# Confidence–Coverage Gating for Early Exit

**Aaroosh Rustagi***
**Lynbrook High School**
aarooshr@gmail.com

**Hsien Xin Peng***
**Ridge High School**
maxpeng123678@gmail.com

**Khushal Murthy**
**Mountain House High School**
khushal.murthy09@gmail.com

**Attrey Koul**
**Foothill High School**
attreykoul1@gmail.com

**Ryan Lagasse**
**Algoverse AI Research**
ryan@algoverseresearch.com

**Kevin Zhu**
**Algoverse AI Research**
kevin@algoverse.us

## Abstract

Smaller Large Reasoning Models (LRMs) have shown remarkable capabilities for their model size. However, due to Chain-of-Thought (CoT) reasoning, these models often produce redundant and verbose reasoning chains when short reasoning suffices, leading to excessive computation and tokens generated. We propose a training-free early exit approach that detects newline-scoped, low confidence connector words and self-truncates at the boundary of the previous step when that step shows sufficient semantic similarity to the original prompt. Our three-pronged, training-free approach can be easily incorporated with open-source LRMs such as DeepSeek-Distill-Qwen-7B, DeepSeek-Distill-Llama-8B and QwQ-32B. Experiments across GSM8K, MATH500, and AMC have resulted in a minimal reduction in average accuracy and a significant decrease in average token count. More broadly, our method highlights the potential of using low-confidence tokens to identify potential self-truncation points for early exiting.

## 1 Introduction

Smaller LRMs have demonstrated exceptional reasoning skills and proficiency in reasoning tasks [4]. The accuracy of many LRMs is enhanced by their ability to enumerate their reasoning before providing a well-reasoned answer [4]. However, these methods often generate large amounts of excess tokens, resulting in significant computational overhead [14]. Connector words such as "Wait" can elicit further reasoning, but can also inadvertently exacerbate excessive token generation by encouraging further unnecessary reasoning or answer verification [14]. Therefore, these connector words are natural indicators of logical shifts in model reasoning.

Using these natural indicators, we propose a training-free method **C**onfidence-**C**overage **G**ating for **E**arly **E**xit (**CCGEE**) that allows LRMs to self-truncate excessively verbose reasoning chains. We segment the model output into a sequence of reasoning steps split by newline characters. We run experiments to determine a bank of connector words, such as "Wait", which are defined in Appendix B, as possible self-truncation points if they appear at the start of these reasoning steps. We also observe that LRMs often repeat sub-phrases of the original question when they output an answer. Therefore, our method decides whether to self-truncate by looking at the log probability of the first token of the connector word and comparing keywords to determine the previous step's semantic similarity with the question. In our method, the model self-truncates when three conditions are satisfied: 1) **Connector Word After Newline**: A common connector word appears immediately

*Equal Contribution

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Efficient Reasoning.

after a newline character, which is also the start of the next reasoning step. 2) **Low Confidence Connector Word**: The first token of the connector word in question is low-confidence. 3) **Question Coverage Metric**: The preceding reasoning step reiterates key words from the original prompt. After all three conditions are met, the model is forced to output a final answer. To summarize, our main contributions are as follows:

1. We demonstrate that low-confidence connector tokens can be used as high-signal indicators of possible self-truncation points.
2. We propose a dynamic and lightweight heuristic (**CCGEE**) for self-truncation that keeps accuracy similar to CoT and reduces average token count significantly.

