of 36 that are also multiples of 4 is **(B)**.

Please output "YES" if the new solution is a novel solutions; otherwise, output "NO". YES or NO?

Figure 6: Example prompt used for novelty (creativity) evaluation

- 1090 • **OpenMath2-LLaMA3.1-8B** (nvidia/OpenMath2-Llama3.1-8B): 1.0
- 1091 • **OREAL-7B** (internlm/OREAL-7B): 0.7
- 1092 • **Qwen-2.5-Math-7B** (Qwen/Qwen2.5-Math-7B-Instruct): 0.7

1093 All generations were performed on NVIDIA RTX A5000 GPUs (24GB VRAM), using up to 8 GPUs
1094 in parallel.

1095 The LLM Evaluators used in our study are listed below:

- 1096 • **Gemini-1.5-Pro**: models/gemini-1.5-pro-002
- 1097 • **GPT-o4-mini**: o4-mini-2025-04-16

1098 We evaluate our methods using a variety of Evaluation strategies, including thresholding, distance-
1099 based prototype matching, and trainable models such as MLP, TabM, and Decision-tree based
1100 XGBOOST. The implementation and hyperparameter settings for each method are summarized
1101 below.

1102 **Threshold (only for Baselines)** We divide the value range of each baseline measure into 200 intervals
1103 and evaluate performance at each threshold. The threshold that achieves the best macro-f1 score on
1104 the Reference Set is selected for final evaluation.

1105 **Prototype (only for CLAWS)** We used an Encoder consisting of two Linear Layers for the prototype-
1106 based evaluation method. The input dimension is reduced to 16 dimensions through the first Linear
1107 Layer and reduced to 8 dimensions through the second Layer. After that, it is expanded to 16
1108 dimensions again and reduced to the dimension corresponding to the final number of classes, and
1109 used as the output. The output of each data was averaged by class and used as the class center value.
1110 Afterwards, the Euclidean distance between each data sample and the class center was calculated to
1111 predict the closest class. We generated prototypes of the Reference Set using 20 different random
1112 seeds and presented the one that achieved the best macro-f1 score.

1113 **MLP** We use a three-layer feed-forward neural network. The model is trained for 10 epochs using
1114 cross-entropy loss with class weights to account for class imbalance. Optimization is performed
1115 using Adam with a learning rate of 0.001. Depending on the number of classes and the input feature
1116 dimension, the model contains up to 133 learnable parameters. We experimented with 20 different
1117 random seeds and presented the one that gave the best macro-f1 score.

1118 **XGBOOST** We use the multiclass error rate for 3-class detection and binary cross-entropy for 2-class
1119 detection.

1120 **TabM** We use TabM, an ensemble of 32 independently parameterized MLPs. Each MLP has 512
1121 parameters. model in the ensemble is trained for 20 epochs using cross-entropy loss. Optimization is
1122 performed using AdamW with a learning rate of $2e-3$ and a weight decay of $3e-4$.

1123 **C Experimental Results**

1124 **C.1 Visualizations for Reference Set**

1125 Figure 7 presents class-wise average scores across different methods on the Reference Set for four
1126 models. Results for Qwen-2.5-Math-7B are shown separately in Figure 4.

1127 **C.2 3-Class Detection Full Table**

1128 The results of 3-class detection for each model using different evaluation strategies (see Section 4.2)
1129 are presented in Tables 4–8. The results for the Threshold and Prototype-based methods are shown
1130 separately in Table 2.

1131 **C.3 2-Class Detection (Hallucination Detection) Full Table**

1132 The results of 2-class detection for each model using different evaluation strategies are presented in
1133 Tables 9–13. The results for the Threshold and Prototype-based methods are shown separately in
1134 Table 3.

1135 **C.4 3-Class Detection for Balanced Reference set Full Table**

1136 The results of 3-class balanced detection using the Threshold and Prototype-based methods are shown
1137 in Table 14, followed by model-specific results using other evaluation strategies in Tables 15–19.

1138 **C.5 Efficiency Comparison between Baseline Methods and CLAWS**

1139 We compare the runtime efficiency of our proposed method, CLAWS, with baseline feature extraction
1140 methods. As CLAWS leverages the attention weights obtained during response generation, it
1141 incurs minimal additional computational cost and is the fastest among all methods. The results are
1142 summarized in Figure 8.

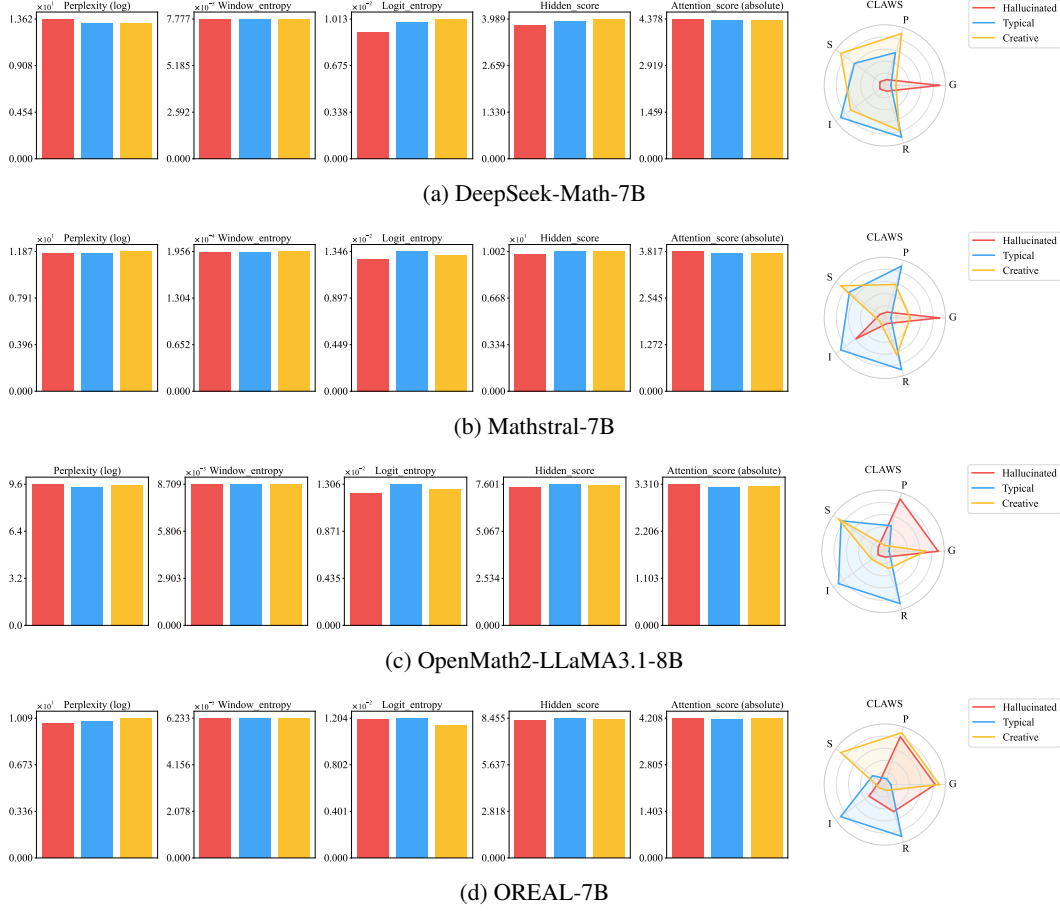


Figure 7: Visualization of class-wise average scores across all evaluation methods for four models on the Reference Set. For CLAWS, the scores are normalized and clipped to the range $[0.1, 0.9]$ to enhance visual clarity.

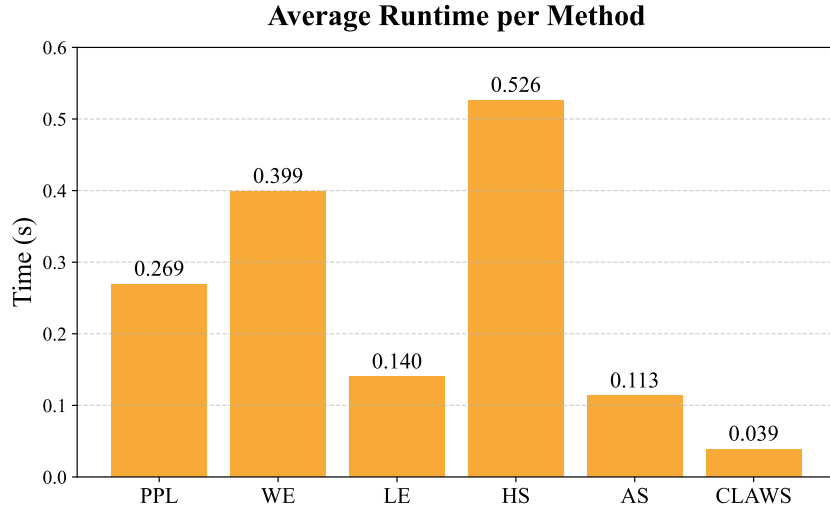


Figure 8: Average runtime required to compute the input features for each method — PPL (Perplexity), WE (Window Entropy), LE (Logit Entropy), HS (Hidden Score), AS (Attention Score), and CLAWS (ours) — which are subsequently used in evaluation strategies (see Section 4 for details). Response R generation time is excluded as it is shared across all methods. Notably, our proposed method CLAWS requires the least computation time.

