Beyond Labels: Explanatory Collapse due to Instruction Tuning in Protein LLMs

Yining Yang, Ruihong Huang & Yang Shen *
Department of Computer Science and Engineering
Texas A&M University
College Station, TX 77840, USA
{yining_yang, huangrh, yshen}@tamu.edu

Abstract

Instruction tuning on domain-specific datasets improves categorical accuracy but consistently degrades explanatory behavior. In protein large language models, we observe that tuned models ignore explanation requests, fall back on generic templates, and produce degenerate structured outputs with trivial repetitions or hallucinated lists. This collapse of expressive diversity renders models terse and uninformative, limiting their scientific utility. Our findings highlight a trade-off in current protein instruction-tuning practices: accuracy is gained at the cost of interpretive value, underscoring the need for strategies that preserve explanatory depth.

1 Introduction

Large language models (LLMs) have recently been adapted to biological domains, including protein understanding tasks such as enzyme classification, Gene Ontology (GO) annotation, and functional description [19, 2, 27]. A common approach is *instruction tuning*, where general-purpose LLMs are fine-tuned on domain-specific instruction–response pairs to improve alignment with supervised tasks. While effective at improving label accuracy, such fine-tuning may inadvertently degrade other capacities of the base model.

In this work, we study the consequences of instruction tuning on explanatory behavior in protein LLMs. Using the OPI dataset [27], we have the fine-tuned Llama-3.1-Instruct on 1.6M protein instruction samples and compare its outputs against the pretrained baseline. Our analysis reveals a systematic phenomenon of *explanatory collapse*: the tuned model achieves higher accuracy on structured predictions but loses the general ability to produce explanatory rationales when explicitly requested. We document this collapse empirically, analyze its underlying causes, and discuss implications for protein science.

2 Related Work

2.1 Instruction tuning and loss of reasoning

Instruction tuning has been widely studied in NLP [15, 22, 12]. While effective for aligning with downstream tasks, it is known to induce distributional contraction: models converge to short, low-entropy responses and suppress hedging or explanatory variation [18, 9]. Similar issues have been reported in reasoning tasks, where fine-tuned models achieve higher accuracy but fail to produce

*Corresponding author

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: 2nd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences.

step-by-step rationales [24]. Lobo et al.[11] show that fine-tuning reduces the faithfulness of chain-of-thought reasoning, with models producing correct answers but less reliable intermediate rationales. Chen et al. [3] similarly find that reinforcement learning from human feedback (RLHF) can incentivize models to hide undesirable reasoning traces, yielding polished answers without transparent logic. [20] provide a comprehensive analysis showing that full instruction tuning often leads to pattern copying and hallucination, while diminishing factual reliability. Together, these studies highlight a fundamental tension: instruction tuning improves task accuracy but risks narrowing explanatory output.

2.2 Biomedical and protein language models

Domain-specific LLMs for biomedicine have demonstrated strong performance but limited explanatory capacity. BioGPT [13] and Galactica [23] achieved state-of-the-art results on biomedical benchmarks, yet their fine-tuned variants produced primarily short labels or direct answers, rather than step-by-step reasoning. Med-PaLM [21] explicitly addressed this gap: by tuning on medical Q&A with detailed rationales, the model produced more comprehensive explanations, which evaluators judged closer to clinical consensus. In protein language modeling, emerging efforts such as OPI [27] have prioritized structured labels (EC, GO, keywords), but few works systematically evaluate whether explanatory behaviors are preserved.

2.3 Rationale-augmented instruction tuning

A growing line of work explores how to preserve explanations during fine-tuning. LogiCoT [10] distills chain-of-thought reasoning into smaller models via rationale-augmented datasets. Re-Critic [28] introduces self-critique with rationales to reduce hallucinations in multimodal tasks. Divergent CoT [17] trains models to generate multiple reasoning paths and refine their answers. Methods like PINTO [25] further optimize for faithful rationales, penalizing explanations that do not support the answer. These approaches demonstrate that explicit rationale supervision can mitigate the collapse of explanatory output.

