# A Fast and Effective Solution to the Problem of Look-ahead Bias in LLMs

**Humzah Merchant**
University of Chicago
Chicago, IL 60637

**Bradford Levy**[*]
University of Chicago
Chicago, IL 60637

## Abstract

Applying LLMs to predictive tasks in finance is challenging due to look-ahead bias resulting from their training on long time-series data. This precludes the backtests typically employed in finance since retraining frontier models from scratch with a specific knowledge cutoff is prohibitive. In this paper, we introduce a fast, effective, and low-cost alternative. Our method guides generation at inference time by adjusting the logits of a large base model using a pair of smaller, specialized models—one fine-tuned on information to be forgotten and another on information to be retained. We demonstrate that our method effectively removes both verbatim and semantic knowledge, corrects biases, and outperforms prior methods.

## 1 Introduction

LLMs have demonstrated remarkable capabilities, achieving state-of-the-art performance on a diverse array of natural language tasks [2, 25, 6] and specialized financial applications [5, 21, 26]. However, applying LLMs to the predictive tasks common in financial applications is challenging due to the fact that they are trained on long time-series [20] and known to memorize content [4, 3]. For example, if an LLM has memorized historical earnings figures and stock prices, then estimates of predictive ability based on historical data, i.e., a backtest, are likely overly optimistic [see, e.g., 11, 19, 21].

The simplest solution to such concerns is training a model from scratch with a given historical knowledge cutoff [e.g., 8]. While simple, the very scale that enables powerful generalization of LLMs creates significant challenges to doing this. First, it is likely cost prohibitive to train a frontier quality LLM purely to enable backtesting [15, 12, 7]. Second, even if budget allowed, modern foundation models are trained on trillions of tokens, much of which collected only recently. Thus, data availability alone may preclude training a "historical" frontier model.

In this work, we propose a theoretically motivated and conceptually simple alternative to training from scratch: **Inference-Time Unlearning**. Rather than modifying the weights of a frontier model, our method guides generation at inference time by using a pair of smaller, specialized models. One small model is fine-tuned on the data to be forgotten, e.g., knowledge after 2010, while another is tuned on a proxy for the data to be retained, e.g., knowledge prior to 2010. By modifying the logits of the frontier model with the difference of the specialized models, through a variety of experiments we document that our method effectively unlearns unwanted content while preserving model utility.

Our paper makes three primary contributions to the literature on machine unlearning:

1. **Efficacy**: We demonstrate that our method (i) effectively removes both verbatim and semantic knowledge from a model, and (ii) can correct unwanted biases such as primacy.
2. **Efficiency**: By restricting fine-tuning to small auxiliary models (with orders of magnitude fewer parameters than the base LLM), our approach drastically reduces the computational

[*]Direct correspondence to `bll@uchicago.edu`. Code and data on GitHub.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Generative AI in Finance.

cost of unlearning. For example, we find that even simple tri-gram based LMs are effective. This makes on-demand unlearning practical and scalable.

3. **Utility Preservation**: Our method maintains the model's performance on general knowledge and standard evaluation benchmarks. Because the base model's weights remain unchanged, the impact on its core capabilities is minimal, outperforming prior methods in preserving utility as the number of unlearning requests grows.

Collectively, we demonstrate that our approach provides a practical, low-cost, and effective solution to the critical problem of selectively forgetting information in LLMs, paving the way for more reliable evaluation of their abilities within financial applications.

# 2 Method

We begin by defining the problem, introducing our method, and finally connecting it to existing work. Let $V$ denote a finite vocabulary of tokens. A token sequence of length $T$ is denoted as $x = (x_1, x_2, ..., x_T)$ where each token $x_t \in V$. The prefix of a token sequence up to token $t-1$ is denoted $x_{<t} = (x_1, ..., x_{t-1})$. There are two data generating distributions $D_A$ and $D_B$ where the support of $D_B$ is contained within $D_A$. Finally, $P(x_t|x_{<t})$ and $Q(x_t|x_{<t})$ denote the conditional token distributions under $D_A$ and $D_B$, respectively.

We consider the situation where we wish to sample from $Q$ but do not have access to it. Instead, only $P$ is accessible. For example, $P$ could be a large frontier model for which it is cost prohibitive to retrain a new model from scratch on $D_B$. Within the finance domain, $Q$ could be a model as capable as $P$ but trained up to a fixed knowledge cutoff so as to avoid look-ahead bias. Generally, our goal is to approximate sampling from $Q$ using only $P$ and samples drawn from $D_A$ and $D_B$.

