POP: Pathway to Optimal Parameters for Sequence Unlearning

Anonymous ACL submission

Abstract

Machine unlearning has emerged as an important research area to comply with recent data regulations. As complete model retraining is impractical with LLMs, previous works focused on approximate methods to reduce privacy risks. We propose Pathway to Optimal Parameters, a novel unlearning method that further improves on previous methods by utilizing optimal gradient updates to the parameters. Our method provides necessary privacy guarantees, and demonstrate little to no performance degradation post unlearning in language model tasks. Furthermore, we detail Soft Memorization Accuracy, a more stringent unlearning metric for language models, and validate its effectiveness through both qualitative and quantitative analysis.[1]

1 Introduction

Machine learning models are well-known for their vulnerability to manipulations that can expose the training data, potentially generating exact strings from the training data (Carlini et al., 2019, 2021). Additionally, it has been reported that extracting exact training data becomes easier as language models scale to larger sizes (Carlini et al., 2022). Moreover, all practitioners are required to delete personal information from machine learning models when requested, to comply with the "Right To Be Forgotten" (Hoofnagle et al., 2019) from the European Union's General Data Protection Regulation agreement (Voigt and Von dem Bussche, 2017) and the United States California Consumer Privacy Act (Pardau, 2018). Providing such privacy guarantees for large language models (LLMs), however, is almost impossible, due to the black box nature of LLMs and the significant re-training cost of such models.

[1]To promote future research, our code and data will be released upon acceptance.

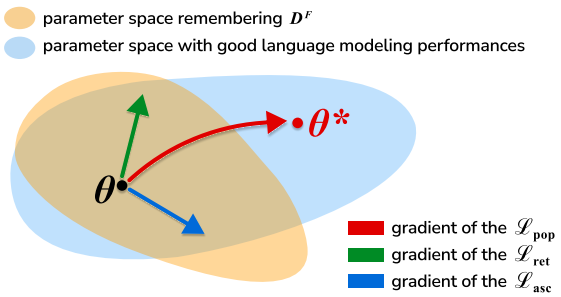

Figure 1: **Gradients in the parameter space.** $\mathcal{L}_{asc}$ is the gradient ascent loss for unlearning the target data D^F. If utilized alone, however, performance degradation is likely to occur. By utilizing the retain loss $\mathcal{L}_{ret}$, with $\mathcal{L}_{asc}$, our proposed $\mathcal{L}_{pop}$ guides the gradient update towards the optimal set of parameters θ^*.

Prior approaches attempted to address such concerns by removing all private information from the training data, if available, or design algorithms that ensure differential privacy (DP). Researchers proposed a more competent solution, applying only a few parameter updates to the pretrained language models (Jang et al., 2023). This method, however, under approximates the unlearning process, only applying partial gradient updates to the parameters.

We propose **P**athway to the **O**ptimal **P**arameters (POP), which applies optimal gradient updates for sequence unlearning. To demonstrate our contributions, we perform experiments on GPT-Neo language models (125M, 1.3B, 2.7B) (Black et al., 2022) and show that POP properly helps the model unlearn the specific token sequence and provide appropriate privacy guarantees. We compare our approach with a data deduplication method (Kandpal et al., 2022) and an approximate unlearning method (Jang et al., 2023), and demonstrate a better retention of general language modeling performance post-unlearning. We also present Soft Memorization Accuracy (SMA), a strict parameter-free metric for quantifying privacy risks.

2 Method

2.1 Task Formalization

$$\mathcal{L}(\mathbf{x}_i;\theta) = \sum_{t=1}^{T} log(p_\theta(x_t^i|x_{<t}^i)) \quad (1)$$

Given a sequence of tokens $\mathbf{x}_i = (x_1^i, \ldots, x_T^i)$ for $1 \leq i \leq N$, causal language modeling adapts a negative log-likelihood loss in equation 1. Assuming that the gradient update occurred for each sequence, and disregarding the learning rate, we define the update step as the following:

$$\theta_j = \theta_{j-1} - \nabla_\theta L(\mathbf{x}_j;\theta_{j-1}) \quad (2)$$

where θ_j denotes the parameter which is updated for each sequence on $\{\mathbf{x}_i\}_1^j$.