Dataset		TEST				AMC				AIME				A(J)HSME			
Strategy	Method	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC
MLP	PPL	41.04	31.64	40.85	59.94	42.58	33.79	<u>39.38</u>	<u>56.92</u>	53.72	<u>35.93</u>	<u>40.08</u>	<u>61.61</u>	36.36	30.81	<u>39.57</u>	<u>57.42</u>
	WE	<u>58.29</u>	<u>42.05</u>	42.32	62.73	44.30	35.70	37.58	56.56	<u>56.51</u>	31.67	35.90	54.68	38.62	33.61	36.95	55.99
	LE	39.99	32.84	35.47	52.45	41.30	<u>35.43</u>	35.48	52.18	46.20	30.22	35.41	53.91	<u>38.85</u>	<u>35.80</u>	36.67	53.31
	HS	48.79	41.59	42.00	62.98	35.61	27.90	34.76	52.59	35.63	28.31	32.55	48.47	34.19	32.00	34.58	50.59
	AS	51.05	37.63	<u>43.02</u>	<u>64.31</u>	36.61	28.87	33.88	50.68	46.90	30.25	34.07	49.98	32.78	27.70	33.63	49.79
	CLAWS	61.43	49.65	<u>50.08</u>	71.41	<u>43.78</u>	34.84	41.01	60.11	57.39	39.88	42.26	62.56	44.11	39.65	42.64	60.12
XGBOOST	PPL	50.32	36.08	37.52	56.82	40.34	32.65	36.65	53.92	<u>55.57</u>	34.55	<u>36.68</u>	<u>55.38</u>	38.30	33.32	37.23	<u>55.15</u>
	WE	59.39	<u>42.20</u>	<u>43.01</u>	<u>63.81</u>	<u>41.53</u>	<u>33.75</u>	<u>38.03</u>	57.36	56.51	31.67	35.92	55.01	<u>38.52</u>	<u>33.52</u>	37.16	56.25
	LE	45.62	34.29	34.82	52.23	37.96	31.41	34.63	51.27	50.71	33.27	35.18	52.83	34.09	30.10	34.80	52.29
	HS	49.14	37.55	38.40	56.61	39.37	32.70	34.52	51.19	47.59	31.38	32.97	49.50	35.47	31.85	34.74	51.74
	AS	53.58	38.76	42.19	63.53	38.09	30.36	34.31	51.09	51.56	31.66	33.97	49.75	33.85	28.86	33.85	50.40
	CLAWS	<u>57.85</u>	45.59	47.98	68.39	44.37	37.67	38.57	<u>56.99</u>	51.12	<u>34.53</u>	38.65	57.31	39.04	34.49	<u>37.23</u>	54.88
TabM	PPL	51.79	37.78	41.03	60.03	<u>43.76</u>	<u>35.20</u>	<u>39.29</u>	56.83	<u>56.84</u>	<u>35.80</u>	<u>41.01</u>	61.73	40.89	35.30	<u>39.67</u>	<u>57.42</u>
	WE	<u>59.39</u>	<u>42.20</u>	<u>42.93</u>	63.27	41.53	33.75	38.00	<u>57.13</u>	56.51	31.67	36.16	54.97	38.62	33.61	37.17	56.13
	LE	42.80	31.66	35.39	53.49	40.44	32.32	35.45	51.89	48.06	30.82	36.39	54.66	37.27	32.04	36.73	53.64
	HS	51.25	36.33	42.64	63.04	38.75	31.17	34.86	52.60	48.77	29.56	32.40	48.14	34.58	29.99	35.34	51.85
	AS	51.05	37.63	42.83	<u>64.13</u>	36.61	28.87	33.97	50.77	46.90	30.25	33.76	49.82	32.78	27.70	33.68	49.85
	CLAWS	60.78	45.03	51.37	72.82	46.46	37.56	41.06	59.26	57.05	37.77	41.44	<u>60.73</u>	<u>40.81</u>	<u>35.10</u>	41.25	59.12

Table 4: Evaluation results for 3-class detection using DeepSeek-Math-7B. Bold values indicate the best performance, underlined values indicate the second-best. Light gray-shaded cells correspond to results where the model performed detection over only 2 out of the 3 target classes, while dark gray-shaded cells indicate cases where the model predicted only 1 out of the 3 target classes.

Dataset		TEST				AMC				AIME				A(J)HSME			
Strategy	Method	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC
MLP	PPL	53.42	32.71	39.61	59.24	36.93	28.62	36.25	52.83	54.76	33.05	34.34	52.57	34.71	26.01	35.11	51.70
	WE	49.45	28.16	37.02	55.53	37.14	23.20	35.41	52.17	<u>65.47</u>	28.70	<u>34.48</u>	<u>52.90</u>	33.06	22.45	33.79	49.80
	LE	49.54	32.89	37.04	56.02	46.20	34.91	36.55	54.41	57.77	31.25	32.55	48.89	38.70	31.36	34.14	51.37
	HS	60.32	40.84	<u>45.48</u>	<u>68.58</u>	38.35	30.42	<u>39.90</u>	<u>58.59</u>	63.39	32.59	33.02	50.08	38.05	30.74	35.76	53.95
	AS	<u>62.82</u>	<u>42.48</u>	44.00	63.55	<u>49.70</u>	<u>37.03</u>	38.63	58.10	61.26	<u>33.63</u>	33.59	50.32	<u>41.86</u>	<u>33.38</u>	<u>37.48</u>	<u>55.85</u>
	CLAWS	64.65	45.34	47.74	70.25	52.01	41.33	41.89	61.03	69.54	39.70	41.51	61.86	45.81	40.58	42.67	60.58
XGBOOST	PPL	55.02	31.81	36.77	56.43	41.89	29.03	34.58	51.87	64.84	<u>31.95</u>	34.91	<u>54.42</u>	38.87	28.83	35.90	53.20
	WE	58.25	34.05	44.34	67.71	<u>45.13</u>	31.05	41.15	61.71	<u>65.90</u>	29.74	<u>35.19</u>	54.33	<u>41.84</u>	<u>31.38</u>	<u>39.03</u>	<u>58.88</u>
	LE	52.67	29.93	33.46	49.70	43.28	30.38	36.59	54.33	63.76	31.62	32.94	49.05	37.13	27.49	34.06	51.18
	HS	<u>59.40</u>	<u>38.04</u>	41.71	62.34	45.00	<u>34.05</u>	37.12	54.68	64.85	30.67	32.82	49.71	38.97	29.97	35.07	52.29
	AS	57.64	34.26	41.54	60.55	41.61	28.15	39.25	<u>57.78</u>	65.34	28.87	33.94	49.11	38.13	28.34	36.23	53.93
	CLAWS	62.05	42.09	<u>44.25</u>	<u>64.43</u>	50.92	38.85	<u>40.31</u>	57.69	69.30	40.42	41.21	59.53	47.04	38.75	40.66	59.06
TabM	PPL	52.72	26.55	38.26	59.43	37.14	23.20	38.09	56.46	65.69	28.80	<u>35.89</u>	<u>54.48</u>	39.68	28.93	32.67	48.80
	WE	52.72	26.55	37.06	55.54	37.72	23.78	35.59	52.25	<u>66.14</u>	29.66	34.96	53.40	45.45	34.94	35.90	53.75
	LE	52.39	26.74	35.96	54.93	37.81	23.92	36.26	53.96	65.45	29.51	32.01	47.47	41.61	31.58	34.33	50.45
	HS	64.38	<u>42.69</u>	<u>45.11</u>	68.38	48.44	<u>35.81</u>	<u>39.97</u>	<u>58.39</u>	65.09	29.99	32.99	50.08	<u>48.20</u>	<u>36.14</u>	<u>39.90</u>	<u>57.52</u>
	AS	58.04	33.81	43.98	<u>66.28</u>	42.35	28.54	39.84	58.55	66.10	<u>30.22</u>	34.19	51.11	45.81	34.68	36.07	52.79
	CLAWS	<u>62.96</u>	43.67	45.16	65.28	50.44	38.13	40.67	58.14	68.05	37.58	40.03	58.42	54.29	41.03	46.14	65.47