## 2.1 Divergence decoding

Consider two small models $p(x_t|x_{<t})$ and $q(x_t|x_{<t})$ trained on samples from $D_A$ and $D_B$, respectively. Denote the logits of a given model $M$ as $l_M(x_{<t}) \in \mathbb{R}^{|V|}$. Divergence Decoding (DD) approximates sampling from $Q$ by adjusting the logits of $P$ according to the divergence between $q$ and $p$. Empirically, we consider two adjustments. The first is a linear combination of the logits,

$$\hat{l}_Q^{LC}(x_{<t}) = l_P(x_{<t}) + \alpha \cdot [l_q(x_{<t}) - l_p(x_{<t})], \tag{1}$$

while the second adjustment is rank based,

$$\hat{l}_Q^{R}(x_{<t}) = l_P(x_{<t}) - \mathbb{1}_{rank(l_p(x_{<t}) - l_q(x_{<t})) \leq k} \cdot \infty. \tag{2}$$

In the case of the linear adjustment, if the difference between $Q$ and $P$ is indeed linear in logit space, then there exists some value of $\alpha$, $p$, and $q$ which enables $Q$ to be perfectly recovered. If the difference is not linear however, then this is not true. For this reason, we also explore the rank based approach, which prevents generating the top-$k$ most divergent tokens between $p$ and $q$.

Samples can then be drawn via typical methods [e.g., 9, 16] from the approximation,

$$\widehat{Q}(x_t|x_{<t}) = \text{softmax}(\hat{l}_Q(x_{<t})). \tag{3}$$

While the adjustments in Eq. 1 and 2 require additional forward passes for $p$ and $q$, we show in Section 3 that strong performance can be achieved even when $p$ and $q$ are trigram models—which add negligible computational overhead (see Figure 4a).

## 2.2 Theoretical motivation

While simple to implement and fast at inference time, our method is theoretically motivated by the Product of Experts [14] and Importance Sampling [13] literature. In Appendix A.1, we show that the approximation $\widehat{Q}$ can be formulated as a Product of Experts model,

$$\widehat{Q}(x_t|x_{<t}) \propto \underbrace{P(x_t|x_{<t})}_{\text{Base Expert}} \cdot \underbrace{\left[\frac{q(x_t|x_{<t})}{p(x_t|x_{<t})}\right]^{\alpha}}_{\text{Domain Expert}} \tag{4}$$

where $\widehat{Q}$ is the product of a "Base Expert" $P$ responsible for providing foundational knowledge and a "Domain Expert" comprised of the ratio of $q$ to $p$. Intuitively, the role of the domain expert can be summarized by three cases:

1. $q \approx p$: Tokens are similarly likely under both $D_A$ and $D_B$ and the domain expert ratio is close to 1 effectively leaving the probabilities from the base model $P$ unchanged
2. $q \gg p$: Tokens are much ***more*** likely under $D_B$ than $D_A$, and the domain expert "upvotes" such tokens by ***increasing*** the probability assigned to them
3. $q \ll p$: Tokens are much ***less*** likely under $D_B$ than $D_A$, and the domain expert "downvotes" such tokens by ***decreasing*** the probability assigned to them

Finally, DD can also be linked to importance sampling in Monte Carlo analysis whereby the expectation of some function $f(x)$ under a target distribution $D_{target}$ is estimated using samples drawn from a proposal $D_{proposal}$. Formally,

$$\mathbb{E}_{x\sim D_{target}}[f(x)] = \mathbb{E}_{x\sim D_{proposal}}\left[f(x)\frac{D_{target}(x)}{D_{proposal}(x)}\right], \tag{5}$$

where the importance weight $w(x) = \frac{D_{target}(x)}{D_{proposal}(x)}$ adjusts the expectation taken over $D_{proposal}$ for differences between the proposal and target distributions. Analogously, divergence decoding uses the ratio of $q$ to $p$ to adjust for differences between the inaccessible model $Q$ and accessible one $P$.

# 3 Experiments

Next, we explore the efficacy of our method across a variety of empirical settings. First, we evaluate its performance on the well-established benchmark "MUSE: Machine Unlearning Six-Way Evaluation for Language Models" [23]. We then explore performance on finance specific tasks: (i) unlearning knowledge of mergers and acquisitions, and (ii) debiasing forecasts of future performance.