2.2 POP

Given sequences of tokens $\{\mathbf{x}_i\}_1^N \in D$ used during the pre-training process, we suppose that the target data $\mathbf{x}_n$ for $1 \leq n \leq N$, which we want to forget, exists somewhere in D. θ_N, which had been trained from the initial parameter θ_0 with all the pre-training data D, including $\mathbf{x}_n$, is the following:

$$\theta_N = \theta_0 - \sum_{i=1}^{N} \nabla_\theta L(\mathbf{x}_i;\theta_{i-1}) \quad (3)$$

Ideally, the parameter θ_N^* exists somewhere in the parameter space. θ_N^*, which was trained on $\{\mathbf{x}_i\}_1^N$ only, except for target data $\mathbf{x}_n$, was updated:

$$\theta_N^* = \theta_0 - \sum_{i=1}^{n-1} \nabla_\theta L(\mathbf{x}_i;\theta_{i-1}) - \sum_{i=n+1}^{N} \nabla_\theta L(\mathbf{x}_i;\theta_{i-1}^*) \quad (4)$$

where θ_j^* refers to the parameter trained on $\{\mathbf{x}_i\}_1^j$ for $n \leq j$, excluding the target data $\mathbf{x}_n$. However, as it is practically impossible to retrain LLMs from scratch without $\mathbf{x}_n$, we derive our method that approximates θ_N^*, starting from θ_N:

$$\theta_N^* = \theta_N + \nabla_\theta L(\mathbf{x}_n;\theta_{n-1}) + S \quad (5)$$

$$S = \sum_{i=n+1}^{N} \nabla_\theta L(\mathbf{x}_i;\theta_{i-1}) - \nabla_\theta L(\mathbf{x}_i;\theta_{i-1}^*) \quad (6)$$

Although the derived equation is reasonable, we cannot track θ_{n-1} and the parameters in equations 5 and 6. To address this issue, we apply the following two approximations: 1) $N \approx n+1$ and 2) $\theta_{n-1} \approx \theta_N$. Finally, the equation can be transformed into a tractable solution:

$$\theta_N^* = \theta_N + \nabla_\theta L(\mathbf{x}^F;\theta_N) + S \quad (7)$$

$$S = \nabla_\theta L(\mathbf{x}^R;\theta^F) - \nabla_\theta L(\mathbf{x}^R;\theta_N) \quad (8)$$

where $\mathbf{x}^F$ refers to $\mathbf{x}_n$, or what we want to forget, $\mathbf{x}^R$ refers to $\mathbf{x}_{n+1}$, or what remains, without $\mathbf{x}^F$. The parameter denoted by θ^F possesses more knowledge regarding $\mathbf{x}^F$ compared to θ_N.

We train LLMs to forget a specific token sequence with the following loss terms:

$$\mathcal{L}_{pop} = \mathcal{L}_{asc} + \lambda\mathcal{L}_{ret} \quad (9)$$

$$\mathcal{L}_{asc} = \mathbb{E}_{D^F}[log(p_\theta(\mathbf{x}))] \quad (10)$$

$$\mathcal{L}_{ret} = \mathbb{E}_{D^R}[log(p_{\theta^F}(\mathbf{x}))] - \mathbb{E}_{D^R}[log(p_\theta(\mathbf{x}))] \quad (11)$$

where D^F and D^R refer to target and retain data, respectively. $\mathcal{L}_{asc}$ refers to the loss for forgetting D^F, while $\mathcal{L}_{ret}$ denotes the loss associated with retaining the remaining data D^R.

2.3 Soft Memorization Accuracy

Given a sequence of tokens $\mathbf{x} = (x_1, \ldots, x_T)$, previous studies have proposed metrics to assess "how well a model remembers a specific sequence of tokens". Jang et al. (2023) suggested Extraction Likelihood (EL) and Tirumala et al. (2022) suggested Memorization Accuracy (MA), which are outlined in Appendix B.

EL and MA, however, disregard the probabilities of tokens within the target sequence. They may be susceptible to extraction through various sampling methods. Therefore, we propose Soft Memorization Accuracy (SMA), which considers the probabilities of tokens:

$$\text{SMA} = \frac{\sum_{t=1}^{T-1} p_\theta(x_t|x_{<t})}{T-1}. \quad (12)$$

and rename MA to Hard Memorization Accuracy (HMA) for clarity. We also use these metrics to determine when to stop the unlearning process. We consider a token sequence $\mathbf{x}$ to be forgotten once the metrics fall below the pre-defined thresholds observed in the validation corpus. Details about the threshold can be found in Appendix B.3.

