Complementarity-Driven Distillation from Multiple Foundation Models for DNA Classification

Anonymous Author(s)
Affiliation
Address
email

Abstract

DNA sequence modeling has advanced with specialized foundation models such as HyenaDNA, yet these models capture only partial genomic cues. In this work, we investigate whether large language models (LLMs)—both subword-tokenized (LLaMA) and byte-level (EvaByte)—provide complementary perspectives when applied to DNA classification. Through experiments on the Human Enhancer Cohn benchmark, we find that DNA-pretrained models and LLMs succeed on largely disjoint subsets of data, revealing genuine cross-family complementarity. Building on this insight, we propose a confidence-guided distillation framework that aggregates supervision only from correct and confident teachers, producing soft labels that safely transfer diverse knowledge. Our method consistently improves both compact DNA-specific models and large byte-level LLMs, achieving gains of up to +2.34 accuracy points while remaining robust against overfitting even under near-perfect training accuracy. These findings highlight that DNA and language models encode orthogonal yet synergistic representations, and that principled distillation can unify them into a single model for robust genomic prediction.

1 Introduction

DNA sequence analysis lies at the heart of modern genomics. With the rapid advances in artificial intelligence, deep learning has emerged as a powerful paradigm for tackling DNA-related tasks such as enhancer detection, and regulatory element classification [Zhou and Troyanskaya, 2015, Kelley et al., 2016]. Motivated by the success of large-scale language modeling [Devlin et al., 2019, Touvron et al., 2023], a growing body of research has introduced DNA-specific foundation models such as DNABERT [Ji et al., 2021], DNABERT2 [Zhou et al., 2023], and HyenaDNA [Nguyen et al., 2023], which adapt Transformer to genomic sequences [Dalla-Torre et al., 2023, Avsec et al., 2021]. More recently, this line of work has been extended with models like Caduceus, which incorporates bi-directional equivariant long-range modeling [Cao et al., 2024], and DNABERT-S, which introduces species-aware embeddings for improved cross-species generalization [Zhang et al., 2024].

Despite these advances, current approaches remain limited. In particular, our understanding of what features these models capture from DNA sequences—and how these features contribute to classification performance—remains incomplete. More critically, DNA-pretrained models, though specialized, may fail to capture alternative structural or semantic cues that could be extracted if DNA sequences are treated as symbolic strings, analogous to natural language [Malusare et al., 2023].

In this work, we explore this perspective by hypothesizing that DNA sequences, viewed as ordered character strings, can be effectively modeled not only by DNA-specific pretraining but also by general-purpose large language models (LLMs). We investigate two complementary directions: (i) a subword-tokenized LLM such as LLaMA [Touvron et al., 2023], which leverages Byte-Pair Encoding (BPE), and (ii) a byte-level pre-trained model such as EvaByte [Zheng et al., 2025], which

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

processes inputs at the character level. The latter is particularly appealing, as it naturally aligns with the character-based nature of DNA sequences, enabling finer-grained representations without reliance on arbitrary subword segmentation [Malusare et al., 2023, Xue et al., 2022, Tay et al., 2022].

Through experiments, we demonstrate that DNA-pretrained models, standard LLMs, and byte-level LLMs capture different aspects of DNA sequences. Importantly, these models succeed and fail on complementary subsets of samples, both in test and training data. Building on this observation, we propose a simple yet safe knowledge distillation framework [Hinton et al., 2015, Furlanello et al., 2018, Zhang et al., 2019] that leverages model confidence to construct a soft-labeled distillation dataset. By averaging the confidences of models that correctly classify a sample, we generate training signals that preserve complementary knowledge. Distilling across DNA-specific and language-based models consistently improves performance, and notably, the resulting models remain robust even when trained to near-perfect accuracy on the distillation set.

Our contributions are threefold:

- We reveal that LLMs, including byte-level models, extract complementary representations from DNA sequences compared to DNA-pretrained models.
- We introduce a confidence-guided distillation dataset construction method that is both simple and resistant to overfitting.
- We empirically validate that this approach enhances DNA classification performance by safely integrating complementary knowledge across model families.