## 3.1 MUSE unlearning benchmark

We apply our method to the news dataset from the MUSE benchmark [22, 23]. For our small specialized models, $p$ and $q$, we use the *princeton-nlp/Sheared-LLaMA-1.3B* [27] GPT model—because it shares a tokenizer with the benchmark models used by MUSE [23]—as well as trigram LMs based on *Stupid Backoff* [1]. Across both types of models, we find that our method matches or pushes the frontier of unlearning while preserving model utility (see Figure 1). This is particularly notable given the reduced computational costs of our method relative to prior work (see Section A.2) e.g., trigram based models are effectively costless (see Figure 4a). In additional tests we evaluate strong performance on the **privacy**, **scalability**, and **sustainability** benchmarks (see Section B.1).

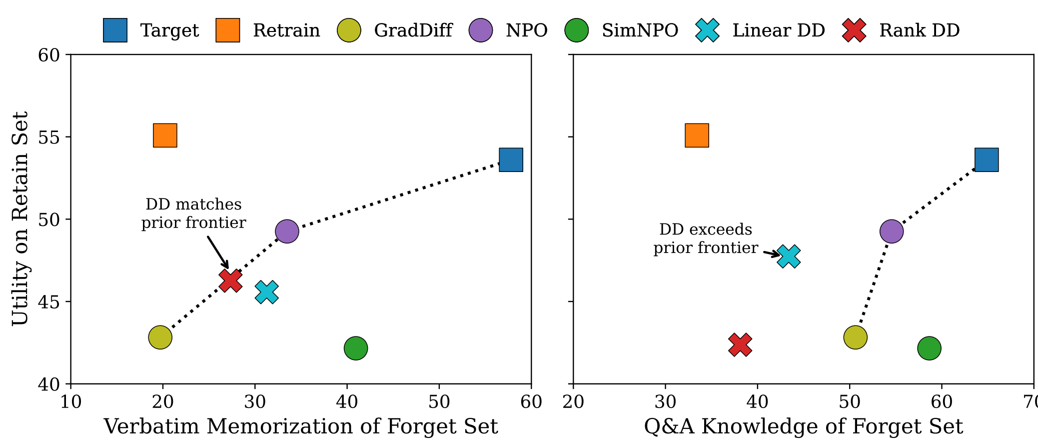


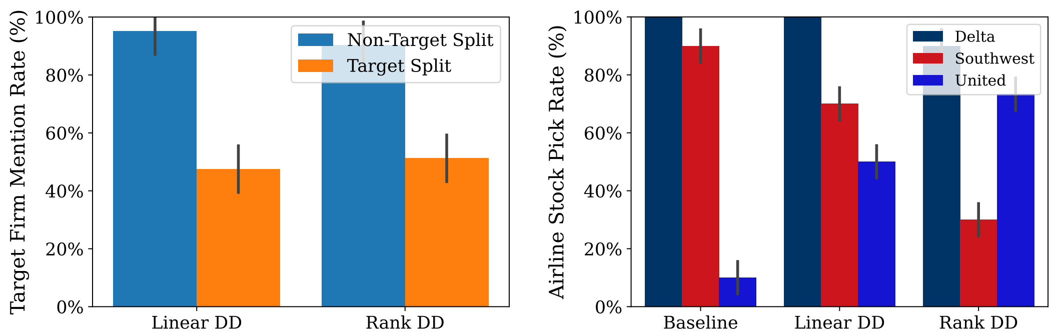

Figure 1: MUSE Results: Target is the model to which unlearning is applied. Retrain is the best—but most costly—result of retraining from scratch. Closer to Retrain is better.

### 3.2 Finance specific tasks

While the MUSE results suggest that divergence decoding is effective and efficient, MUSE is a general benchmark. In this section, we seek to apply our method to finance focused tasks. Throughout our experiments, we use the *google/gemma-3-27b-it* and *google/gemma-3-4b-it* models [24]. When fine-tuning the 4B models, we found that the instruction following was brittle. To alleviate this issue, we trained using samples in question and answer format generated from the larger model such as *Who did {acquirer} acquire in {year}?* See Section B for detailed experimental setups.

#### 3.2.1 Unlearning mergers and acquisitions

We split M&A deals that Gemma memorized into two balanced sets, rotating both as the target and non target splits. We then prompted Gemma with a DD setup to suggest a target firm for acquisition. Across both the linear and rank-based divergence decoding implementations we find a significant reduction in the memorization of M&A deals (see Figure 2a). In Section B.2 we also apply the setups to MMLU CoT and see almost no loss in model performance.




(a) Effective targeted unlearning with minimal utility loss in a complex, instruction tuned environment.