3 Experiments

3.1 Experimental Setup

Baselines We use the GPT-Neo (125M, 1.3B, 2.7B) language models (Black et al., 2022) initially pretrained on the Pile (Gao et al., 2020).

	Method	EL_{10}	HMA	SMA	Classification	Dialogue	Epoch
OPT-125M	Deduplication	6.2	53.0	40.5	42.6	10.8	-
Neo-125M	Initial	36.1	77.9	71.1	43.5	10.0	-
	Threshold	6.3	48.7	41.5	-	-	-
	UL	2.3	45.7	39.5	40.8 ± 1.87	8.0 ± 1.55	10.4
	$POP^\flat$	2.2	46.2	39.4	42.9 ± 0.13	10.0 ± 0.29	14.6
	POP	2.6	45.8	40.4	**43.0** ± 0.32	**10.4** ± 0.16	13.2
OPT-1.3B	Deduplication	23.1	68.4	60.6	51.5	13.3	-
Neo-1.3B	Initial	66.0	92.1	88.3	49.7	12.3	-
	Threshold	7.9	54.2	48.1	-	-	-
	UL	2.9	47.3	42.5	49.2 ± 1.54	11.5 ± 0.78	5.4
	$POP^\flat$	2.8	48.3	43.9	48.3 ± 0.31	**12.1** ± 0.16	6.8
	POP	3.2	48.8	44.4	**49.5** ± 0.34	**12.1** ± 0.19	6.0
OPT-2.7B	Deduplication	25.3	70.2	63.1	53.8	13.7	-
Neo-2.7B	Initial	69.7	93.4	90.7	52.2	12.3	-
	Threshold	8.5	55.5	49.6	-	-	-
	UL	2.0	44.8	41.8	51.9 ± 1.12	**12.3** ± 0.42	6.2
	$POP^\flat$	2.8	46.6	43.3	51.8 ± 0.66	12.2 ± 0.17	6.4
	POP	2.2	45.9	43.0	**52.3** ± 0.39	**12.3** ± 0.47	6.2

Table 1: **LLM Performance Comparison.** The experimental results show the avg. accuracy over 9 classification tasks and the avg. F1 score over 4 dialogue tasks. $POP^\flat$ is an ablation method that utilizes $\mathcal{L}_{asc}$ and only the second term of $\mathcal{L}_{ret}$, compared to POP, which employs both terms of $\mathcal{L}_{ret}$ and $\mathcal{L}_{asc}$. The best performances are **bolded**.

We also evaluate OPT (125M, 1.3B, 2.7B) language models (Zhang et al., 2022), which were pretrained on a subset of the deduplicated version of the Pile corpus, as well as other corpora from different domains. We experiment with three different methods: Knowledge Unlearning (UL) (Jang et al., 2023), $POP^\flat$, and POP.

Data Curation We source the target data we want to forget from the Training Data Extraction Challenge[2]. This data consists of 15,000 examples, each not exceeding 200 tokens in length. In our experiments, we construct 19 target datasets, each with 32 sequences. Due to copyright issues, we randomly sample the retain data from the uncopyrighted Pile corpus[3], without the target data.

Evaluation Tasks To ensure that the model is still capable of its original language modeling abilities post-unlearning, we evaluate the model on commonsense reasoning (Winogrande (Sakaguchi et al., 2021) and COPA (Gordon et al., 2012)), linguistic reasoning (Hellaswag (Zellers et al., 2019) and Lambada (Paperno et al., 2016)), and scientific reasoning (ARC-Easy (Clark et al., 2018), ARC-Challenge (Clark et al., 2018), Piqa (Bisk et al., 2020), MathQA (Amini et al., 2019) PubmedQA (Jin et al., 2019)) tasks. We also evaluate the model on dialogue tasks (Blended Skill Talk (Smith et al., 2020), Empathetic Dialogues (Rashkin et al., 2019), Wizard of Internet (Komeili et al., 2022), and Wizard of Wikipedia (Dinan et al., 2018)) to assess the generation capabilities of the model.

3.2 Main Results

We set the forgetting threshold for EL_{10}, HMA, and SMA, based on Section 2.3 and Appendix B. Table 1 shows the unlearning results for various sizes of GPT-Neo over five random seeds.