Table 5: Evaluation results for 3-class detection using Mathstral-7B

Dataset		TEST				AMC				AIME				A(J)HSME			
Strategy	Method	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC
MLP	PPL	32.70	24.95	31.40	47.66	40.22	28.52	32.40	50.26	37.94	27.23	30.08	44.11	39.52	28.74	33.17	51.08
	WE	54.52	38.02	37.93	57.74	45.19	32.85	35.74	52.96	36.87	23.13	33.27	49.26	43.96	33.30	34.99	51.99
	LE	37.41	27.93	33.52	48.99	42.73	30.66	32.41	47.40	41.31	29.13	30.58	44.97	41.39	30.77	33.39	49.88
	HS	57.11	<u>40.73</u>	<u>46.31</u>	<u>66.72</u>	50.02	36.02	<u>40.98</u>	<u>58.74</u>	49.78	<u>35.33</u>	<u>36.12</u>	<u>52.65</u>	<u>48.00</u>	<u>35.80</u>	<u>39.40</u>	<u>56.58</u>
	AS	<u>57.30</u>	40.33	42.38	63.12	<u>50.74</u>	<u>36.76</u>	37.99	54.72	<u>49.81</u>	34.53	35.05	51.70	46.34	34.75	35.91	52.32
	CLAWS	65.32	47.69	52.80	73.42	58.31	43.82	48.59	67.67	58.11	42.09	44.17	61.44	52.72	42.31	45.52	64.63
XGBOOST	PPL	47.78	33.14	34.54	52.60	42.29	31.00	33.16	49.65	46.38	32.22	32.55	50.81	40.45	31.02	34.02	51.25
	WE	<u>58.84</u>	<u>40.14</u>	<u>43.09</u>	<u>63.58</u>	49.28	<u>36.50</u>	<u>40.17</u>	<u>58.33</u>	43.16	28.33	<u>35.66</u>	<u>52.80</u>	<u>47.00</u>	<u>36.16</u>	<u>38.35</u>	<u>57.20</u>
	LE	47.12	33.47	34.59	50.79	44.35	33.34	34.88	52.15	44.58	30.92	33.01	48.48	39.81	30.93	33.49	49.91
	HS	54.90	38.46	41.22	60.99	48.40	35.39	37.51	54.81	48.65	<u>34.10</u>	34.91	51.78	45.77	35.26	37.85	55.67
	AS	57.43	40.10	42.07	61.57	<u>49.95</u>	36.40	37.26	54.48	<u>48.73</u>	33.54	35.32	50.98	44.56	33.59	35.40	52.01
	CLAWS	63.20	46.23	48.16	67.22	54.56	40.24	42.68	61.61	56.87	40.82	42.51	59.67	50.40	38.13	42.63	61.47
TabM	PPL	35.59	26.46	31.59	47.64	40.66	28.92	32.32	49.46	39.10	27.92	29.78	43.42	32.90	22.11	35.43	52.59
	WE	58.46	40.05	40.11	60.38	48.15	35.64	36.65	54.54	45.85	30.55	35.01	51.67	33.20	22.42	34.08	50.31
	LE	47.65	33.30	35.52	52.44	45.31	33.11	34.86	52.12	44.65	31.01	31.60	45.48	33.40	22.78	35.11	52.27
	HS	<u>58.51</u>	<u>41.39</u>	<u>46.16</u>	<u>66.42</u>	<u>52.12</u>	<u>37.80</u>	<u>41.05</u>	<u>58.90</u>	<u>50.78</u>	<u>35.73</u>	<u>36.14</u>	<u>52.56</u>	<u>41.22</u>	<u>31.66</u>	36.10	54.01
	AS	57.09	39.55	42.58	63.30	51.33	37.50	38.17	55.22	47.05	32.05	35.16	51.83	39.04	28.53	<u>37.16</u>	<u>55.69</u>
	CLAWS	66.45	46.83	52.95	73.17	58.70	42.78	47.83	66.93	58.92	40.95	44.10	60.67	54.20	41.15	46.42	64.95

Table 6: Evaluation results for 3-class detection using OpenMath2-LLaMA3.1-8B

Dataset		TEST				AMC				AIME				A(J)HSME			
Strategy	Method	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC
MLP	PPL	59.88	35.79	37.54	56.31	58.17	41.28	41.52	62.85	68.61	35.39	35.91	57.86	<u>48.78</u>	<u>36.16</u>	<u>41.35</u>	<u>59.53</u>
	WE	53.83	26.76	33.11	49.04	47.70	25.56	32.56	49.16	73.07	29.92	33.51	51.24	39.22	23.70	34.50	50.36
	LE	53.76	26.72	33.25	49.11	47.70	25.56	33.25	51.68	73.07	29.92	<u>34.31</u>	52.45	46.63	31.35	35.59	53.16
	HS	57.25	37.18	<u>39.90</u>	<u>58.83</u>	48.60	<u>34.10</u>	39.34	56.63	62.86	31.67	33.29	50.56	39.22	23.70	33.37	50.04
	AS	62.98	<u>38.54</u>	39.78	55.90	<u>54.91</u>	36.53	37.32	55.57	<u>68.89</u>	33.82	34.24	<u>52.77</u>	49.86	35.87	36.03	53.61
	CLAWS	<u>60.12</u>	41.87	42.39	61.48	51.30	<u>37.14</u>	<u>40.19</u>	<u>59.47</u>	66.67	<u>34.20</u>	33.83	50.25	48.76	37.00	41.53	59.78
XGBOOST	PPL	56.08	31.31	35.98	54.06	51.46	30.83	36.59	56.38	72.38	33.10	33.81	50.45	44.41	29.62	35.92	53.99
	WE	59.33	33.74	36.28	54.38	<u>55.08</u>	33.66	36.80	55.13	73.10	30.26	33.47	51.08	50.95	34.42	<u>37.19</u>	<u>54.80</u>
	LE	54.29	28.42	33.05	48.22	50.04	28.78	33.53	50.16	72.80	32.68	33.59	50.18	43.16	27.57	33.80	51.14
	HS	<u>61.54</u>	<u>38.44</u>	<u>39.26</u>	<u>58.61</u>	57.30	38.33	<u>38.29</u>	<u>56.62</u>	70.42	33.57	<u>33.90</u>	<u>53.02</u>	52.60	36.87	37.58	55.27
	AS	53.83	26.76	36.00	53.50	47.70	25.56	35.93	53.68	<u>73.07</u>	29.92	34.63	55.29	39.22	23.70	34.03	50.52
	CLAWS	62.41	40.83	42.20	62.33	54.03	<u>37.02</u>	38.32	57.29	70.94	<u>33.49</u>	33.87	49.05	49.02	<u>35.11</u>	36.37	54.38
TabM	PPL	53.83	26.76	37.84	57.08	47.70	25.56	38.36	58.89	<u>73.07</u>	29.92	35.26	56.47	39.22	23.70	38.30	57.69
	WE	59.16	34.17	35.25	52.75	55.08	33.66	36.22	54.09	73.10	30.26	33.35	50.55	<u>50.95</u>	34.42	36.90	54.54
	LE	53.83	26.76	31.97	46.90	47.70	25.56	32.06	49.43	<u>73.07</u>	29.92	32.20	47.67	39.22	23.70	34.54	52.61
	HS	<u>62.28</u>	<u>38.75</u>	<u>40.26</u>	<u>58.73</u>	57.60	38.44	39.42	57.29	71.32	<u>33.38</u>	33.59	<u>51.54</u>	52.55	36.74	<u>38.95</u>	56.34
	AS	53.83	26.76	39.79	58.17	47.70	25.56	37.33	55.62	<u>73.07</u>	29.92	33.94	49.19	39.22	23.70	39.28	56.80
	CLAWS	62.86	39.09	42.41	63.48	<u>55.53</u>	<u>37.45</u>	<u>38.75</u>	<u>58.23</u>	71.21	33.92	<u>35.08</u>	51.07	<u>51.27</u>	<u>36.49</u>	38.07	<u>57.23</u>

Table 7: Evaluation results for 3-class detection using OREAL-7B

Dataset		TEST				AMC				AIME				A(J)HSME			
Strategy	Method	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC
MLP	PPL	<u>54.27</u>	46.34	47.90	67.59	50.01	<u>38.01</u>	43.58	<u>61.96</u>	47.58	36.13	41.91	62.02	44.92	37.00	<u>38.23</u>	<u>57.72</u>
	WE	45.67	39.58	38.31	54.97	43.32	32.95	36.49	53.87	40.26	30.78	34.22	51.26	42.11	33.31	34.35	50.53
	LE	29.48	23.10	28.27	41.43	36.53	26.33	33.96	49.47	31.62	25.13	31.19	47.00	<u>32.30</u>	<u>23.38</u>	<u>35.14</u>	<u>50.73</u>
	HS	54.58	42.77	46.98	65.86	<u>50.29</u>	36.64	41.90	60.61	<u>43.11</u>	32.96	36.01	52.38	<u>46.24</u>	36.21	37.80	56.01
	AS	43.80	38.05	40.62	59.13	43.04	32.73	35.52	52.88	39.07	30.08	32.98	47.71	39.86	30.64	34.43	51.33
	CLAWS	53.78	<u>45.32</u>	50.44	67.98	52.51	39.65	43.06	63.20	42.76	36.59	<u>39.32</u>	<u>56.49</u>	53.75	41.05	42.29	62.61
XGBOOST	PPL	48.70	<u>41.51</u>	41.49	59.98	46.82	37.65	36.90	54.52	<u>44.53</u>	36.78	37.82	55.39	44.21	<u>34.63</u>	35.26	53.28
	WE	<u>49.96</u>	39.15	<u>43.47</u>	<u>63.00</u>	<u>49.93</u>	36.57	38.63	57.69	41.13	32.17	36.10	52.48	<u>45.39</u>	32.74	<u>36.82</u>	<u>55.51</u>
	LE	38.22	32.42	33.83	50.94	40.59	31.06	34.20	51.68	35.19	28.98	33.38	49.97	37.81	28.36	34.17	50.89
	HS	46.98	39.95	41.80	57.98	48.68	<u>37.94</u>	<u>39.70</u>	<u>58.48</u>	41.31	33.40	50.87	34.65	44.80	33.46	36.41	54.44
	AS	42.60	35.24	37.91	55.60	42.92	32.76	35.51	53.54	38.85	30.92	33.74	49.56	40.00	29.96	34.70	52.49
	CLAWS	52.35	43.33	47.72	65.98	50.30	38.98	40.66	61.11	45.18	38.95	<u>39.75</u>	55.86	47.54	36.02	39.41	59.45
TabM	PPL	54.53	<u>42.74</u>	47.13	65.44	50.83	<u>38.45</u>	43.07	<u>60.61</u>	48.30	<u>37.01</u>	43.37	60.80	45.58	<u>34.28</u>	<u>38.54</u>	<u>57.10</u>
	WE	48.15	37.73	40.35	58.70	46.62	33.89	36.59	55.06	38.66	30.12	34.82	51.49	45.58	31.17	34.81	52.06
	LE	33.24	26.04	32.55	50.24	41.43	28.20	34.61	52.39	31.50	25.36	33.98	50.30	38.81	25.93	35.89	52.41
	HS	51.47	41.39	45.50	63.64	50.08	36.79	41.58	61.00	43.75	34.48	36.73	53.74	<u>46.01</u>	33.00	37.82	55.41
	AS	45.76	35.86	38.65	56.76	43.53	31.74	36.03	54.26	37.84	29.68	33.06	49.16	41.37	28.96	35.60	54.28
	CLAWS	<u>51.93</u>	42.92	<u>45.82</u>	<u>63.83</u>	48.69	38.90	39.23	59.61	<u>45.14</u>	38.28	<u>38.15</u>	<u>55.26</u>	47.31	35.97	39.23	59.28