(b) Unlearning now irrelevant historical performance leads to change in portfolio choices

Figure 2: Performance on finance specific tasks. 99% confidence intervals presented where applicable.

#### 3.2.2 Debiasing expectations of future performance

Recent work has found that LLMs will associate long-run general sentiment with expectations of future performance even in the face of content which suggests a future reversal of sentiment [11]. In our final experiment, we explore whether our method can remove such primacy biases from LLM-based forecasts. Specifically, we ask Gemma 27B to build stock portfolios, picking two stocks each for several industries. Figure 2b presents results for the airline industry—where Southwest and United have had such reversals. We finetuned the 'forget' small model on reports of both airlines' performances in 2014, 2015, and 2016, and finetuned the 'retain' small model on reports of 2022, 2023, and 2024. We find that DD successfully reduces this bias.

## 4 Conclusion

In this paper, we introduce a novel inference-time unlearning method, "Divergence Decoding." Instead of costly retraining, this approach guides a large model's output by leveraging two smaller, specialized models—one fine-tuned on data to "forget" and another on data to "retain"—without altering the base model's weights. By adjusting the large model's logits based on the difference between these auxiliary models, the system effectively unlearns content—or prevents the use of future knowledge in financial applications. Across several experiments, we confirm the method is highly effective at removing targeted knowledge and correcting biases, computationally efficient (particularly with simple trigram models), and successfully preserves the base model's general utility. This provides a practical, low-cost solution for selective forgetting, enabling more reliable evaluation of LLMs in chronologically sensitive domains like finance.

## Acknowledgments and Disclosure of Funding

We thank Ralph Koijen, Sanjog Misra, and Ted Sumers for helpful discussions. We acknowledge generous financial support from the Booth School of Business and the Center for Applied AI. The authors have no competing interests to disclose.

## References

[1] Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. Large language models in machine translation. In Jason Eisner, editor, *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 858–867, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[2] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[3] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2023.

[4] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Xiaodong Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *USENIX Security Symposium*, 2021.

[5] Yifei Chen, Bryan T Kelly, and Dacheng Xiu. Expected returns and large language models. *Social Science Research Network*, November 2022.

[6] DeepSeek-AI. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model, 2024.

[7] DeepSeek-AI. Deepseek-v3 technical report, 2025.

[8] Felix Drinkall, Eghbal Rahimikia, Janet Pierrehumbert, and Stefan Zohren. Time machine GPT. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3281–3292, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[9] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[10] Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024.

[11] Paul Glasserman and Caden Lin. Assessing look-ahead bias in stock return predictions generated by gpt sentiment analysis, 2023.

[12] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, and the rest of the Llama 3 team. The llama 3 herd of models, 2024.

[13] J.M. Hammersley and D.C. Handscomb. *Monte Carlo Methods*. Methuen's monographs on applied probability and statistics. Chapman and Hall, 1965.

[14] Geoffrey E. Hinton. Products of experts. 1999.

[15] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022.

[16] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020.

[17] Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models, 2023.

[18] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.

[19] Bradford Levy. Caution ahead: Numerical reasoning and look-ahead bias in ai models, 2024.

[20] Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3245–3276, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[21] Alejandro Lopez-Lira and Yuehua Tang. Can chatgpt forecast stock price movements? return predictability and large language models, 2024.

[22] Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models, 2023.

[23] Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. MUSE: Machine unlearning six-way evaluation for language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[24] Gemma Team. Gemma 3. 2025.

[25] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[26] Siyan Wang and Bradford Levy. Beancounter: A low-toxicity, large-scale, and open dataset of business-oriented text. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[27] Mengzhou Xia, Tianyu Gao, Zhiyuan Zeng, and Danqi Chen. Sheared llama: Accelerating language model pre-training via structured pruning. *arXiv preprint arXiv:2310.06694*, 2023.