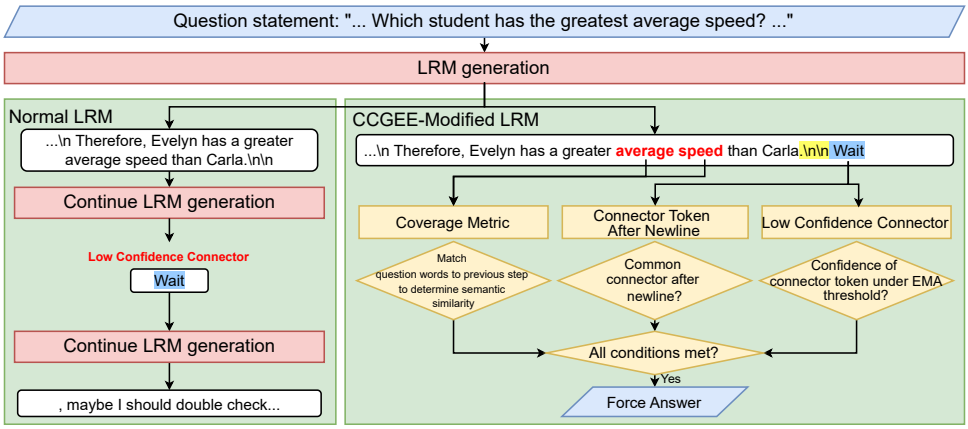


Figure 1: Our method allows LRMs to self-truncate, reducing average token count. At low-confidence connector tokens, vanilla LRMs continue generation, while our method catches when an answer has likely been reached using our three-pronged approach, forcing an answer and saving compute.

## 2 Related Works

**Existing LRMs** State-of-the-art (SOTA) LRMs such as OpenAI's o1 [11], and DeepSeek-R1 [4] have improved significantly in their ability to perform reasoning tasks [1][14]. Models distilled from Deepseek-R1, such as the Qwen-7B and Llama-8B variants surpass the SOTA models QwQ-32B [17] and OpenAI-o1-mini on certain metrics [4]. Similar to larger SOTA models, these smaller LRMs tend to reason extensively even after providing an answer.

**Self-Truncation and Early Exit** Many methods employ early exit strategies for more efficient transformer inference by exiting at intermediate layers [16][5][18][6][21]. Some more recent papers employ self-truncation methods to stop inference early and force the model to output a final answer[3]. Certaindex and Dynasor [7] are a model-agnostic metric that probe the model at regular intervals to get a preliminary answer, and then use recurrences of trial answers as a signal to exit. DEER [20] assesses self-truncation points but only on connector words, and uses the confidence of the generated answer as a signal to exit early. s1 test-time scaling [14] can end reasoning early or force a model to continue output to fit a token budget.

**Confidence-Based Methods** Many works use model confidence as a metric to determine whether or not an early exit can be executed. [15] [9]. ConCISE [15] reinforces the token confidence and exits early when the post-answer confidence is high enough. DEER [20] exits early when trial answer token confidence is above a certain threshold, and finally,

Existing methods use high confidence as a metric to determine when to terminate inference. Our method, however, scans for low confidence in connector words to determine whether an early exit is feasible, highlighting the value of low confidence tokens in determining when a model is reasoning excessively.

## 3 Methodology

**Reasoning Step Deconstruction** We generate responses for LRMs and store the log probabilities of the generated tokens. We denote the reasoning steps of the LRM's response as the text between subsequent newline characters due to the cleaner segmentation of model reasoning. The heuristic only looks at tokens that strictly come at the start of each reasoning step.

**Confidence Evaluation** Our method processes the first few tokens in each step and checks if the first word is a connector word. If it is, we classify this connector token as low confidence if it falls under the Exponential Moving Average (EMA) of past eligible connector words. This allows the low confidence threshold to change depending on each response's differing connector token confidences, allowing our method to generalize across different problems, datasets and models. We calculate EMA sequentially, where $EMA_i$ represents the EMA after the $i$-th eligible connector token, $C_i$ represents the confidence of the current connector token and $\alpha$ is a constant hyperparameter. We set $\alpha$ to 0.5 after initial, small-scale experiments, but with further finetuning, accuracy could be improved.

$$EMA_i = (C_i\alpha) + (EMA_{i-1} \cdot (1 - \alpha)) \tag{1}$$

If the current connector token's log probability is less than a hyperparameter ratio $\beta$ multiplied by the EMA of the previous connector words:

$$C_i < \beta \cdot EMA_{i-1} \tag{2}$$

Then, the token is classified as low confidence and is further evaluated for keyword similarity to the original prompt by using a coverage metric.