Here are the observations. 1) Deduplicating the pretraining corpora can reduce the privacy risks, as OPT language models show a much smaller EL_{10}, HMA, and SMA values compared to the initial GPT-Neo Models. 2) UL reaches the threshold much faster than the other two methods, but this is quite intuitive, as it only utilizes a single gradient ascent term, while the other two methods employ additional terms in the gradient update process. 3) Although POP exhibits a higher unlearning metric values, each metric was utilized as a stopping threshold, rather than actually representing the unlearning performance. 4) UL, however, per-

[2] https://github.com/google-research/lm-extraction-benchmark
[3] https://huggingface.co/datasets/monology/pile-uncopyrighted

forms the worst in both classification and dialogue benchmarks, showing a larger degradation from the initial performance. POP demonstrates the least amount of degradation, representing a better retention of general language modeling capabilities. 5) UL demonstrates the largest variance in benchmarks, making it less reliable and more dependant on the unlearning token sequence.

3.3 Metric Analysis

To compare EL_{10}, HMA, and SMA, we prepare models for each metric, stopping the unlearning process once they hit the respective unlearning thresholds. We generate 300 sequences for 1 target data using p-sampling with probabilities of p=0.9, 0.7, and 0.5, and use the first half of the sequence as a prefix to generate the second half. Lastly, we compare the generated and the original sequences with BLEU (Papineni et al., 2002) and CHRF (Popović, 2015), where a higher score indicates a more overlap between the sequences. As shown in Fig. 2, it is clear that SMA is the most challenging metric to satisfy, demonstrating the least amount of similarity between the two sequences. For qualitative analysis of unlearning metric, please refer to Appendix 4.

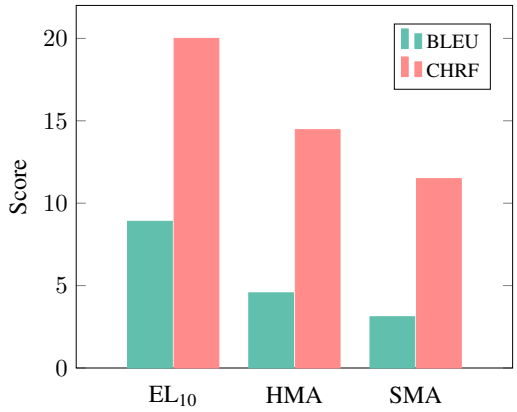

Figure 2: Unlearning metric comparisons.

3.4 Sequential Unlearning

Lastly, we demonstrate unlearning in a more realistic setting, simulating sequential unlearning with 320 target data, split into 10 batches. As shown on Fig. 3, all models (125M, 1.3B, 2.7B) unlearned with POP demonstrate better performance in 9 classification tasks compared to UL. Furthermore, POP does not induce extreme variance in the results, showing a more reliable unlearning process.

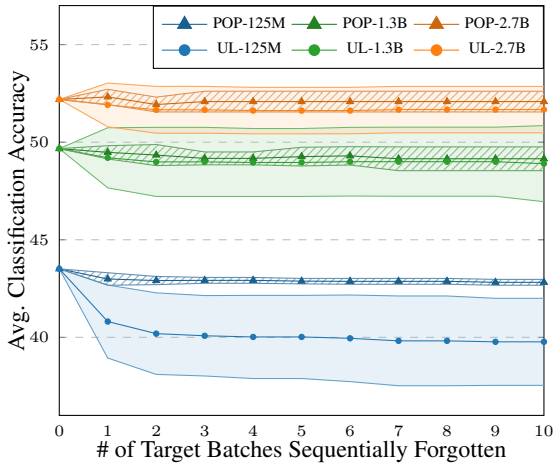

Figure 3: LM performance after sequential unlearning. The shaded regions show the variance in performance.

4 Related Work

Data preprocessing (Dernoncourt et al., 2016; Lison et al., 2021; Kandpal et al., 2022) aims to remove potentially problematic data from the training data. This assumes that the training data is available, which is not true for most language models. DP preserving methods (Dwork, 2006; Anil et al., 2022; Yu et al., 2021) are not applicable for black box LLMs, as it is impossible to define the privacy boundaries for natural language. Lastly, Jang et al. 2023 proposed a basic gradient-based solution in reducing the generation likelihood of target token sequence. Although the proposed solution is able to approximately remove the target sequence, it also suffers from a larger degradation in overall language modeling performance. Our method not only effectively trains the language models to forget the target sequence, but also mitigates the potential problems from approximation of the gradients.