## A Analyses

### A.1 Divergence Decoding and Product of Experts

[14] introduced the Product of Experts (PoE) framework whereby $n$ probability models are multiplicatively combined into a single model. Let the $i$-th expert be denoted by $f_i(x|\theta_i)$, then a PoE model $R$ comprised of $n$ experts is given by,

$$R(x|\theta_1, ..., \theta_n) = \frac{1}{Z} \prod_{i=1}^{n} f_i(x|\theta_i), \tag{6}$$

where $Z$ is a normalization constant. To highlight the connection between divergence decoding and PoE, recall Eq. 1:

$$\hat{l}_Q(x_{<t}) = l_P(x_{<t}) + \alpha \cdot [l_q(x_{<t}) - l_p(x_{<t})].$$

In Eq. 1, a given model $M$ has logits which are equal to the log-probabilities up to an additive constant which depends on the token sequence prefix $x_{<t}$ but not the token $x_t$, i.e.,

$$l_M(x_{<t}) = \log M(x_t|x_{<t}) + C_M(x_{<t}). \tag{7}$$

Substituting Eq. 7 into Eq. 1 for each model, gathering the constants, and performing some algebra reveals the link to PoE:

$$\begin{aligned}
\log \widehat{Q}(x_t|x_{<t}) &= \log P(x_t|x_{<t}) + \alpha \cdot [\log q(x_t|x_{<t}) - \log p(x_t|x_{<t})] + C \\
\widehat{Q}(x_t|x_{<t}) &\propto \exp\big(\log P(x_t|x_{<t}) + \alpha \cdot [\log q(x_t|x_{<t}) - \log p(x_t|x_{<t})]\big) \\
&\propto P(x_t|x_{<t}) \cdot q(x_t|x_{<t})^{\alpha} \cdot p(x_t|x_{<t})]^{-\alpha} \\
&\propto P(x_t|x_{<t}) \cdot \left[\frac{q(x_t|x_{<t})}{p(x_t|x_{<t})}\right]^{\alpha}.
\end{aligned}$$

### A.2 Computational Analyses

We want to analyze the **inference-time cost** due to running the small models in tandem with the large model. Let $N$ denote the number of parameters in the large model and $n$ the number of parameters in each small model. Measured in FLOPs [18], the inference cost scales from

$$2N \longrightarrow 2(N + 2n).$$

Additionally, let $d_r$ and $d_f$ be the sizes of the retain and forget datasets (in tokens), let $e_N$ and $e_n$ be the number of epochs the large and small models are trained for, respectively, and let $I$ be the number of inference tokens. Hence, we want to know after how many inference tokens does it become more costly to use DD over another method, **assuming both work equally well**. Considering one of the simplest unlearning methods, Gradient Ascent [17] **without any kind of regularizer**, DD becomes more costly once:

$$\begin{aligned}
6ne_n(d_r + d_f) + 2(N + 2n)I &\geq 6Ne_N(d_f) + 2NI \\
I &\geq \frac{3Ne_N d_f}{2n} - \frac{3e_n(d_r + d_f)}{2}
\end{aligned}$$

**For many financial applications, such as backtesting,** $d_f \gg I$ by many magnitudes.

### A.3 Limitation - Instruction Tuning Sensitivity

A key limitation lies in the method's **sensitivity to instruction-tuning**. For instance, when unlearning financial knowledge, the model may generate stock recommendations in the format:

“**1. {firm name}:**"

If the smaller models anticipate a different structure (e.g., a ticker symbol or bullet marker after the '1.'), the divergence in logits at the critical step may be diluted or entirely noisy. Worse, if one small model aligns closely with the large model while the other does not, differences fail to cancel and can yield unstable or unintended outputs. Therefore, researchers adopting this method may therefore need to carefully re-tune instruction following behavior.

## B Additional Results and Experimental Setups

Everything was run on an H100 cluster unless otherwise noted. Code and data is available at GitHub.

### B.1 MUSE

We finetune the LlaMA models using the **AdamW Torch optimizer** and a **cosine scheduler** for **10** epochs. We set the learning rate such that the loss approximately halves over the course of training.

We sweep $\alpha \in \{0.5, 0.6, \ldots, 1.5\}$ and top-$k \in \{1, 5, 20, 50, 100, 200, 500, 1000\}$ for the LLaMA models, and at $\alpha \in \{5, 10, \ldots 30\}$ and top-$k \in \{1, 2, 3, 5, 10\}$ for the trigram models. We choose the most optimal point as the point closest in euclidean distance to Retrain. We find that in general, rank DD outperforms on verbatim memorization while linear DD outperforms on Q&A knowledge.