**Coverage Evaluation** The coverage metric is calculated by comparing two sets of words. One set is created from the original prompt by finding the first sentence that contains a common question word, such as "Calculate" (Appendix A). Once this common question word is found, we take the last five words of the same sentence, with the choice of five being chosen from initial experiments due to time constraints. We then filter the last five words using the NLTK English Stopwords Library [12] to remove filler words. Our second set is the words from the previous step. With these two sets, our coverage score is calculated by the overlap of the two sets. Specifically:

$$S(L_1, L_2) \;=\; \frac{|L_1 \cap L_2|}{|L_1|}.$$

**Early Exiting** If $S(L_1, L_2)$ is larger than a hyperparameter constant threshold $\tau$, then the model self-truncates and answers. We force self-truncation by appending `"Therefore, I think that is the correct answer.\n\n**Final Answer**\n\n"` to the model's current response. We use this as it is a common pattern that LRMs use when outputting the answer.

## 4 Experiments

### 4.1 Experimental Setup

**Models, Datasets, Baselines** Using Chain-of-Thought prompting, we ran all problems through QwQ-32B, DeepSeek-R1-Distill-LLaMA-8B, and DeepSeek-R1-Distill-Qwen-7B. We evaluated on multiple math reasoning benchmarks: GSM8K[2], MATH500[8], and AMC[13]. Math-Verify[10] was used for grading model responses.

**Implementation Details** For all experiments, greedy decoding is used. We set `max_new_tokens` to 4096 for GSM8K and MATH500, and 8192 for AMC. Hyperparameters used for the EMA ratio ($\beta$), and coverage ($\tau$) are detailed below. All experiments were done on a H200 SXM.

**Hyperparameter Determination** To determine the best hyperparameters for $\beta$ and $\tau$, we performed a grid search on a representative subset of 270 problems, with 90 problems each from GSM8K, MATH500, and AIME[19] in order to prevent overfitting to easier problems in GSM8K and MATH500. Using the results, we generated Pareto frontiers, which are shown in Figure 2. We present the Pareto point with $\beta = 0.95$ and $\tau = 0.55$ in Table 1.

Table 1: Average accuracy decreases by 1.42% while resulting in a 21.4% average token reduction.

| | | GSM8K | | MATH500 | | AMC | |
|---|---|---|---|---|---|---|---|
| Model | Method | Acc. | Avg Tok. | Acc. | Avg Tok. | Acc. | Avg Tok. |
| DeepSeek-Distill-Qwen-7B | CoT | 88.93% | 1440.3 | 82.40% | 2481.6 | 71.08% | 4790.6 |
| | CCGEE | 88.40% | 1119.9 | 83.40% | 2090.2 | 72.29% | 3868.6 |
| DeepSeek-Distill-Llama-8B | CoT | 81.80% | 960.3 | 74.60% | 2879.4 | 56.63% | 5849.8 |
| | CCGEE | 81.12% | 777.8 | 75.80% | 2329.6 | 51.81% | 4622.2 |
| QwQ-32B | CoT | 92.34% | 2689.5 | 82.40% | 3275.4 | 72.29% | 6710.8 |
| | CCGEE | 89.16% | 1955.3 | 81.40% | 2648.0 | 66.27% | 4700.0 |

### 4.2 Results

We conducted experiments on the datasets listed above, and our findings are listed in Table 1. With only GSM8K and MATH500, average accuracy decreases by 0.53%, but results in a 20.4% average token reduction. Even with the challenging AMC dataset added, average accuracy only decreases by 1.42% while resulting in an even greater 21.4% average token reduction.

### 4.3 Ablation Study

In order to demonstrate the importance of our three conditions, we conduct three ablation studies: (1) Using low-confidence connector words without coverage thresholding (2) Using coverage thresholding without connector words (3) Using connector words and coverage thresholding but without low confidence. Figure 2 shows that only CCGEE was able to achieve the same accuracy as the baseline with a token reduction.