2 Method

2.1 Preliminary

We consider three types of models: a DNA-pretrained model (HyenaDNA), a subword-based large language model (LLaMA), and a byte-level language model (EvaByte). For LLaMA and EvaByte, we follow a generative prompt-based setup: each DNA sequence is provided in the form of `Use your background knowledge about DNA enhancers. Classify the following DNA sequence as an enhancer (O) or not (X):{DNA sequence}, Answer:`, and the model is trained to generate a binary answer (`O` for positive and `X` for negative). In contrast, HyenaDNA is trained following the original design in Nguyen et al. [2023], where a classifier head is attached to the backbone and optimized with supervised cross-entropy loss.

2.2 Motivation

During evaluation, we observe that the three models (HyenaDNA, LLaMA, and EvaByte) often succeed on different subsets of test samples, as shown in Figure 1. Although there exists a substantial overlap among correctly predicted samples, the disagreement across models is considerably larger than the discrepancy observed when comparing HyenaDNA models trained with different random seeds, as shown in Appendix B. This indicates that LLaMA and EvaByte capture aspects of DNA sequences that HyenaDNA fails to recognize, and vice versa.

Interestingly, such complementary behavior is not restricted to test data. Even among training samples, the models show differences in which instances they predict correctly or incorrectly. This finding suggests that each model focuses on distinct features of the input sequences during training, leading to complementary strengths. Building on this observation, we propose to leverage training samples themselves for distillation, enabling integration of the diverse knowledge captured by different models.

2.3 Distillation Framework

Based on the above observation, we construct a confidence-guided distillation dataset. Specifically, from the training set, we extract all samples that are correctly predicted by at least one of the three models. For each such sample, we record the confidence scores of the models that made correct predictions and compute the average of these scores to form a soft label. This process yields a soft-labeled dataset that captures complementary signals across models.

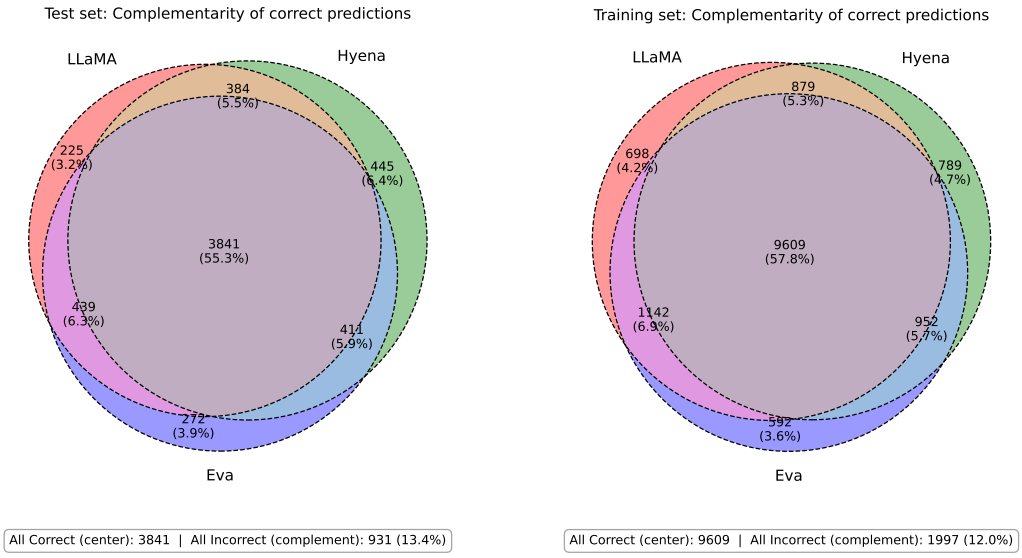

Figure 1: Complementarity of correct predictions among LLaMA, Hyena, and Eva on the test (left) and training (right) sets. Numbers indicate counts with percentages relative to the total; the complement region (all incorrect) is shown below each plot. Note that pairwise disagreements across model families substantially exceed the seed variance of a single family, evidencing genuinely complementary competence.