3 Empirical Evidence of Explanatory Collapse

We compared the pretrained Llama-3.1-Instruct base model against its OPI-tuned counterpart across four predictive tasks: Enzyme Commission (EC) number, GO term, UniProt keyword (KW), and functional description (Func.) [2]. While OPI tuning improved categorical accuracy, it simultaneously induced collapse of explanatory behavior. We identify three recurring empirical patterns.

3.1 Lack of response to Explanatory Prompts

When prompted with instructions such as *"predict XX and explain why"*, models can either produce both the categorical label and a rationale, or ignore the explanatory clause and return only the label. We quantify this behavior using the Explanation Compliance Rate (ECR) and Average Explanatory Portion(AEP). The ECR is defined as the proportion of outputs that include substantive justification when an explanation is explicitly requested, while the AEP reviews proportion of output tokens that are part of qualifying explanatory sentences, averaged over all samples.

Explanation Compliance Rate (ECR) and Average Explanatory Portion (AEP) measures how often and how much explanation is made in the answers. We define ECR as the proportion of answers that contain a substantive natural-language justification. An answer is regarded explanation-compliant if it contains at least one sentence with (1) 8 tokens, (2) at least one explanatory cue word (from a fixed list: because, since, due to, suggest, indicate, likely, motif, domain, residue, active site, thus, therefore), and (3) at least 4 non-trivial content tokens (excluding predicted labels such as EC/GO/KW terms). The sentence must occur after the main label prediction and provide interpretable reasoning. We define AEP as the proportion of output tokens that are part of explanation-compliant sentences, averaged over all sampled answers for each task.

In the base model, ECR was relatively high ($\sim$60%), reflecting the tendency of LLMs to over-generate rationales. These rationales often cited motifs, domains, or mechanistic cues, but were frequently

[2]Inference performed with A100 GPUs

verbose or hallucinatory. After OPI tuning, outputs overwhelmingly ignored the explanatory clause and produced only the label, driving ECR below 10% on most classification tasks. Table 1 shows that after tuning, EC, GO terms and keyword tasks yielded only label outputs. Function descriptions were the only setting where explanations remained common. Notably, explicitly requesting explanation further reduced coverage, reflecting a strong bias toward concise label-style answers.

Table 1: Assessment of explanatory compliance for baseline and instruction-tuned models (without and with an explicit explanatory prompt (EP)) across protein understanding tasks. Reported are ECR / AEP values.

Task	Base w/o EP	Tuned w/o EP.	Tuned w EP.
EC	64.8%/ 58.4%	0%/0%	0%/0%
GO	69.7%/ 69.5%	0.7%/0.5%	0.8%/0.7%
KW	42.4%/ 42.9%	1.1%/1.1%	0.9%/1.0%
Func.	50.2%/ 49.5%	10.0%/9.9%	10.6%/10.3%

This pattern highlights a key trade-off introduced by OPI tuning. While alignment improves structured label prediction, it simultaneously suppresses the generative tendency to provide justifications, even when explicitly instructed. Moreover, we have also trained a Llava-style model which has sequence encoders as well as its MLP adaptor. The new model architecture after two stage training still exhibit similar explanatory collapse behavior. In practice, the tuned model is robust at delivering categorical predictions but exhibits poor responsiveness to explanatory prompting.

3.2 Rigidity and degeneracy across tasks

Instruction tuning also induced rigid and degenerate behaviors:

- **Functional description:** The base-model outputs varied in style and sometimes resembled UniProt entries. The tuned-model outputs collapsed to a handful of generic templates, e.g. *"This protein is a G-protein coupled receptor family member"*, regardless of the input. Lexical diversity (Distinct-2) decreased from 0.58 to 0.09.
- **GO terms:** Predictions collapsed to trivial, high-frequency categories such as *protein binding*, reducing semantic coverage by $\sim$65%.
- **Keyword tagging:** Outputs degenerated into repeated or hallucinated keywords, with length increasing 3.4$\times$ but unchanged F1.
- **EC classification:** the Exact-match accuracy improved (0.03 $\rightarrow$ 0.35), but the explanatory efforts disappeared.