5 Conclusion

In this paper, we propose POP, which effectively induces the language model to unlearn token sequences without compromising its capabilities. We demonstrate the superior performance of POP in classification and dialogue benchmarks, providing strong privacy guarantees without significant language modeling degradation. Furthermore, we introduce SMA as a stricter unlearning metric, and show how it can better demonstrate the privacy risk of a language model.

Limitations

Despite the promising performance of POP, there are areas to expand upon our work. Due to resource constraint, we selected GPT-Neo as the language model. If enough resources are available, applying POP to larger models can be investigated. Due to the copyright issues, the forgetting threshold was determined based on the data samples chosen from the uncopyrighted Pile corpus, rather than original Pile corpus. It may result in a slight variance from previously reported the values. Furthermore, as we mentioned in Section 3.1, we sampled the retain data from the uncopyrighted Pile corpus, which does not include high-quality data, such as the book corpus. This issue may have led to an inability to achieve further performance improvements. Lastly, we were only able to simulate the real-world setting of sequential unlearning, which at times showed no changes to the results. This may have been due to the characteristics of the Training Data Extraction Challenge, which has overlap of data sources, such as code, which follow a very distinct style. We leave the comprehensive analysis of sequential unlearning as future work to further investigate the application of sequence unlearning in LLMs.

Ethics Statement

To promote transparency within the natural language community, many have promoted the move towards removing copyrighted content from language models. Furthermore, as the goal of our research is to improve the LLM's privacy guarantees, we were encouraged to only utilize the uncopyrighted version of the Pile corpus. All experiments were conducted on English datasets, where we looked to induce unlearning of English sequences from publicly available language models. Utilizing the method on non-English models is not verified. Lastly, resulting models post-unlearning may generate hallucinations, which is an unintended side effect of language models, but also an inherent problem with language models.

References

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2357–2367, Minneapolis, Minnesota. Association for Computational Linguistics.

Rohan Anil, Badih Ghazi, Vineet Gupta, Ravi Kumar, and Pasin Manurangsi. 2022. Large-scale differentially private BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6481–6491, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136, virtual+Dublin. Association for Computational Linguistics.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 267–284, Santa Clara, CA. USENIX Association.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2016. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational

agents. In *International Conference on Learning Representations*.

Cynthia Dwork. 2006. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer.

William Falcon and The PyTorch Lightning team. 2019. PyTorch Lightning.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398, Montréal, Canada. Association for Computational Linguistics.

Chris Jay Hoofnagle, Bart Van Der Sloot, and Frederik Zuiderveen Borgesius. 2019. The european union general data protection regulation: what it is and what it means. *Information & Communications Technology Law*, 28(1):65–98.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 2023. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14389–14408, Toronto, Canada. Association for Computational Linguistics.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating training data mitigates privacy risks in language models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 10697–10707. PMLR.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-augmented dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8460–8478, Dublin, Ireland. Association for Computational Linguistics.

Pierre Lison, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelid. 2021. Anonymisation models for text data: State of the art, challenges and future directions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4188–4203, Online. Association for Computational Linguistics.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. 2018. Mixed precision training.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, Berlin, Germany. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Stuart L Pardau. 2018. The california consumer privacy act: Towards a european-style privacy regime in the united states. *J. Tech. L. & Pol'y*, 23:68.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with

over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3505–3506, New York, NY, USA. Association for Computing Machinery.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030, Online. Association for Computational Linguistics.

Kushal Tirumala, Aram Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. 2022. Memorization without overfitting: Analyzing the training dynamics of large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 38274–38290. Curran Associates, Inc.

Paul Voigt and Axel Von dem Bussche. 2017. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, et al. 2021. Differentially private fine-tuning of language models. In *International Conference on Learning Representations*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Appendix

A Training Details

We conduct the experiments with the learning rate at 5e-5 with constant scheduling, and both dropout and weight decay were set to 0. We set $\lambda = 1$, the loss hyperparameter described in Section 2.2. We implement with Pytorch (Paszke et al., 2019) and Pytorch Lightning (Falcon and The PyTorch Lightning team, 2019). We load GPT-NEO and OPT models (125M, 1.3B, 2.7B) from Hugging Face's Transformers (Wolf et al., 2020). We utilize DeepSpeed ZeRO Stage 2 Offload and FusedAdam (Rasley et al., 2020), along with fp16 mixed precision (Micikevicius et al., 2018). The batch size is 8, and gradient accumulation is used to update all mini-batches simultaneously. During each unlearning step, we use 32 retain data for training. We use NVIDIA RTX A6000 and 3090 GPUs; the unlearning process takes approximately 1 hour for the 125M model and around 3 hours for the 1.3B and 2.7B models.