Table 8: Evaluation results for 3-class detection using Qwen-2.5-Math-7B

Dataset		TEST				AMC				AIME				A(J)HSME			
Strategy	Method	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC
MLP	PPL	53.26	54.05	60.78	63.40	<u>58.89</u>	<u>55.25</u>	<u>57.58</u>	60.10	59.93	57.91	<u>60.78</u>	63.60	<u>63.44</u>	<u>58.87</u>	<u>59.02</u>	<u>61.55</u>
	WE	<u>65.53</u>	<u>63.87</u>	62.80	65.76	50.79	52.21	56.97	<u>60.84</u>	60.76	54.86	53.73	55.66	57.90	56.12	55.34	59.09
	LE	47.54	48.57	50.43	50.62	55.57	53.13	53.23	53.82	53.84	51.23	52.66	53.65	58.38	54.61	54.77	55.73
	HS	61.89	62.14	<u>63.48</u>	66.54	53.86	48.59	52.13	52.39	42.56	42.32	48.47	46.41	55.68	49.34	50.96	51.83
	AS	61.26	61.37	62.99	<u>67.08</u>	53.47	49.26	50.75	50.82	52.64	48.97	50.77	50.12	55.74	49.52	50.03	49.90
	CLAWS	70.69	70.51	76.44	78.35	63.48	61.29	62.59	65.67	<u>60.50</u>	<u>57.34</u>	61.15	<u>62.70</u>	65.12	61.11	61.64	64.64
XGBOOST	PPL	56.31	56.08	52.18	59.45	<u>54.76</u>	<u>53.53</u>	66.15	55.92	<u>57.65</u>	53.32	<u>40.16</u>	<u>56.75</u>	<u>58.61</u>	<u>56.18</u>	70.04	58.75
	WE	<u>65.53</u>	<u>63.87</u>	<u>60.55</u>	<u>68.13</u>	50.79	52.21	68.78	60.84	59.51	49.52	38.21	55.66	53.27	53.41	<u>70.15</u>	<u>59.51</u>
	LE	50.45	50.49	44.87	52.12	52.12	50.32	62.19	51.76	53.25	50.66	36.64	53.36	54.86	52.15	65.98	53.49
	HS	57.91	57.84	49.82	58.86	53.23	49.99	61.94	50.51	46.49	44.95	31.93	46.78	56.00	51.74	64.72	50.91
	AS	61.26	61.37	57.39	67.27	53.47	49.26	61.88	50.82	52.64	48.97	34.84	50.44	55.74	49.52	64.52	49.93
	CLAWS	66.15	66.02	66.06	73.21	58.95	56.00	<u>68.65</u>	<u>59.39</u>	55.34	<u>52.80</u>	43.24	58.22	61.12	56.83	71.01	59.99
TabM	PPL	59.01	58.98	54.53	63.40	<u>58.42</u>	<u>57.20</u>	<u>69.78</u>	60.62	62.95	58.42	<u>45.75</u>	63.60	<u>62.21</u>	<u>59.47</u>	<u>71.05</u>	<u>61.55</u>
	WE	<u>65.53</u>	<u>63.87</u>	<u>60.57</u>	<u>68.13</u>	50.79	52.21	68.78	<u>60.84</u>	59.51	49.52	38.17	55.64	53.46	53.58	70.20	59.56
	LE	47.53	48.49	41.98	50.62	55.79	52.92	63.51	53.82	52.33	50.51	36.12	53.71	58.74	54.71	66.79	55.29
	HS	60.90	61.22	54.83	65.20	53.59	48.45	63.08	52.67	42.46	42.23	33.84	46.98	55.16	48.54	67.30	53.87
	AS	62.95	62.43	57.29	67.33	53.15	50.65	61.96	50.82	56.37	50.39	35.48	50.12	54.40	49.92	64.47	49.94
	CLAWS	70.80	70.40	68.96	77.03	61.18	58.52	71.23	62.34	<u>61.27</u>	<u>57.73</u>	46.94	<u>63.37</u>	64.01	59.84	74.78	64.70

Table 9: Evaluation results for hallucination detection using DeepSeek-Math-7B. Bold and underlined values indicate the best and second-best performance, respectively. Light gray-shaded cells indicate cases where the model predicted only a single class in a 2-class detection setting.

Dataset		TEST				AMC				AIME				A(J)HSME			
Strategy	Method	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC
MLP	PPL	62.44	58.35	59.79	63.82	55.80	55.65	57.72	58.83	58.87	51.43	<u>52.67</u>	<u>55.24</u>	52.06	52.02	53.64	54.01
	WE	52.42	44.27	55.67	56.78	38.24	35.98	53.10	52.71	<u>66.08</u>	44.36	52.01	53.63	50.57	50.63	51.02	50.27
	LE	54.35	53.02	55.60	58.33	55.17	55.27	55.44	57.35	50.54	43.69	48.03	46.99	52.10	52.04	52.61	54.60
	HS	68.29	<u>65.92</u>	<u>69.13</u>	<u>72.56</u>	53.29	53.85	<u>59.92</u>	61.18	65.09	51.03	49.89	49.29	50.87	50.79	53.46	54.78
	AS	<u>69.28</u>	65.39	65.30	68.56	<u>58.59</u>	<u>58.45</u>	59.61	<u>62.04</u>	63.78	<u>52.56</u>	50.74	51.46	<u>55.17</u>	<u>55.20</u>	<u>56.01</u>	<u>57.89</u>
	CLAWS	72.60	70.23	71.58	76.46	63.97	63.91	64.43	67.13	69.36	55.98	60.50	65.42	57.48	57.41	61.13	63.78
XGBOOST	PPL	59.16	51.63	39.99	57.91	48.13	46.97	48.19	52.02	<u>66.72</u>	<u>53.03</u>	25.93	53.45	47.97	48.07	54.48	53.54
	WE	61.35	52.24	<u>52.44</u>	70.96	48.34	46.73	59.67	65.49	66.05	44.75	<u>26.79</u>	<u>54.60</u>	47.97	48.12	<u>59.81</u>	62.04
	LE	55.98	48.86	33.54	50.88	49.70	48.54	50.70	54.97	63.63	47.62	22.38	47.37	45.96	46.08	50.91	51.38
	HS	<u>64.41</u>	<u>58.97</u>	49.50	66.57	<u>53.31</u>	<u>52.71</u>	52.16	57.07	65.91	48.86	23.06	48.36	<u>50.24</u>	<u>50.32</u>	53.49	53.79
	AS	63.29	54.88	52.28	66.47	47.93	46.39	56.53	<u>61.52</u>	66.37	46.10	25.84	52.21	45.81	45.96	56.78	56.92
	CLAWS	68.93	64.74	54.22	<u>70.14</u>	57.86	57.42	<u>57.63</u>	60.81	68.62	55.25	34.92	59.35	56.64	56.66	61.68	<u>61.07</u>
TabM	PPL	52.72	39.83	41.56	63.13	37.14	34.80	52.24	58.69	65.69	43.19	<u>28.13</u>	<u>55.72</u>	32.90	33.16	51.18	52.35
	WE	53.00	40.24	48.29	67.44	38.28	36.03	55.86	62.65	<u>66.08</u>	<u>44.36</u>	26.90	<u>54.81</u>	33.56	33.82	55.86	56.94
	LE	53.17	41.09	37.32	57.25	38.42	36.18	51.03	56.34	65.43	43.89	21.57	46.00	34.25	34.49	52.13	53.35
	HS	<u>70.78</u>	<u>65.92</u>	<u>55.96</u>	<u>69.92</u>	<u>58.19</u>	<u>57.72</u>	54.09	59.79	65.07	44.02	24.20	51.09	<u>50.56</u>	<u>50.65</u>	54.18	54.76
	AS	65.28	57.77	54.07	69.09	52.23	50.99	<u>57.18</u>	62.07	<u>66.85</u>	<u>47.91</u>	25.53	51.90	49.70	49.82	<u>56.85</u>	<u>57.84</u>
	CLAWS	71.24	66.85	58.34	74.90	59.83	59.33	57.96	<u>62.41</u>	70.49	54.75	34.28	59.23	58.69	58.70	60.09	61.38