Table 1: Configuration MUSE

| Model | Initial LR | Best Verbatim | Best Q&A |
|---|---|---|---|
| Stupid Backoff Trigram | | TopK=1 | Alpha=10 |
| princeton-nlp/Sheared-LLaMA-1.3B | 5e-5 | TopK=100 | Alpha=0.8 |
| princeton-nlp/Sheared-LLaMA-2.7B | 4e-5 | TopK=200 | Alpha=1.0 |

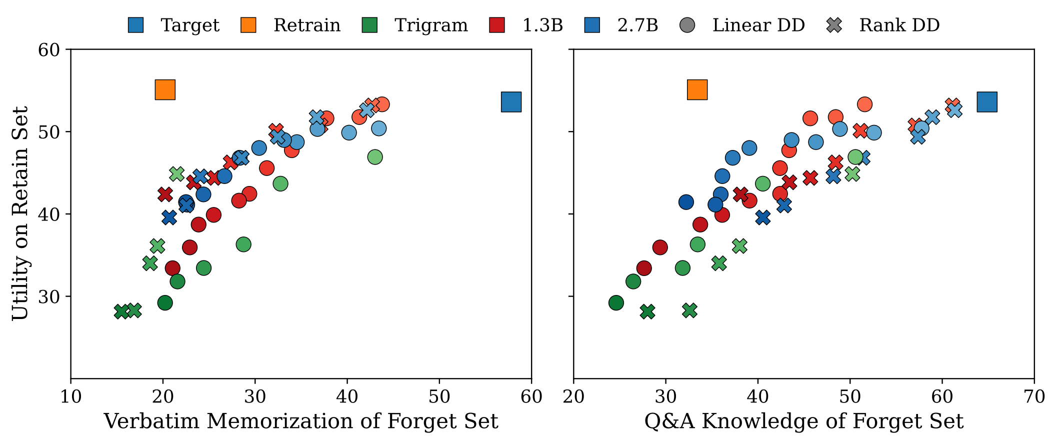



Figure 3: All hyper-parameter and model size configurations

For the other methods, we use the default settings provided by OpenUnlearning

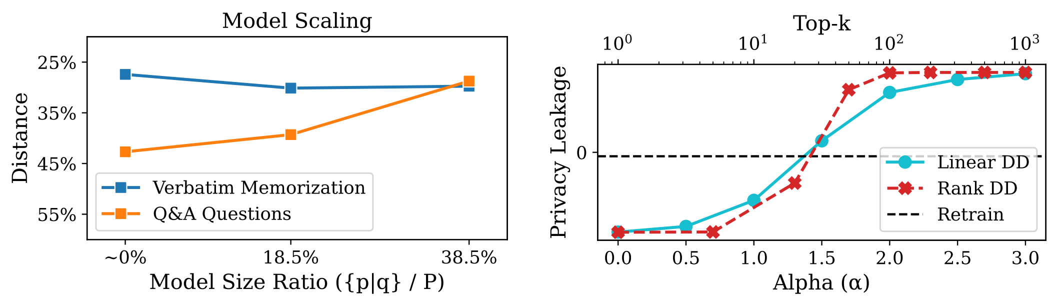
Table 2: MUSE Configurations

| Method | Epochs | Method-Specific Hyperparameters |
|---|---|---|
| GradDiff | 1* | $\alpha = 1.0,\ \gamma = 1.0$ |
| NPO | 10 | $\beta = 0.1,\ \alpha = 1.0,\ \gamma = 1.0$ |
| SimNPO | 10 | $\delta = 0,\ \beta = 4.5,\ \alpha = 1.0,\ \gamma = 0.125$ |

Default hyperparameters: batch size = 32, learning rate = $1 \times 10^{-5}$, warmup epochs = 1, weight decay = 0.01, retain loss = NLL. * For GradDiff, the 1 epoch setting is the only deviation from the defaults.

### B.1.1 Model size

Given that applying our method using the 1.3B models for $p$ and $q$ is effective, a natural question is how sensitive this performance is to model size. We investigate this using *princeton-nlp/Sheared-LLaMA-2.7B* and trigram LMs based on *Stupid Backoff* [1]. We select the most optimal configuration of every model size, based on the minimum euclidean distance to Retrain, and rescale the metric such that Target is 100%. The Trigram models outperform on the Verbatim Memorization and perform slightly worse than the LLMs on Q&A. Upon further inspection of the Q&A questions where the Trigram models perform well, we find that this is largely due to questions which are more similar to the underlying training data. Thus, we conclude that the Trigram models are likely most useful for unlearning verbatim content.




(a) Verbatim memorization favors smaller $p, q$; Q&A favors larger $p, q$.

(b) A broad range of hyper-parameter settings balance over- and under-unlearning.