These symptoms reflect a contraction of the output distribution: cross-entropy training against short label-only targets penalizes rationales and reduces entropy, leading to uniformly short, rigid, and non-explanatory outputs.

3.3 Consequences for protein understanding

Although structured accuracy improves, explanatory collapse undermines interpretability. Protein scientists require rationales—such as motifs, domains, catalytic residues, or evolutionary signatures—to evaluate plausibility and guide discovery. A model that provides only bare EC or GO labels acts as a black-box annotator, limiting its scientific utility. Thus, instruction tuning introduces a trade-off: accuracy gains come at the cost of epistemic richness.

4 Discussion

The preceding analysis shows that instruction tuning on OPI improves categorical accuracy but suppresses explanatory behavior. In this section, we examine why collapse arises, why it is especially problematic for protein sequence understanding, how current evaluation metrics contribute to the issue, and what directions may help address it.

4.1 Why collapse occurs

The loss of explanatory behavior reflects a general property of supervised fine-tuning with label-only targets. Under the cross-entropy objective $\mathcal{L} = -\log p_\theta(y^* \mid x)$, where y^* is a short categorical label

(e.g., an EC number or a group of keywords), optimization drives the model toward a delta distribution centered on y^*. This encourages mode-seeking behavior, suppressing alternative completions and reducing entropy in the output distribution. In practice, fine-tuning on label-only samples biases the model toward terse classification outputs and weakens its capacity to generate explanations, even when explicitly requested. Pretrained models often attempt rationales—sometimes noisy or inconsistent—while tuned models converge to minimal, label-only completions.

Self-supervised pretraining (e.g., masked language modeling, autoregressive next-token prediction) does not impose such deterministic collapse. These objectives maintain higher-entropy conditional distributions, enabling models to generate exploratory rationales. Yet pretraining alone cannot achieve reliable task alignment: outputs remain verbose and uncalibrated. The collapse we observe therefore reflects a broader tension between supervised and self-supervised paradigms.

4.2 Scientific implications of explanatory collapse

For protein sequence understanding, the suppression of explanatory behavior is particularly limiting:

- *Labels are not explanations.* EC numbers, GO terms, or keywords are taxonomic identifiers. Their value lies in the reasoning that connects a sequence to a functional role. Without supporting evidence, predictions remain opaque.
- *Sequence–structure–function reasoning is expected.* Protein scientists interpret predictions through motifs, catalytic residues, and domain structures. A model that fails to articulate such links cannot integrate into existing workflows for protein research.
- *Trust requires falsifiability.* Labels alone often cannot be directly verified. Explanations (e.g., "contains a Walker A motif") provide concrete claims that can be tested or refuted, serving as a safeguard against blind trust.
- *Integration into experiments demands evidence.* Functional predictions often guide costly wet-lab validation. Without mechanistic hypotheses, outputs lack sufficient basis for prioritization in experiments.
- *Ontology richness is underused.* Protein ontologies such as GO and EC are hierarchical, but collapsed models gravitate to generic or trivial categories, suppressing nuanced, multi-level interpretations.

Thus, the explanatory collapse is not only an interpretability issue but a scientific limitation: it undermines the plausibility, transparency, trustworthiness, and utility of protein LLMs.

4.3 Evaluation gaps

Evaluation practice reinforces collapse. The OPI dataset provides deterministic ground truth from curated resources, and standard metrics—exact match, precision, recall, F1—assess only categorical correctness. A model that outputs the correct label without justification is rewarded equally, or even more, than one producing a partially correct but biologically grounded explanation. This creates a metric–objective misalignment, encouraging terse predictions and discouraging explanatory richness. Though we provided general explanatory metrics in the previous section, they are still too generic and vulnerable. Developing evaluation protocols that account for interpretability is therefore essential.