Table 10: Evaluation results for hallucination detection using Mathstral-7B

Dataset		TEST				AMC				AIME				A(J)HSME			
Strategy	Method	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC
MLP	PPL	40.48	41.94	46.39	44.60	48.50	45.26	49.84	49.79	41.54	42.24	44.95	42.76	50.68	45.38	45.14	43.00
	WE	59.94	59.40	58.43	61.08	54.28	53.64	53.62	55.24	51.44	51.03	52.21	52.93	54.54	52.81	52.34	53.08
	LE	45.18	46.17	53.63	54.10	52.39	50.32	51.29	51.79	44.98	45.19	46.27	44.54	53.68	49.98	49.79	49.22
	HS	<u>64.07</u>	<u>64.21</u>	<u>69.38</u>	<u>71.21</u>	<u>60.17</u>	<u>58.12</u>	<u>61.24</u>	<u>63.68</u>	52.88	<u>53.21</u>	<u>53.96</u>	<u>55.40</u>	<u>60.69</u>	<u>57.26</u>	<u>58.61</u>	<u>60.95</u>
	AS	62.94	62.67	63.22	66.27	58.37	56.76	56.78	59.16	<u>53.17</u>	52.77	52.56	53.79	56.96	54.02	53.65	55.54
	CLAWS	71.11	70.76	77.32	78.66	66.07	64.48	66.29	69.34	61.15	60.82	61.13	63.52	63.75	60.69	64.97	67.89
XGBOOST	PPL	50.29	50.23	45.17	51.55	49.68	48.63	57.68	48.70	48.47	48.42	46.56	48.88	50.25	48.29	62.56	49.13
	WE	<u>63.59</u>	62.14	<u>61.51</u>	<u>68.35</u>	54.17	55.17	<u>68.09</u>	<u>63.29</u>	43.86	42.25	50.17	<u>54.96</u>	54.85	<u>55.10</u>	<u>69.61</u>	<u>60.47</u>
	LE	51.35	51.07	46.90	53.15	51.61	50.97	59.25	51.71	47.46	47.15	45.91	47.80	48.39	46.90	60.89	47.56
	HS	60.00	59.81	54.59	64.55	56.74	55.33	64.64	57.95	51.68	<u>51.65</u>	<u>50.50</u>	52.95	<u>56.39</u>	53.87	68.48	57.27
	AS	63.52	<u>63.06</u>	57.72	66.34	<u>58.27</u>	<u>57.02</u>	63.49	57.37	<u>51.99</u>	51.46	49.91	54.12	56.36	53.87	64.44	53.77
	CLAWS	67.70	67.18	67.37	72.71	62.63	61.48	69.63	64.47	59.28	58.87	58.90	62.40	62.85	60.48	73.57	64.44
TabM	PPL	44.17	44.84	44.73	47.61	49.54	47.44	57.60	47.80	44.17	44.14	43.89	43.50	48.89	45.42	60.63	46.06
	WE	59.94	59.40	51.84	61.08	54.28	53.64	61.87	55.35	51.44	51.03	49.40	52.96	54.54	52.81	63.59	53.08
	LE	52.78	52.51	48.32	55.00	52.50	51.56	59.79	51.76	47.48	47.23	43.22	45.68	51.77	50.04	65.28	52.06
	HS	<u>65.03</u>	<u>64.90</u>	<u>63.72</u>	<u>71.21</u>	<u>60.35</u>	<u>58.89</u>	<u>69.00</u>	<u>63.68</u>	<u>54.66</u>	<u>54.59</u>	<u>50.40</u>	<u>55.40</u>	<u>60.25</u>	<u>57.57</u>	<u>70.61</u>	<u>60.95</u>
	AS	62.90	62.26	57.64	66.27	58.20	57.27	64.66	59.16	50.65	49.97	48.80	53.79	55.84	53.76	64.66	54.94
	CLAWS	69.62	69.19	69.18	74.51	65.20	63.98	74.83	69.17	61.15	60.82	61.13	63.52	65.13	62.79	75.60	67.35

Table 11: Evaluation results for hallucination detection using OpenMath2-LLaMA3.1-8B

Dataset		TEST				AMC				AIME				A(J)HSME			
Strategy	Method	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC
MLP	PPL	63.58	57.86	58.36	60.87	61.39	57.22	60.84	63.86	72.77	50.52	54.10	57.11	57.81	<u>56.53</u>	62.58	63.95
	WE	53.83	40.14	53.34	55.76	47.70	38.35	50.38	49.42	73.07	44.88	49.57	48.51	39.22	35.55	46.93	43.33
	LE	53.83	40.14	48.33	46.47	47.70	38.35	47.30	45.62	73.07	44.88	48.70	45.18	39.22	35.55	50.10	50.16
	HS	59.85	56.95	59.13	61.23	57.76	<u>56.68</u>	59.33	60.93	65.99	49.85	<u>51.27</u>	51.93	55.95	55.86	58.06	59.23
	AS	66.27	<u>61.13</u>	<u>60.03</u>	<u>63.33</u>	<u>59.15</u>	56.22	55.25	57.60	66.15	<u>50.58</u>	51.16	<u>52.63</u>	<u>57.78</u>	57.30	<u>58.53</u>	60.12
	CLAWS	<u>64.71</u>	61.29	63.95	66.93	56.54	56.08	<u>59.53</u>	<u>61.73</u>	68.43	51.28	51.04	50.94	56.14	56.45	58.49	<u>60.57</u>
XGBOOST	PPL	59.22	49.94	36.97	54.44	54.35	47.79	43.30	56.57	72.00	49.27	18.63	48.58	49.75	47.51	49.10	54.52
	WE	60.40	50.83	38.54	57.13	56.09	49.39	43.11	57.40	73.10	45.26	18.48	49.66	52.84	50.64	<u>50.98</u>	<u>58.15</u>
	LE	56.21	44.73	33.26	48.61	51.53	44.15	38.09	49.52	72.89	49.29	19.12	50.26	44.64	41.85	45.93	50.64
	HS	64.38	<u>58.48</u>	<u>43.79</u>	<u>60.96</u>	61.46	58.64	46.34	59.01	72.17	53.12	19.66	<u>50.69</u>	57.84	56.87	51.96	58.27
	AS	53.83	40.14	38.01	56.00	47.70	38.35	40.78	54.58	<u>73.07</u>	44.88	<u>19.58</u>	53.26	39.22	35.55	44.96	50.35
	CLAWS	<u>64.31</u>	58.82	44.59	62.34	<u>57.95</u>	<u>55.58</u>	<u>44.36</u>	<u>57.67</u>	70.94	<u>50.29</u>	19.47	48.47	<u>53.84</u>	<u>53.17</u>	47.38	53.54
TabM	PPL	59.36	49.44	41.76	60.30	56.70	50.27	<u>50.12</u>	63.76	72.67	49.88	20.65	56.42	51.83	49.60	56.57	64.01
	WE	60.40	50.83	36.82	55.76	56.09	49.39	42.67	55.17	73.10	45.26	18.97	50.41	52.84	50.64	50.67	56.85
	LE	53.83	40.14	33.18	51.29	47.70	38.35	35.34	47.56	<u>73.07</u>	44.88	18.84	52.45	39.22	35.55	45.74	<u>50.69</u>
	HS	<u>65.67</u>	<u>59.53</u>	46.96	61.59	62.04	58.94	51.10	<u>60.95</u>	72.72	51.80	<u>20.36</u>	51.17	57.68	<u>56.55</u>	<u>56.12</u>	59.23
	AS	53.83	40.14	42.08	<u>61.99</u>	47.70	38.35	42.90	57.50	<u>73.07</u>	44.88	20.29	52.54	39.22	35.55	49.02	56.70
	CLAWS	66.82	61.04	<u>46.92</u>	63.98	<u>60.38</u>	<u>57.71</u>	44.02	58.87	72.66	<u>51.27</u>	19.93	51.23	<u>57.67</u>	56.93	52.52	<u>59.28</u>

Table 12: Evaluation results for hallucination detection using OREAL-7B

Dataset		TEST				AMC				AIME				A(J)HSME			
Strategy	Method	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC
MLP	PPL	66.79	<u>66.76</u>	<u>69.52</u>	<u>73.18</u>	<u>72.67</u>	62.12	65.22	70.57	59.03	59.04	65.01	65.82	<u>76.35</u>	58.21	58.34	65.10
	WE	48.63	44.28	56.64	58.59	68.80	50.49	53.17	57.00	49.56	49.56	50.83	50.02	63.09	47.25	52.70	53.42
	LE	41.52	38.90	45.78	40.81	54.04	42.87	47.37	43.32	39.42	39.40	43.82	40.81	60.20	42.24	47.40	41.58
	HS	<u>67.34</u>	66.26	68.94	72.35	73.89	<u>59.54</u>	<u>61.08</u>	<u>68.01</u>	<u>52.22</u>	<u>52.24</u>	53.83	54.20	74.35	54.64	54.44	58.05
	AS	59.31	58.50	61.44	62.87	61.52	50.92	53.04	55.88	48.03	48.02	49.07	47.88	64.78	47.88	51.44	53.56
	CLAWS	69.13	68.11	73.19	74.80	68.24	57.50	58.89	65.89	51.41	51.43	<u>53.89</u>	<u>54.39</u>	77.81	<u>56.34</u>	<u>55.41</u>	<u>61.42</u>
XGBOOST	PPL	61.27	<u>59.63</u>	70.49	65.17	71.32	55.49	84.41	60.70	<u>53.64</u>	<u>53.62</u>	57.96	57.56	75.31	51.90	87.48	56.02
	WE	<u>61.82</u>	59.59	<u>71.87</u>	<u>69.07</u>	74.08	<u>56.51</u>	85.10	63.88	49.14	49.09	52.66	53.82	79.36	54.39	<u>87.86</u>	<u>60.02</u>
	LE	50.31	47.63	54.19	46.79	64.87	49.60	78.87	50.02	44.29	44.25	48.77	48.05	68.83	47.76	83.42	47.38
	HS	60.24	58.39	71.75	65.48	<u>73.30</u>	57.63	<u>86.31</u>	65.38	51.06	51.04	52.35	52.96	76.28	53.50	86.34	56.05
	AS	53.53	50.47	65.72	59.94	70.02	51.53	82.11	55.60	44.77	44.71	48.40	48.59	<u>76.44</u>	51.33	86.24	52.55
	CLAWS	65.68	64.26	76.73	71.49	70.33	55.37	87.57	<u>64.43</u>	54.20	54.19	<u>56.73</u>	<u>56.92</u>	74.70	<u>53.65</u>	89.57	61.67
TabM	PPL	54.59	51.04	77.93	74.38	<u>71.03</u>	46.78	88.28	70.25	42.73	42.66	64.11	65.78	77.71	46.49	90.16	65.19
	WE	44.28	39.16	66.10	63.23	70.53	45.80	82.92	59.29	35.37	35.28	50.26	50.65	<u>77.97</u>	47.21	85.75	54.04
	LE	42.61	38.82	56.70	42.63	62.39	45.80	77.26	44.41	38.08	38.04	47.82	40.96	66.08	44.04	81.65	41.85
	HS	63.90	62.04	76.98	<u>72.73</u>	73.50	<u>55.38</u>	<u>87.61</u>	<u>68.01</u>	<u>51.06</u>	<u>51.04</u>	53.91	54.20	77.91	<u>53.08</u>	88.03	58.05
	AS	41.98	36.51	66.28	57.65	69.89	44.17	82.10	55.16	33.49	33.40	48.82	48.13	<u>77.56</u>	45.83	86.34	53.51
	CLAWS	<u>63.38</u>	<u>61.74</u>	<u>77.83</u>	71.17	69.70	55.45	87.38	63.99	53.38	53.38	<u>56.58</u>	<u>56.14</u>	78.89	53.44	<u>89.54</u>	<u>61.99</u>