Finally, we use this distillation dataset to further train the models (HyenaDNA, LLaMA, and EvaByte). This simple yet effective framework allows each model to benefit from the unique perspectives of the others. The overall procedure is summarized in Algorithm 1.

Algorithm 1 Confidence-Guided Distillation

Input: Training set $\mathcal{D}_{\text{train}}$, teacher models $\mathcal{M}$
Initialize distillation set $\mathcal{D}_{\text{distill}} \leftarrow \varnothing$
for each sample $(x, y) \in \mathcal{D}_{\text{train}}$ **do**
    Collect predictions and confidences from all teachers
    Keep only teachers that predict y correctly
    if at least one correct teacher exists **then**
        Soft label $\tilde{y}(x) \leftarrow$ average confidence of correct teachers
        Add $(x, \tilde{y}(x))$ to $\mathcal{D}_{\text{distill}}$
    end if
end for
Return $\mathcal{D}_{\text{distill}}$

3 Experiments

3.1 Experimental Setup

We conduct all experiments on the Human Enhancer Cohn dataset from GenomicBenchmarks [Gresova et al., 2022], a widely used benchmark for enhancer identification. The task is formulated as a binary classification problem, distinguishing enhancer from non-enhancer sequences, with an equal number of samples per class to ensure balance. The dataset contains 10,421 positive and 10,422 negative samples in the training set, and 3,474 samples per class in the test set.

We evaluate three model families: HyenaDNA (1.6M parameters), a DNA-pretrained model with a classifier head; EvaByte (6.5B parameters), a byte-level large language model; and LLaMA3-8B (8B parameters), a subword-tokenized large language model. All models are first trained and evaluated on the Human Enhancer Cohn dataset using the same train/test split.

After this supervised training stage, we perform confidence-guided distillation. Specifically, distillation is applied to HyenaDNA, EvaByte, and LLaMA separately, starting from their baseline models trained on the Human Enhancer Cohn dataset, while keeping all other settings unchanged.

3.2 Baselines

As baselines, we use the three model families introduced above—HyenaDNA, EvaByte, and LLaMA3-8B—each trained individually on the dataset. In addition, we consider simple ensemble heuristics: (i) majority voting across the three models and (ii) a confidence-based strategy that selects the prediction from the most confident model. Finally, we compare these baselines against our confidence-guided distillation applied to each model family individually, as summarized in Table 1.

3.3 Main Results

Table 1 summarizes the effect of distillation compared to ensemble heuristics. Both HyenaDNA and EvaByte improve after distillation; LLaMA3-8B shows the largest gains. Majority voting and confidence-based ensembling achieve slightly higher accuracy in aggregate, but they require multiple models at inference time. By contrast, our confidence-guided distillation compresses complementary knowledge into a single student, offering comparable accuracy with much greater efficiency and deployability.

Table 1: Effect of confidence-guided distillation on the Human Enhancer Cohn dataset. Distilled rows indicate models fine-tuned from the corresponding baseline. Δ denotes the accuracy gain over baseline.

Model	Variant	Pretraining	Acc. (%)	Δ
HyenaDNA	Baseline	DNA	73.06	–
	Distilled	DNA	73.33	+0.27
EvaByte	Baseline	Natural Language	71.60	–
	Distilled	Natural Language	**73.50**	+1.90
LLaMA3-8B	Baseline	Natural Language	70.06	–
	Distilled	Natural Language	72.40	**+2.34**
Majority Voting Ensemble			73.07	–
Confidence-based Ensemble			**73.70**	–

3.4 No Collapse under High Memorization

A key finding is that even as both EvaByte and LLaMA achieve very high training accuracy (98.36% and 95.6%, respectively), their test accuracy does not collapse and stays stable around 72–74%. Even though the models memorize the training data almost perfectly, their test accuracy remains stable without degradation. This indicates that our distillation targets do not induce overfitting collapse, but instead provide robust generalization under high memorization. Figure 2 illustrates the training trajectory.