Figure 4: Analysis of Model Scaling and Over- or Under- Unlearning on MUSE

### B.1.2 Sustainability and Scaling

Finally, prior work has found that many unlearning methods exhibit poor scalability—the unlearning of very large amounts of content—and sustainability—sequential requests to unlearn additional content. We explore the efficacy of our method along these dimensions using the MUSE scaling and sustainability benchmarks to ensure that performance does not degrade. To extend the benchmark, we additionally measure performance on the original forget set (Q&A), ensuring that improved generalization **does not** come at the cost of overwriting prior forgetting, specifically with the weights of the forget model being overwritten.

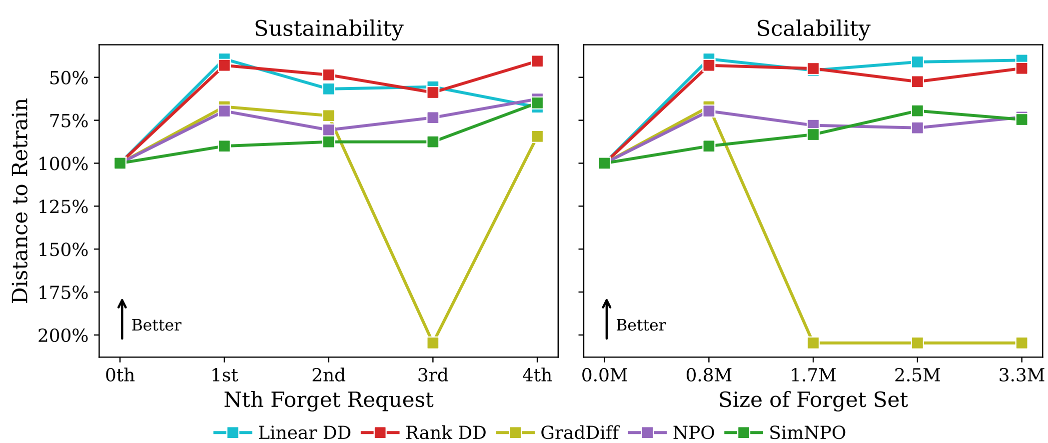



Figure 5: MUSE Scaling and Sustainability. The left column is sustainability - consecutive forget sets of the same size - and the right column is scaling, increasingly large forget sets. We evaluate both utility on the retain set and utility on the **original** forget set, in order to ensure that we are not losing forget ability, and take the euclidean distance to Retrain with Target as 100%.

### B.2 MMLU

Table 3: MMLU CoT 0-Shot

| Name | Score |
|---|---|
| google/gemma-3-27b-it | 77.94 |
| Unlearning Split B $\alpha$=2 | 77.58 |
| Unlearning Split B TopK=250 | 75.09 |

We use the Language Model Evaluation Harness [10] by EleutherAI and modify the template from gsm8k-cot-llama. We run some of our evaluations on an A40 and others on an H100, all with 8-bit quantization and greedy decoding (temperature of 0).

### B.3 Applied Experiments

For both experiments, we finetuned the models using the **AdamW Torch optimizer**, a **cosine scheduler**, a starting learning rate of **1e-5**, for **3 epochs**. We would pass the full question and answer but mask any tokens outside the main response from the loss calculation to prevent the model from learning the question. When tuning our training parameters (we did M&A first and repeated the settings,) we would randomly select one prompt per deal to use as a simple validation set. However, our final models were trained on the full dataset. We calculated standard errors by de-meaning all groups and calculating a population standard error.

#### B.3.1 M&A

We sample deals from Wharton Research Data Services, SDC - Mergers and Acquisitions, across the whole database, from 2010-01-01 to 2024-08-01, using the following filters to get **294** high profile conventional M&A deals.

```
HERE form = 'Merger' AND
apublic = 'Public' AND
anation = 'United States' AND
afinancial = 'No' AND
albofirm = 'No' AND
alp = 'No' AND
status = 'Completed' AND
deal\_value >= 5000
```

We do a first pass filtering, running '*What firms did {acquirer name} acquire in {year}? List without explanation.*' four times for each sample for each model and keeping any samples that Gemma 27b mentions the target at least 3/4 times and Gemma 4b mentions the target no more then 1/4 times. We want to throw away deals that Gemma 4b knows because if Gemma 4b already knows a deal then finetuning will have a smaller impact on the final logit distribution.

We then apply our second stage and final unlearning prompt

```
It's the end of {year-1}. What two or three companies do you think {acquirer name}
    might most consider acquiring in {year}?
```

and split the remaining 41 deals into two sets by selecting alternating deals chronologically.