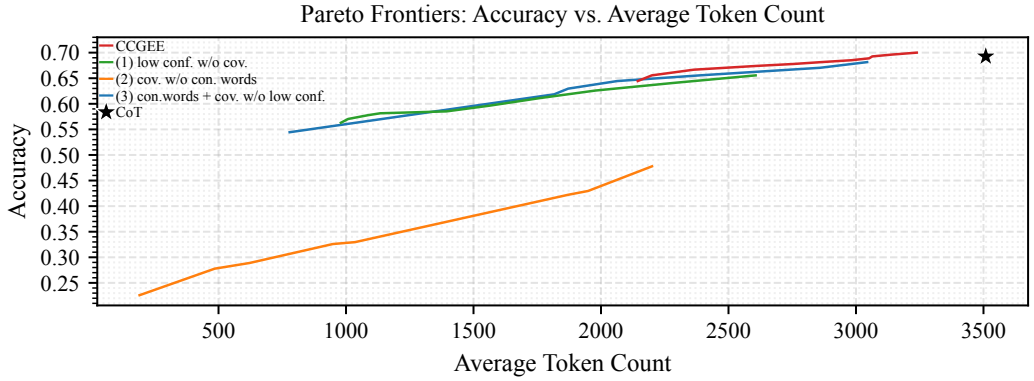


Figure 2: Pareto frontiers calculated with grid search using representative subset. CCGEE offers the highest accuracy (accuracy > 0.65) for a given average token per problem. CCGEE was able to achieve a slightly higher accuracy than CoT with a token reduction.

## 5 Limitations

Further evaluations on non-math-related reasoning benchmarks are necessary to further gauge the generalizability of CCGEE. More detailed analysis on finding the best hyperparameters is necessary, as the value of $\alpha$ and the usage of the last five words for the coverage metric have not been rigorously proven to be optimal values.

## 6 Conclusion

We propose a new method that uses low-confidence connectors, in addition to a question coverage metric, to decide when a model has likely reached the correct answer. We demonstrate that our

method is viable across different state-of-the-art LRMs and 3 math datasets with different difficulties. On top of this, we provide a Pareto frontier to demonstrate the tradeoff between token count and model accuracy, yielding as much as a 21.4% decrease in average token count with only a small 1.42% average decrease in accuracy when the much harder AMC dataset was included.

## 7 References

## References


[1] Sébastien Bubeck et al. *Sparks of Artificial General Intelligence: Early experiments with GPT-4*. 2023. arXiv: 2303.12712 [cs.CL]. URL: https://arxiv.org/abs/2303.12712.

[2] Karl Cobbe et al. *Training Verifiers to Solve Math Word Problems*. 2021. arXiv: 2110.14168 [cs.LG]. URL: https://arxiv.org/abs/2110.14168.

[3] Muzhi Dai, Chenxu Yang, and Qingyi Si. *S-GRPO: Early Exit via Reinforcement Learning in Reasoning Models*. 2025. arXiv: 2505.07686 [cs.AI]. URL: https://arxiv.org/abs/2505.07686.

[4] DeepSeek-AI et al. *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. 2025. arXiv: 2501.12948 [cs.CL]. URL: https://arxiv.org/abs/2501.12948.

[5] Alexander Yom Din et al. *Jump to Conclusions: Short-Cutting Transformers With Linear Transformations*. 2024. arXiv: 2303.09435 [cs.CL]. URL: https://arxiv.org/abs/2303.09435.

[6] Siqi Fan et al. *Not All Layers of LLMs Are Necessary During Inference*. 2024. arXiv: 2403.02181 [cs.CL]. URL: https://arxiv.org/abs/2403.02181.

[7] Yichao Fu et al. *Efficiently Scaling LLM Reasoning with Certaindex*. 2025. arXiv: 2412.20993 [cs.LG]. URL: https://arxiv.org/abs/2412.20993.

[8] Dan Hendrycks et al. *Measuring Mathematical Problem Solving With the MATH Dataset*. 2021. arXiv: 2103.03874 [cs.LG]. URL: https://arxiv.org/abs/2103.03874.