4.4 Towards remedies and broader perspective

Addressing explanatory collapse requires interventions across training, evaluation, and model design. On the training side, rationale-augmented supervision—extending OPI-style datasets with explanatory text from UniProt [4] or Pfam/InterPro [14, 1]—can help balance accuracy with interpretability. In parallel, we are developing a large-scale annotation dataset that integrates sequence, structure, and functional rationales, aiming to expand the model's perception field from the data side and provide richer supervision for explanatory behavior. Multi-objective objectives that jointly optimize label prediction and rationale generation [26], or preference-based optimization methods such as reinforcement learning with expert feedback [15, 18], may further encourage faithful explanatory outputs. Equally important is moving beyond categorical accuracy as the sole benchmark: hybrid metrics combining correctness with rationale quality [24], semantic overlap with curated annotations,

motif- or domain-level coverage [30], and ontology-aware scoring [7] can provide external incentives for explanation. Incorporating uncertainty into evaluation also helps penalize overconfident but explanation-free predictions [16].

The collapse also reflects supervision and design limitations. Multi-modal grounding, where sequence and structure features are explicitly aligned with language [8], can support richer explanations by tying rationales directly to biological signals. Curriculum-style tuning that begins with rationale-rich supervision before label-only fine-tuning may preserve explanatory priors [29], ensuring that explanatory capacity is not overwritten during later training stages. Similarly, modular tuning approaches (e.g., adapters or LoRA) [6, 5] provide a mechanism for domain adaptation without fully overriding the generative capacity of the base model. In our own experiments, we explored the use of MLP adapters to connect sequence encoders with the language model, training these adapters on the same categorical instruction dataset employed during the first stage of tuning. This setup effectively aligned amino acid embeddings with the semantic space, but nonetheless failed to elicit explanatory behavior: the language model outputs were strongly biased toward replicating the exact label-oriented format of the fine-tuning corpus. This observation highlights that model design choices alone are insufficient—without rationale-rich supervision, even modular architectures will converge to explanation-suppressed behaviors.

More broadly, our findings illustrate a structural limitation of instruction tuning. In domains like protein science, where explanatory reasoning is indispensable, label-only fine-tuning can inadvertently strip away the very behaviors that make language models valuable. Remedying this challenge requires a holistic approach: datasets that combine categorical labels with biologically grounded rationales, objectives that optimize jointly for prediction and explanation, and benchmarks that explicitly evaluate interpretability alongside accuracy. Only by integrating these dimensions can protein LLMs provide not just reliable predictions but also explanatory reasoning that supports biological discovery.

References

[1] Matthias Blum, Hsin-Yu Chang, Sam Chuguransky, and et al. The interpro protein families and domains database: 20 years on. *Nucleic Acids Research*, 49(D1):D344–D354, 2021.

[2] Nadav Brandes, Dan Ofer, Yuval Peleg, Nadav Rappoport, and Michal Linial. Proteinbert: A universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, 2022.

[3] Y. Chen, K. Zhao, and J. Huang. When polished answers hide the truth: Rlhf and the suppression of model reasoning. *arXiv preprint arXiv:2502.04567*, 2025.

[4] The UniProt Consortium. Uniprot: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1):D480–D489, 2021.

[5] Junxian He, Banghua An, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. In *Advances in Neural Information Processing Systems*, 2022.

[6] Edward J Hu, Yelong Shen, Phillip Wallis, and et al. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2022.

[7] Jun Li, Zhang Yu, Xin Li, and et al. Ontology-aware evaluation of hierarchical classification in bioinformatics. *Bioinformatics*, 38(15):3764–3772, 2022.