Table 13: Evaluation results for hallucination detection using Qwen-2.5-Math-7B

		TEST				AMC				AIME				A(J)HSME			
Model	Method	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC
Deepseek	PPL	<u>35.22</u>	<u>35.22</u>	34.40	52.03	37.57	37.57	<u>35.04</u>	<u>53.16</u>	42.57	42.57	37.65	57.34	<u>36.21</u>	<u>36.21</u>	<u>34.83</u>	<u>52.30</u>
	WE	30.54	30.54	<u>35.34</u>	<u>53.44</u>	28.46	28.46	34.20	51.60	25.80	25.80	33.15	49.40	28.70	28.70	34.35	51.95
	LE	31.89	31.89	32.98	48.91	32.78	32.78	33.35	49.95	29.32	29.32	32.65	47.62	33.06	33.06	33.37	49.94
	HS	27.36	27.36	32.21	45.16	31.40	31.40	33.25	49.58	32.09	32.09	33.14	49.21	<u>32.88</u>	<u>32.88</u>	<u>33.58</u>	<u>50.53</u>
	AS	31.48	31.48	33.81	49.69	32.11	32.11	33.31	49.91	30.56	30.56	33.56	50.40	33.29	33.29	33.55	50.47
	CLAWS	46.30	46.30	41.34	62.03	<u>35.90</u>	<u>35.90</u>	35.87	54.62	<u>36.93</u>	<u>36.93</u>	<u>36.66</u>	<u>55.95</u>	36.43	36.43	35.97	54.89
Mathstral	PPL	29.11	29.11	32.86	48.70	32.37	32.37	33.39	49.88	27.67	27.67	32.46	47.79	31.55	31.55	33.04	49.29
	WE	<u>38.76</u>	<u>38.76</u>	<u>36.14</u>	<u>54.71</u>	<u>36.23</u>	<u>36.23</u>	<u>35.05</u>	<u>53.11</u>	34.21	34.21	<u>33.81</u>	<u>50.74</u>	<u>34.19</u>	<u>34.19</u>	<u>34.41</u>	<u>52.05</u>
	LE	30.60	30.60	32.83	48.54	32.05	32.05	32.99	49.08	26.00	26.00	31.56	44.85	30.59	30.59	32.62	48.09
	HS	27.80	27.80	33.82	49.35	26.85	26.85	33.12	48.96	<u>19.96</u>	<u>19.96</u>	<u>31.70</u>	<u>45.22</u>	25.67	25.67	32.73	48.38
	AS	24.86	24.86	31.88	44.16	28.19	28.19	32.16	46.26	28.62	28.62	32.91	48.53	29.09	29.09	32.16	46.82
	CLAWS	42.50	42.50	40.40	60.71	38.13	38.13	37.08	56.45	<u>31.86</u>	<u>31.86</u>	34.23	51.84	38.04	38.04	37.05	56.43
OpenMath2	PPL	29.78	29.78	33.01	49.23	27.45	27.45	32.15	46.56	25.40	25.40	31.73	44.40	23.79	23.79	31.37	43.75
	WE	33.85	33.85	34.18	51.55	33.45	33.45	34.29	51.89	31.01	31.01	32.73	48.51	29.53	29.53	33.14	49.50
	LE	<u>36.34</u>	<u>36.34</u>	<u>34.53</u>	<u>52.32</u>	40.00	40.00	<u>36.16</u>	<u>55.15</u>	31.18	31.18	33.42	50.00	38.06	38.06	<u>35.53</u>	<u>53.93</u>
	HS	25.49	25.49	31.53	44.33	28.34	28.34	32.25	46.91	<u>36.30</u>	<u>36.30</u>	<u>34.92</u>	<u>52.61</u>	28.43	28.43	32.30	47.38
	AS	23.92	23.92	31.19	43.04	29.84	29.84	32.75	48.20	38.59	38.59	35.52	54.10	32.32	32.32	33.77	50.30
	CLAWS	41.90	41.90	38.92	58.51	<u>37.66</u>	<u>37.66</u>	36.93	56.36	24.86	24.86	33.22	49.63	<u>33.47</u>	<u>33.47</u>	35.60	54.23
OREAL	PPL	29.02	29.02	32.41	47.47	23.55	23.55	31.56	44.09	31.64	31.64	33.65	50.00	23.87	23.87	31.48	44.25
	WE	25.69	25.69	32.14	46.91	27.60	27.60	33.08	49.37	30.21	30.21	33.34	50.00	27.38	27.38	33.00	49.07
	LE	33.34	33.34	33.86	50.84	34.64	34.64	<u>34.96</u>	<u>53.16</u>	29.33	29.33	33.37	49.47	<u>35.79</u>	<u>35.79</u>	<u>35.33</u>	<u>53.88</u>
	HS	<u>30.03</u>	<u>30.03</u>	32.87	48.60	26.77	26.77	31.91	45.89	<u>33.93</u>	<u>33.93</u>	<u>33.76</u>	<u>50.53</u>	27.64	27.64	32.10	46.58
	AS	26.10	26.10	<u>33.10</u>	47.75	31.65	31.65	34.15	51.58	25.07	25.07	33.61	<u>50.53</u>	30.21	30.21	33.59	49.22
	CLAWS	25.27	25.27	32.99	48.31	<u>34.08</u>	<u>34.08</u>	35.16	53.48	34.85	34.85	34.69	52.66	37.49	37.49	35.52	54.04
Qwen-2.5	PPL	27.04	27.04	34.14	50.00	27.75	27.75	33.72	49.52	25.76	25.76	33.34	49.53	35.59	31.09	33.92	50.47
	WE	<u>34.62</u>	<u>34.62</u>	<u>34.91</u>	<u>52.96</u>	32.83	32.83	34.20	51.39	31.12	31.12	33.01	49.06	29.08	28.56	33.01	49.11
	LE	45.25	45.25	39.53	59.24	<u>40.54</u>	<u>40.54</u>	<u>36.60</u>	<u>55.60</u>	39.56	39.56	36.05	54.56	41.31	39.10	35.66	54.32
	HS	27.84	27.84	34.67	51.72	30.39	30.39	35.66	53.58	18.97	18.97	33.48	50.31	36.80	31.55	35.08	53.11
	AS	26.88	26.88	32.66	47.17	31.59	31.59	34.13	51.27	<u>32.32</u>	<u>32.32</u>	<u>34.20</u>	<u>51.73</u>	37.59	34.97	34.28	51.52
	CLAWS	31.34	31.34	33.39	49.63	40.88	40.88	38.04	57.63	23.76	23.76	31.66	45.28	<u>38.27</u>	<u>36.63</u>	<u>35.56</u>	<u>53.95</u>

Table 14: Evaluation results for 3-class detection on a balanced dataset. The evaluation strategies used are Threshold and Prototype. Bold values indicate the best performance, underlined values indicate the second-best, and gray-shaded cells correspond to results where the model performed detection over only 2 out of the 3 target classes.