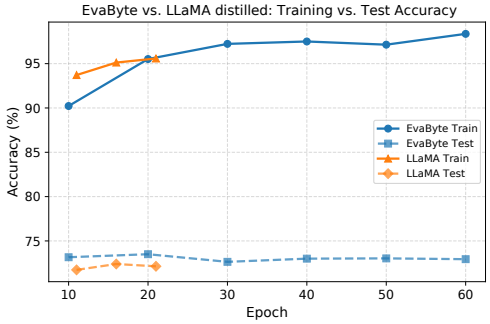

Figure 2: EvaByte and LLaMA under confidence-guided distillation.

4 Conclusion and Discussion

We investigated the complementarity between DNA-pretrained models and language-pretrained LLMs for DNA classification, showing that they capture orthogonal genomic signals. Building on this, we proposed a confidence-guided distillation strategy that integrates only reliable teacher signals, yielding consistent improvements—modest for HyenaDNA and substantial for EvaByte and LLaMA—while remaining robust against overfitting.

A key limitation of our approach is that distillation requires an additional training phase, which adds computational overhead and may constrain scalability. Looking forward, it would be valuable to move beyond post-hoc distillation and design unified architectures that can directly fuse heterogeneous sources of complementary knowledge. More broadly, our findings suggest that language-based LLMs can provide non-redundant genomic insights, pointing toward new opportunities for leveraging foundation models to advance trustworthy biomedical AI.

References

Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwińska, Kristin R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18(10):1196–1203, 2021.

Shuxiao Cao, Zhiwei Hong, Stephen A. Baccus, Christopher Ré, et al. Caduceus: Bi-directional equivariant long-range dna sequence modeling. *arXiv preprint arXiv:2405.07990*, 2024.

Hugo Dalla-Torre, Tristan Bepler, and et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *bioRxiv*, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 4171–4186, 2019.

Tommaso Furlanello, Zachary C Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born-again neural networks. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

Katarina Gresova, Vlastimil Martinek, David Cechak, Petr Simecek, Panagiotis Alexiou, et al. Genomic benchmarks: A collection of datasets for genomic sequence classification. *bioRxiv*, 2022. URL `https://www.biorxiv.org/content/10.1101/2022.06.08.495248`.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37 (15):2112–2120, 2021.

David R Kelley, Jasper Snoek, and John L Rinn. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*, 26(7):990–999, 2016.

Aditya Malusare, Harish Kothandaraman, et al. Understanding the natural language of dna using encoder-decoder foundation models with byte-level precision. *arXiv preprint arXiv:2311.02333*, 2023.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Callum Birch-Sykes, Michael Wornow, Aman Patel, Clayton Rabideau, Stefano Massaroli, Yoshua Bengio, Stefano Ermon, Stephen A Baccus, and Christopher Ré. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *arXiv preprint arXiv:2306.15794*, 2023.

Yi Tay, Mostafa Dehghani, Vinh Q Tran, Xavier Garcia, Dara Bahri, Zhen Xiao, Naveen Arivazhagan, and Donald Metzler. Charformer: Fast character transformers via gradient-based subword tokenization. In *International Conference on Learning Representations (ICLR)*, 2022.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. Byt5: Towards a token-free future with pre-trained byte-to-byte models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Your classifier is secretly an energy based model and you should treat it like one. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Hao Zhang, Zhihan Zhou, Yanrong Ji, and Han Liu. Dnabert-s: Pioneering species differentiation with species-aware dna embeddings. *arXiv preprint arXiv:2406.01914*, 2024.

Lin Zheng, Xueliang Zhao, Guangtao Wang, Chen Wu, David Dong, Angela Wang, Mingran Wang, Yun Du, Haige Bo, Amol Sharma, Bo Li, Kejie Zhang, Changran Hu, Urmish Thakker, and Lingpeng Kong. Evabyte: Efficient byte-level language models at scale. `https://hkunlp.github.io/blog/2025/evabyte`, 2025.

Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods*, 12(10):931–934, 2015.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006*, 2023.

A Experimental Details

We follow the standard GenomicBenchmarks split: training and test sets are provided, and we further reserve 20% of the training set for validation. All models are trained with AdamW optimizer, a linear warmup of 100 steps, and early stopping with patience of 3 based on validation performance. Final results are reported on the held-out test split using the checkpoint with the best validation score.