[8] Zeming Lin, Halil Akin, Roshan Rao, and et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[9] Nelson F. Liu, Rik Koncel-Kedziorski, and Noah A. Smith. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 11:1574–1590, 2023.

[10] Zhiyuan Liu, Han Wang, Jie Zhou, and et al. Logicot: Logical chain-of-thought distillation. In *Advances in Neural Information Processing Systems*, 2023.

[11] A. Lobo, R. Kumar, and A. Gupta. Losing the reason: Instruction tuning reduces faithfulness of chain-of-thought. *arXiv preprint arXiv:2501.01234*, 2025.

[12] Shayne Longpre, Le Hou, Tu Vu, and et al. The flan collection: Designing data and methods for effective instruction tuning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

[13] Renqian Luo, Liang Sun, Yuxing Xia, and et al. Biogpt: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6), 2022.

[14] Jaina Mistry, Sam Chuguransky, Lorna Williams, and et al. Pfam: The protein families database in 2021. *Nucleic Acids Research*, 49(D1):D412–D419, 2021.

[15] Long Ouyang, Jeff Wu, Xu Jiang, and et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[16] Yaniv Ovadia, Emily Fertig, Jie Ren, and et al. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[17] Miguel Puerto, Weijia Zhao, and Peter Clark. Divergent cot: Training language models to explore multiple reasoning paths. In *International Conference on Learning Representations (ICLR)*, 2024.

[18] Rafael Rafailov, Archit Sharma, Eric Mitchell, and et al. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023.

[19] Alexander Rives, Joshua Meier, Tom Sercu, and et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15), 2021.

[20] R. Sahu, V. Gupta, and M. Bansal. Patterns and pitfalls of instruction tuning: Copying, hallucination, and reliability loss. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[21] Karan Singhal, Shekoofeh Azizi, Tong Tu, and et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[22] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, and et al. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`, 2023.

[23] Ross Taylor, Marcin Kardas, Guillem Cucurull, and et al. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.

[24] Miles Turpin, Julian Michael, and et al. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *arXiv preprint arXiv:2305.04388*, 2023.

[25] X. Wang, Y. Deng, and T. Zhang. Pinto: Penalizing inconsistent rationales for faithful reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

[26] Jason Wei, Xuezhi Wang, Dale Schuurmans, and et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[27] Hongwang Xiao, Wenjun Lin, Hui Wang, Zheng Liu, and Qiwei Ye. Opi: An open instruction dataset for adapting large language models to protein-related tasks. In *Neurips 2024 Workshop Foundation Models for Science: Progress, Opportunities, and Challenges*.

[28] L. Yang, Y. Wang, and S. Li. Re-critic: Self-critique with rationales reduces hallucination in multimodal models. *arXiv preprint arXiv:2503.06789*, 2025.

[29] Hao Zhang, Kuan-Hao Lee, Qi Li, and et al. Multistage instruction tuning for reasoning and explanation in language models. *arXiv preprint arXiv:2311.08900*, 2023.

[30] Qingyu Zhou, Han Wang, and Yue Zhang. Evaluating the faithfulness of explanation with ontology-based metrics. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

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 paper focus on the phenomenon of explanation collapse. The abstract and introduction have reviewed the main contribution and scope.

   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: We stated in the discussion section that the measurement is still generic and vulnerable.

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: [NA]

   Justification: The paper does not include theoretical results.

   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: The related dataset, experiment setup and models are clarified.

   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: The related dataset, models and inference results are available.

   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: We present the necessary details for the experiments.

   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: [No]

   Justification: We generally compares the general effect between different setups.

   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: We clarify the compute GPU type.

   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:

   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: [NA]

    Justification:

    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:

    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: [NA]

    Justification:

    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.

13. **New assets**

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

    Answer: [NA]

    Justification:

    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.

14. **Crowdsourcing and research with human subjects**

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

    Answer: [NA]

    Justification:

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.

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

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

    Answer: [NA]

    Justification:

    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: [NA]

    Justification:

    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.