Dataset		TEST				AMC				AIME				A(J)HSME			
Strategy	Method	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC
MLP	PPL	30.04	30.04	38.58	57.19	23.69	23.69	<u>38.57</u>	56.47	30.73	30.73	<u>40.75</u>	60.01	27.11	27.11	38.43	56.75
	WE	33.76	33.76	40.38	60.54	32.25	32.25	35.88	54.05	31.12	31.12	36.99	55.19	31.64	31.64	36.92	56.42
	LE	35.27	35.27	36.45	54.32	26.92	26.92	34.54	51.59	29.76	29.76	36.82	54.31	<u>34.95</u>	<u>34.95</u>	34.99	51.25
	HS	<u>41.43</u>	<u>41.43</u>	<u>44.90</u>	<u>62.33</u>	27.66	27.66	34.81	51.37	28.28	28.28	36.62	53.21	28.84	28.84	33.96	49.65
	AS	38.64	38.64	41.96	60.54	<u>34.57</u>	<u>34.57</u>	34.69	51.74	<u>32.82</u>	<u>32.82</u>	35.22	51.83	33.22	33.22	33.77	50.43
	CLAWS	44.98	44.98	49.40	67.35	41.38	41.38	43.04	60.77	41.75	41.75	42.96	62.88	35.00	35.00	42.14	60.25
XGBOOST	PPL	39.12	39.12	37.18	55.02	35.85	35.85	35.79	53.32	36.68	36.68	35.70	52.40	35.90	35.90	35.24	52.73
	WE	33.20	33.20	40.18	<u>60.71</u>	30.24	30.24	<u>37.15</u>	56.72	22.32	22.32	34.51	51.56	31.12	31.12	<u>36.72</u>	56.17
	LE	36.50	36.50	35.10	52.32	34.29	34.29	34.37	51.49	32.17	32.17	33.29	49.80	33.12	33.12	34.87	52.15
	HS	36.52	36.52	37.35	54.52	35.71	35.71	34.28	51.47	35.93	35.93	37.00	<u>54.95</u>	35.13	35.13	35.19	51.92
	AS	<u>41.05</u>	<u>41.05</u>	<u>41.24</u>	59.10	34.01	34.01	34.79	51.88	31.78	31.78	<u>37.18</u>	53.95	32.39	32.39	33.68	50.52
	CLAWS	43.24	43.24	46.66	63.42	37.50	37.50	37.80	<u>55.29</u>	42.11	42.11	41.87	59.88	<u>35.53</u>	<u>35.53</u>	37.34	<u>54.77</u>
TabM	PPL	34.50	34.50	38.91	57.59	33.63	33.63	38.62	<u>56.63</u>	33.89	33.89	41.36	59.97	<u>36.75</u>	<u>36.75</u>	38.43	56.73
	WE	30.20	30.20	40.38	60.66	29.00	29.00	37.21	56.72	25.12	25.12	36.46	54.90	29.53	29.53	36.91	56.41
	LE	37.40	37.40	35.83	54.04	33.79	33.79	34.81	51.50	<u>39.61</u>	<u>39.61</u>	37.83	55.14	35.52	35.52	35.20	52.10
	HS	40.49	40.49	<u>44.16</u>	<u>60.99</u>	32.64	32.64	35.74	52.89	34.37	34.37	35.28	52.06	30.71	30.71	33.73	49.32
	AS	<u>41.17</u>	<u>41.17</u>	42.68	60.83	<u>34.17</u>	<u>34.17</u>	34.72	51.63	32.61	32.61	37.28	54.15	32.51	32.51	33.70	50.42
	CLAWS	46.42	46.42	47.89	64.08	37.89	37.89	38.87	56.16	42.70	42.70	45.24	61.62	38.92	38.92	40.27	58.51

Table 15: Evaluation results for 3-class detection using DeepSeek-Math-7B on a balanced dataset. Bold values indicate the best performance, underlined values indicate the second-best. Light gray-shaded cells correspond to cases where the model performed detection over only 2 out of the 3 target classes, while dark gray-shaded cells indicate cases where the model predicted only 1 out of the 3 target classes.

Dataset		TEST				AMC				AIME				A(J)HSME			
Strategy	Method	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC
MLP	PPL	29.20	29.20	41.65	59.58	16.67	16.67	35.93	52.16	16.48	16.48	36.69	<u>53.09</u>	27.26	27.26	35.73	52.53
	WE	27.15	27.15	36.88	51.56	16.67	16.67	35.44	52.37	22.70	22.70	31.40	45.42	19.64	19.64	33.94	49.79
	LE	37.59	37.59	37.39	54.32	30.70	30.70	35.27	53.14	<u>32.48</u>	<u>32.48</u>	<u>37.23</u>	52.87	16.67	16.67	33.33	50.00
	HS	<u>45.58</u>	<u>45.58</u>	<u>43.92</u>	<u>62.70</u>	29.61	29.61	<u>37.66</u>	<u>55.65</u>	19.92	19.92	36.00	52.67	<u>29.42</u>	<u>29.42</u>	<u>35.91</u>	53.20
	AS	38.49	38.49	42.73	62.29	<u>36.60</u>	<u>36.60</u>	36.38	54.85	31.81	31.81	35.14	52.07	25.42	25.42	35.28	<u>53.28</u>
	CLAWS	46.78	46.78	45.64	63.92	44.79	44.79	44.69	62.51	33.72	33.72	43.75	58.92	39.02	39.02	41.95	60.36
XGBOOST	PPL	33.41	33.41	34.60	51.21	35.84	35.84	34.78	51.49	<u>38.15</u>	<u>38.15</u>	<u>39.75</u>	55.00	34.06	34.06	34.72	51.35
	WE	34.18	34.18	37.59	56.00	33.95	33.95	<u>36.71</u>	<u>56.02</u>	32.49	32.49	33.80	50.17	29.30	29.30	<u>37.49</u>	<u>56.35</u>
	LE	29.74	29.74	32.30	46.74	34.67	34.67	35.20	52.35	30.95	30.95	35.16	52.07	33.39	33.39	33.54	50.56
	HS	<u>37.82</u>	<u>37.82</u>	<u>38.99</u>	<u>56.58</u>	36.03	36.03	35.64	52.11	32.52	32.52	34.83	49.69	33.65	33.65	34.11	50.75
	AS	36.54	36.54	37.89	54.51	<u>36.50</u>	<u>36.50</u>	36.47	53.84	31.42	31.42	31.34	45.44	<u>34.36</u>	<u>34.36</u>	35.45	52.44
	CLAWS	40.20	40.20	45.46	60.50	40.14	40.14	40.23	57.46	42.73	42.73	41.03	<u>54.54</u>	34.60	34.60	38.38	56.42
TabM	PPL	<u>42.24</u>	<u>42.24</u>	38.69	55.69	36.79	36.79	35.46	51.75	<u>37.77</u>	<u>37.77</u>	36.86	<u>53.20</u>	34.14	34.14	35.34	52.11
	WE	31.48	31.48	36.95	51.88	30.15	30.15	35.69	52.18	31.55	31.55	32.46	47.29	22.68	22.68	33.75	49.64
	LE	28.89	28.89	36.88	53.46	29.67	29.67	35.14	52.70	26.54	26.54	<u>37.27</u>	52.38	28.46	28.46	34.87	52.11
	HS	34.64	34.64	39.16	<u>58.56</u>	31.46	31.46	37.67	55.25	20.70	20.70	33.21	47.77	28.15	28.15	35.51	52.67
	AS	38.21	38.21	40.78	58.34	<u>36.79</u>	<u>36.79</u>	37.88	55.80	31.80	31.80	34.94	50.89	35.48	35.48	35.89	53.82
	CLAWS	45.74	45.74	48.76	66.40	41.22	41.22	43.57	61.17	38.58	38.58	44.41	60.86	38.99	38.99	40.32	57.89

Table 16: Evaluation results for 3-class detection using Mathstral-7B on a balanced dataset

Dataset		TEST				AMC				AIME				A(J)HSME			
Strategy	Method	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC
MLP	PPL	29.07	29.07	35.24	50.63	<u>29.70</u>	<u>29.70</u>	33.56	49.09	23.99	23.99	34.08	48.25	<u>31.77</u>	<u>31.77</u>	<u>42.10</u>	<u>59.40</u>
	WE	32.54	32.54	34.84	50.68	28.22	28.22	34.47	50.89	16.67	16.67	34.71	51.53	31.38	31.38	36.24	53.50
	LE	31.28	31.28	34.74	49.56	25.55	25.55	32.21	47.23	18.22	18.22	30.51	43.57	28.08	28.08	35.97	52.38
	HS	33.89	33.89	<u>43.40</u>	<u>61.26</u>	29.16	29.16	<u>39.11</u>	<u>56.93</u>	34.24	34.24	<u>37.61</u>	<u>51.56</u>	28.82	28.82	36.98	54.43
	AS	<u>40.56</u>	<u>40.56</u>	41.38	60.35	27.80	27.80	36.07	52.92	27.55	27.55	31.38	46.84	23.34	23.34	30.92	45.34
	CLAWS	49.51	49.51	49.09	68.00	43.34	43.34	46.49	64.34	<u>27.91</u>	<u>27.91</u>	43.29	60.89	43.42	43.42	43.28	60.58
XGBOOST	PPL	31.64	31.64	34.48	50.89	29.55	29.55	32.14	48.11	<u>35.35</u>	<u>35.35</u>	38.34	<u>53.75</u>	<u>35.58</u>	<u>35.58</u>	<u>37.00</u>	53.21
	WE	<u>37.97</u>	<u>37.97</u>	<u>39.76</u>	<u>57.89</u>	34.03	34.03	35.96	54.12	34.70	34.70	35.60	51.49	29.54	29.54	35.44	<u>53.48</u>
	LE	31.75	31.75	34.09	49.66	33.41	33.41	34.72	51.37	32.31	32.31	35.01	51.03	33.04	33.04	34.44	50.80
	HS	32.30	32.30	35.72	53.62	<u>37.06</u>	<u>37.06</u>	<u>38.54</u>	<u>55.35</u>	32.31	32.31	35.01	51.03	33.38	33.38	35.13	52.06
	AS	36.01	36.01	37.54	54.58	35.51	35.51	34.71	51.48	31.35	31.35	31.95	47.58	33.35	33.35	33.55	51.22
	CLAWS	49.96	49.96	48.76	66.23	42.00	42.00	42.66	60.32	36.91	36.91	<u>37.91</u>	54.72	40.50	40.50	42.87	60.46
TabM	PPL	23.49	23.49	33.31	47.71	24.11	24.11	32.64	47.23	21.31	21.31	33.87	46.70	23.14	23.14	34.57	49.12
	WE	36.25	36.25	37.61	55.32	<u>34.02</u>	<u>34.02</u>	35.06	52.59	<u>31.37</u>	<u>31.37</u>	35.68	<u>53.01</u>	<u>36.77</u>	<u>36.77</u>	35.67	52.81
	LE	32.59	32.59	34.74	50.21	32.19	32.19	35.17	50.86	21.50	21.50	28.69	40.00	32.75	32.75	32.78	49.01
	HS	38.79	38.79	<u>42.21</u>	<u>60.36</u>	30.87	30.87	<u>37.02</u>	<u>54.82</u>	30.62	30.62	<u>36.79</u>	50.53	30.88	30.88	<u>36.98</u>	<u>53.80</u>
	AS	<u>40.56</u>	<u>40.56</u>	41.36	59.98	33.06	33.06	35.25	51.71	29.54	29.54	31.96	47.37	32.42	32.42	33.62	49.69
	CLAWS	45.87	45.87	50.47	69.09	41.45	41.45	47.00	63.95	41.43	41.43	44.73	59.39	40.84	40.84	44.47	61.44