A.1 HyenaDNA: Training and Evaluation Details

For HyenaDNA, DNA sequences are tokenized at the character level ({A,C,G,T,N}) with maximum length 500. The model is initialized from a pretrained HyenaDNA checkpoint. Training uses AdamW (learning rate 6×10^{-4}, weight decay 0.1, batch size 256) and cross-entropy loss. Unlike the original HyenaDNA setup that trains for 100 epochs, we apply early stopping on validation loss and select the best checkpoint.

A.2 LLaMA3-8B: Training and Evaluation Details

For LLaMA3-8B, we adopt a generative prompt-based setup, where each DNA sequence is presented as: `Use your background knowledge about DNA enhancers. Classify the following DNA sequence as an enhancer (O) or not (X): {DNA sequence}, Answer:` and the model is trained to autoregressively predict the label (O/X). We apply LoRA (rank $r = 8$, $\alpha = 16$, dropout 0.05) to the attention projection modules (q_proj, k_proj, v_proj, up_proj, $down\_proj$). Optimization uses learning rate 1×10^{-5} and batch size 16. We train for up to 100 epochs with early stopping on validation accuracy.

A.3 EvaByte: Training and Evaluation Details

EvaByte is trained with the same prompt-based causal language modeling setup as LLaMA3-8B, where each DNA sequence is followed by a binary label (O/X) to be predicted autoregressively. The main differences are in the hyperparameters: we use batch size 4 (instead of 16), and learning rate 1×10^{-5} without weight decay. LoRA (rank $r = 8$, $\alpha = 16$, dropout 0.05) is applied to the attention projection modules as in the LLaMA3-8B experiments.

B Complementarity Beyond Seed Variance

To verify that the observed complementarity is not simply due to randomness in training or seed variance, we compared overlaps between HyenaDNA models trained with different seeds and between HyenaDNA and other model families. Table 2 shows that the disagreement between HyenaDNA runs (seed0 vs. seed42) is relatively small, while the disagreement between HyenaDNA and EvaByte is substantially larger. This confirms that the complementarity arises from model family differences rather than stochastic variation.

Compared to the $\sim$10% disagreement across HyenaDNA seeds, cross-family comparisons with EvaByte exhibit nearly double the disagreement rate (around 20%). This demonstrates that large

Table 2: Comparison of prediction overlaps on the Human Enhancer Cohn dataset. "Both Correct" denotes samples where both models predict correctly, "Disagree" denotes samples where one model is correct and the other is not, and "Both Wrong" denotes samples where both fail. Percentages are relative to the full test set.

Pair of Models	**Both Correct**	**Disagree**	**Both Wrong**
Hyena (seed0, 72.25%) vs. Hyena (seed42, 73.13%)	67.7%	9.9%	22.3%
Hyena (seed42, 73.13%) vs. EvaByte (71.43%)	61.2%	22.2%	16.6%
Hyena (seed0, 72.25%) vs. EvaByte (71.43%)	62.4%	19.9%	18.7%

language models such as EvaByte capture distinct genomic cues that DNA-pretrained models like HyenaDNA do not, indicating genuine complementarity rather than random variation. Therefore, our distillation framework leverages orthogonal knowledge across model families, beyond what can be obtained by simply retraining the same architecture with different seeds.

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: Abstract and introduction includes paper's contributions.

   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: Section 4 Conclusion and Discussion includes limitations of the 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:[NA]

Justification: This paper does not include theorem.

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: Section 3 Experiments include those information.

   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: [No]

Justification: All experiments and method are explained in detail in the manuscript, so that they can be reproduced easily. Furthermore, the code will be available in camera-ready version.

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: Section 3 Experiments include those information.

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: The experiments are conducted on a single seed but the results are justified with same results across various model architecture.

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: Section 3 Experiments include those information.

   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: This research follows the guidelines.

   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: [No]

    Justification: There is no societal impact of the work performed.

    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: This research does not include risk for misuse.

    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: They are properly cited in references.

    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: This paper does not release new assets.

    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: This paper does not involve crowdsourcing nor research with human subjects.

    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: This paper does not involve crowdsourcing nor research with human subjects.

    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: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

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.