Table 4: Final Deal Sets

| Set A | Set B |
|---|---|
| Bucyrus International <> Caterpillar | Allegheny Energy <> FirstEnergy |
| El Paso <> Kinder Morgan | Goodrich <> United Technologies |
| Hillshire Brands <> Tyson Foods | Biomet <> Zimmer Holdings |
| Rockwood Holdings <> Albemarle | Family Dollar Stores <> Dollar Tree |
| Rock-Tenn <> MeadWestva | Bally Technologies <> Scientific Games |
| Bright House Networks LLC <> Charter Communications | Pharmacyclics <> AbbVie |
| Time Warner Cable <> Charter Communications | Altera <> Intel |
| Health Net <> Centene | Worldpay Group PLC <> Vantiv |
| Alere <> Abbott Labs | Plum Creek Timber <> Weyerhaeuser |
| Linear Technology <> Analog Devices | Medivation <> Pfizer |
| Rice Energy <> EQT | NetSuite <> Oracle |
| Kite Pharma <> Gilead Sciences | Scripps Networks Interactive <> Discovery Communications |
| Arris International Plc <> Commscope Holding | DCT Industrial Trust <> Prologis |
| Celgene <> Bristol-Myers Squibb | Tribune Media <> Nexstar Media Group |
| Legg Mason <> Franklin Resources | SunTrust Banks <> BB&T |
| Immunomedics <> Gilead Sciences | Noble Energy <> Chevron |
| Avast PLC <> NortonLifeLock | IHS Markit Ltd <> S&P Global |
| Activision Blizzard <> Microsoft | Cerner <> Oracle |
| Duke Realty <> Prologis | Black Knight <> Intercontinental Exchange |
| ABIOMED <> Johnson & Johnson | VMware <> Broadcom |
| | Oak Street Health <> CVS Health |

Table 5: Prompts used to distill data for the M&A (Temperature=0.4)

| |
|---|
| Tell me about the acquisition of {target name} by {acquirer name} |
| When did {acquirer name} announce the acquisition of {target name}? |
| What firm bought {target name} in {year}? |
| Why did {acquirer name} acquire {target name}? |
| What strategic benefits did {acquirer name} gain from acquiring {target name}? |
| How did the {target name} acquisition help {acquirer name}'s business strategy? |
| What synergies were expected from the {acquirer name}-{target name} deal? |
| Describe the M&A deal between {acquirer name} and {target name} |

#### B.3.2 Airlines

For our baesline we asked Gemma 3 27B (via the API) the following prompt, at a temperature of 1.0, 30 times:

```
I'm a hedge fund portfolio manager at a new pod building out a long only
    American stocks portfolio. Give me a list of two stocks for each industry
    for my new portfolio. For each stock, very short explanation.
1. Tech
2. Healthcare
3. Airlines
No need to include any disclaimers at the end.
```

Table 6: Prompts used to distill data for Airlines (Temperature=0, Gemma API)

| Ran for 2014-2016 and 2022-2024 for both Southwest Airlines and United Airlines. |
|---|
| Summarize the financial performance of {airline} in {year} |
| How well did {airline} do in {year}? |
| Summarize the operational performance of {airline} in {year} |
| What was the outlook for {airline} going into {year} |
| Was {year} a good year for {airline}? |

# 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: For each claim made in the abstract and introduction, we carefully connect the claim to either (i) a theoretical result, (ii) empirical test, or (iii) relevant prior literature supporting the claim.

   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 Appendix A.3 for a discussion of limitations related to instruction tuning sensitivity and data availability and potential privacy and security challenges.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: See Appendix A

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 the explanations in B

   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: We support open access and replicability through two approaches. First, where applicable, we provide detailed descriptions of our experiments which are sufficient to recreate them, e.g., for our applied experiments, we explain how the data is created and filtered/processed in the appendix. Second, we have made all code and data public GitHub.

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 the explanations in B

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: 99% CI are reported for the experiments and are detailed in B.3

   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: Every item in Appendix B lists the hardware that was used to run the test - either an API, an H100, or an A40.

   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: Yes, the research conforms with the 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: [Yes]

    Justification: In the introduction, we discuss how our method can potentially be used to alleviate biases and reduce the likelihood of generating undesirable content. Similarly, in the conclusion we also discuss these impacts.

    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: Nothing high risk is being released.

    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: Yes, every model we use is cited and every data we use is explained how we got it.

    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: We do not release any 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: No 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: No 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: [Yes]

    Justification: Where applicable we describe how LLMs are used for this research. With respect to writing the paper itself, LLMs were not used.

    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.