Table 17: Evaluation results for 3-class detection using OpenMath2-LLaMA3.1-8B on a balanced dataset

Dataset		TEST				AMC				AIME				A(J)HSME			
Strategy	Method	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC
MLP	PPL	<u>30.55</u>	<u>30.55</u>	39.80	56.69	37.16	37.16	41.26	59.01	22.47	22.47	41.00	<u>54.61</u>	40.38	40.38	<u>40.95</u>	56.56
	WE	16.71	16.71	32.98	49.07	18.17	18.17	34.96	52.38	26.00	26.00	35.41	52.23	20.66	20.66	34.70	50.55
	LE	25.62	25.62	37.46	<u>55.19</u>	16.67	16.67	36.41	<u>53.56</u>	<u>27.58</u>	<u>27.58</u>	37.50	54.35	16.67	16.67	33.09	48.46
	HS	26.31	26.31	34.00	50.79	<u>33.04</u>	<u>33.04</u>	<u>39.84</u>	<u>57.51</u>	24.03	24.03	41.83	55.60	35.34	35.34	38.18	56.24
	AS	<u>16.67</u>	<u>16.67</u>	34.39	49.49	29.69	29.69	35.08	51.92	23.90	23.90	32.64	46.50	<u>34.91</u>	<u>34.91</u>	41.78	60.46
	CLAWS	37.54	37.54	<u>38.59</u>	54.61	27.03	27.03	38.90	55.98	30.64	30.64	33.68	47.74	<u>39.49</u>	<u>39.49</u>	<u>40.95</u>	56.25
XGBOOST	PPL	<u>33.63</u>	<u>33.63</u>	<u>34.65</u>	<u>51.86</u>	38.34	38.34	37.58	54.99	36.88	36.88	34.53	50.26	35.93	35.93	37.37	53.25
	WE	24.25	24.25	33.50	49.75	25.34	25.34	36.10	53.91	28.33	28.33	35.06	52.47	25.40	25.40	<u>36.88</u>	54.24
	LE	33.58	33.58	34.26	50.43	32.49	32.49	33.80	50.07	36.88	36.88	<u>36.89</u>	49.59	32.76	32.76	34.25	51.18
	HS	32.32	32.32	33.35	48.58	<u>35.14</u>	<u>35.14</u>	<u>36.56</u>	<u>54.71</u>	28.70	28.70	34.12	50.62	35.73	35.73	35.48	52.58
	AS	27.82	27.82	33.13	48.70	33.44	33.44	34.73	51.62	<u>36.34</u>	<u>36.34</u>	34.41	<u>50.94</u>	29.16	29.16	34.10	50.26
	CLAWS	39.05	39.05	40.73	59.03	34.99	34.99	36.47	53.69	34.14	34.14	36.99	50.62	<u>35.84</u>	<u>35.84</u>	36.48	<u>53.33</u>
TabM	PPL	36.95	36.95	42.66	59.17	38.03	38.03	40.84	57.92	37.60	37.60	40.37	54.71	39.07	39.07	42.25	57.69
	WE	21.91	21.91	32.71	47.72	22.22	22.22	33.44	50.59	25.16	25.16	33.97	50.85	20.66	20.66	34.26	51.71
	LE	32.76	32.76	34.28	50.15	33.43	33.43	34.97	52.29	<u>34.93</u>	<u>34.93</u>	<u>38.28</u>	52.85	<u>36.14</u>	<u>36.14</u>	37.31	55.85
	HS	27.86	27.86	34.51	50.49	<u>36.48</u>	<u>36.48</u>	<u>39.10</u>	<u>56.42</u>	31.06	31.06	36.33	52.99	34.38	34.38	<u>38.41</u>	<u>56.11</u>
	AS	31.94	31.94	35.29	51.14	34.14	34.14	35.55	52.82	33.61	33.61	34.50	50.43	29.68	29.68	37.19	55.81
	CLAWS	44.21	44.21	44.09	62.58	34.23	34.23	37.20	53.25	29.62	29.62	33.35	46.67	35.21	35.21	36.83	53.68

Table 18: Evaluation results for 3-class detection using OREAL-7B on a balanced dataset

Dataset		TEST				AMC				AIME				A(J)HSME			
Strategy	Method	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC	F1 _w	F1 _m	AP _m	AUROC
MLP	PPL	<u>43.21</u>	<u>43.21</u>	48.48	67.07	<u>38.35</u>	<u>38.35</u>	45.18	63.38	36.12	36.12	41.57	60.18	<u>37.57</u>	38.13	<u>40.67</u>	<u>58.45</u>
	WE	23.73	23.73	35.91	52.21	35.18	35.18	35.83	52.69	32.29	32.29	35.18	51.72	32.97	32.69	34.81	50.79
	LE	20.88	20.88	29.52	40.91	27.44	27.44	34.87	50.93	22.47	22.47	30.83	44.62	27.76	26.99	32.09	46.89
	HS	42.56	42.56	45.23	62.86	31.55	31.55	40.19	59.28	32.19	32.19	35.77	52.09	33.65	32.88	38.15	54.69
	AS	39.11	39.11	37.77	56.80	26.66	26.66	36.45	54.02	26.01	26.01	31.89	47.33	34.46	33.71	34.39	51.12
	CLAWS	43.56	43.56	<u>46.59</u>	<u>64.77</u>	38.36	38.36	<u>41.98</u>	<u>61.53</u>	<u>33.57</u>	<u>33.57</u>	<u>37.93</u>	<u>55.03</u>	38.59	<u>35.12</u>	41.92	60.47
XGBOOST	PPL	40.16	<u>40.16</u>	43.09	60.01	37.57	37.57	38.43	55.33	37.99	37.99	39.15	56.60	36.22	35.16	36.07	52.68
	WE	<u>41.38</u>	41.38	40.81	58.87	37.89	37.89	38.51	56.66	<u>35.50</u>	<u>35.50</u>	36.17	52.77	37.33	36.20	<u>36.72</u>	<u>54.31</u>
	LE	33.93	33.93	33.71	51.48	32.65	32.65	34.00	50.92	29.51	29.51	32.34	48.28	34.55	33.51	34.05	50.89
	HS	39.46	39.46	<u>40.97</u>	<u>59.22</u>	<u>38.29</u>	<u>38.29</u>	<u>39.60</u>	<u>57.37</u>	33.19	33.19	33.33	49.91	<u>38.94</u>	37.34	36.09	53.49
	AS	36.63	36.63	37.97	54.72	38.21	38.21	35.99	53.15	30.86	30.86	33.03	47.91	38.05	35.96	34.76	52.47
	CLAWS	42.08	36.63	37.97	54.72	39.02	39.02	39.92	58.99	35.24	35.24	<u>37.61</u>	<u>54.24</u>	39.06	<u>37.01</u>	38.00	57.28
TabM	PPL	40.13	40.13	47.62	66.36	<u>39.52</u>	<u>39.52</u>	43.87	62.15	<u>34.74</u>	<u>34.74</u>	41.71	59.99	35.79	<u>36.55</u>	39.37	<u>57.42</u>
	WE	38.23	38.23	38.37	55.63	35.77	35.77	37.37	54.90	30.03	30.03	35.10	51.88	32.93	32.90	35.90	52.83
	LE	26.16	26.16	30.58	46.48	29.14	29.14	32.54	50.05	27.55	27.55	31.08	46.65	28.17	27.31	32.96	48.78
	HS	<u>40.42</u>	<u>40.42</u>	<u>44.41</u>	61.94	38.08	38.08	<u>41.85</u>	59.30	34.03	34.03	35.72	52.50	<u>36.22</u>	34.52	36.96	53.35
	AS	36.75	36.75	38.01	56.30	36.24	36.24	36.48	53.82	29.81	29.81	33.21	48.37	34.77	33.98	34.55	51.32
	CLAWS	41.14	41.14	44.03	<u>61.98</u>	40.16	40.16	40.10	<u>59.94</u>	36.52	36.52	<u>37.83</u>	<u>54.64</u>	40.03	37.98	<u>38.50</u>	57.69

Table 19: Evaluation results for 3-class detection using Qwen-2.5-Math-7B on a balanced dataset