[9] Jiameng Huang et al. *Efficient Reasoning for Large Reasoning Language Models via Certainty-Guided Reflection Suppression*. 2025. arXiv: 2508.05337 [cs.CL]. URL: https://arxiv.org/abs/2508.05337.

[10] Hynek Kydlíček. *Math-Verify: Math Verification Library*. Version 0.6.1. 2025. URL: https://github.com/huggingface/math-verify.

[11] "Learning to reason with llms." In: (). URL: https://openai.com/index/%20learning-to-reason-with-llms,%202024..

[12] Edward Loper and Steven Bird. *NLTK: The Natural Language Toolkit*. 2002. arXiv: cs/0205028 [cs.CL]. URL: https://arxiv.org/abs/cs/0205028.

[13] AI-MO. *AIMO Validation AMC Dataset*. https://huggingface.co/datasets/AI-MO/aimo-validation-amc. Accessed March 29, 2025. 2024.

[14] Niklas Muennighoff et al. *s1: Simple test-time scaling*. 2025. arXiv: 2501.19393 [cs.CL]. URL: https://arxiv.org/abs/2501.19393.

[15] Ziqing Qiao et al. *ConCISE: Confidence-guided Compression in Step-by-step Efficient Reasoning*. 2025. arXiv: 2505.04881 [cs.LG]. URL: https://arxiv.org/abs/2505.04881.

[16] Tal Schuster et al. *Confident Adaptive Language Modeling*. 2022. arXiv: 2207.07061 [cs.CL]. URL: https://arxiv.org/abs/2207.07061.

[17] Qwen Team. *QwQ-32B: Embracing the Power of Reinforcement Learning*. Mar. 2025. URL: https://qwenlm.github.io/blog/qwq-32b/.

[18] Florian Valade. *Accelerating Large Language Model Inference with Self-Supervised Early Exits*. 2024. arXiv: 2407.21082 [cs.CL]. URL: https://arxiv.org/abs/2407.21082.

[19] Hemish Veeraboina. *AIME Problem Set 1983-2024*. Kaggle, 2023. URL: https://www.kaggle.com/datasets/hemishveeraboina/aime-problem-set-1983-2024.

[20] Chenxu Yang et al. *Dynamic Early Exit in Reasoning Models*. 2025. arXiv: 2504.15895 [cs.CL]. URL: https://arxiv.org/abs/2504.15895.

[21] Bowen Zheng et al. *A Hybrid Early-Exit Algorithm for Large Language Models Based on Space Alignment Decoding (SPADE)*. 2025. arXiv: 2507.17618 [cs.CL]. URL: https://arxiv.org/abs/2507.17618.

## A Question–Trigger List

The common question words are: ['find', 'what', 'how', 'which', 'calculate', 'determine', 'solve', 'compute', 'evaluate', 'show', 'prove', 'verify', 'simplify', 'express', 'convert', 'transform', 'derive', 'obtain', 'get', 'approximate', 'estimate', 'compare', 'identify', 'name', 'list', 'state', 'describe', 'define', 'classify', 'categorize', 'sort', 'arrange', 'order', 'rank', 'maximize', 'minimize', 'optimize', 'draw', 'sketch', 'plot', 'graph', 'construct', 'build', 'create', 'check', 'test', 'confirm', 'validate', 'demonstrate', 'illustrate', 'explain']

## B Common Connector Words

The common connector words are: ['wait', 'hmm', 'just', 'another', 'alternatively', 'similarly']

## 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: Yes, the abstract and introduction accurately reflect the paper's contributions as shown by Table 1.

   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: See section 5.

   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 propose any 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: Yes, see Sections 3 and 4

   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: We do not release any code.

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: Yes, see Section 4

   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: No, the paper does not report any error bars.

   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: [No]

   Justification: While we do not have memory or time of execution statistics readily available, we do mention that all our experiments were run on a H200 SXM GPU.

   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 in the paper conforms to the NeurIPS Code of Ethics.

   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: There is no societal impact of the work performed in this paper.

    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 pose any such risks.

    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: The paper does not use any existing assets.

    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: The 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: The 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: The 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.