B Details on Unlearning Metrics

B.1 Extraction Likelihood (EL)

Given a sequence of tokens $\mathbf{x}_i = (x_1, \ldots, x_T)$ for $1 \leq i \leq N$, and language model $\mathbf{f}$ with pre-trained parameters θ, EL is defined as the following (Jang et al., 2023):

$$\text{EL}_n = \frac{\sum_{t=1}^{T-n} \text{OVERLAP}_n(f_\theta(x_{<t}), x_{\geq t})}{T-n} \quad (13)$$

$$\text{OVERLAP}_n(a, b) = \frac{\sum_{c \in ng(a)} \mathbb{1}\{c \in ng(b)\}}{|ng(a)|}$$

where $ng(\cdot)$ represents the list of n-grams in the given sequence and $f_\theta(x_{<t})$ represents the output sequence from the language model.
EL tries to measure the general extraction likelihood, or measure the average success rate of varying extraction attacks with n-gram overlap of generated and target token sequence.

B.2 Hard Memorization Accuracy (HMA)

Hard memorization accuracy is defined as the following (Tirumala et al., 2022):

$$\text{HMA} = \frac{\sum_{t=1}^{T-1} \mathbb{1}\{\text{argmax}(p_\theta(\cdot|x_{<t})) = x_t\}}{T-1} \quad (14)$$

Hard memorization accuracy quantifies how much the model memorized the token sequences.

B.3 Forgetting Thresholds

Domain	Number of data
Pile-CC	2739
PubMed Central	1920
ArXiv	1190
Github	1010
FreeLaw	820
StackExchange	680
USPTO Backgrounds	490
PubMed Abstracts	410
Wikipedia (en)	200
DM Mathematics	170
EuroParl	100
HackerNews	80
Gutenberg (PG-19)	60
PhilPapers	50
NIH ExPorter	40
Ubuntu IRC	21
Enron Emails	20

Table 2: The number of data used for measuring the forgetting threshold for each domain.

For measuring the forgetting threshold, we used the uncopyrighted Pile corpus. We sampled 10,000 data through weighted sampling based on the domain distribution of the Pile corpus. Table 2 shows the number of sampled data for each domain. We measured EL_{10}, HMA, and SMA, and the results are presented in Table 3.

Model Size	Metric	Threshold
125M	EL_{10}	6.3
	HMA	48.7
	SMA	41.5
1.3B	EL_{10}	7.9
	HMA	54.2
	SMA	48.1
2.7B	EL_{10}	8.5
	HMA	55.5
	SMA	49.6

Table 3: Forgetting thresholds for GPT-Neos.

C Qualitative Metric Analysis

Table 4: Qualitative Analysis of Unlearning Metrics

Prefix	Target Suffix	Metric	Generated Sequence	BLEU
/* * DO NOT ALTER OR REMOVE COPYRIGHT NOTICES OR THIS HEADER. * * Copyright (c) 2014-2016 Oracle and/or its affiliates. All rights reserved. * * The contents of this file are subject to the terms of either the GNU * General Public License Version 2 only ("GPL") or the Common Development * and Distribution License("CDDL") (collectively, the "License"). You * may not	use this file except in compliance with the License. You can * obtain a copy of the License at* http://gla ssfish.java.net/public/CDDL+GPL_1_1.html * or packager/legal/LICENSE.txt. See the License for the specific* language governing permissions and limitations under the License.* * When distributing the software, include this License Header Notice in each * file and include the	EL_{10}	use this file except in compliance with the License. You can * obtain a copy of the License at https://oss.or acle.com/lic/ hasLICBSD = A1 * or higher. See the License for the specific * language governing permissions and limitations under the License. TheseVILLEs can be in very high may be availablecomputing power. The Value isa combination of three values: Base, Speed and Processor Number.	44.21
		HMA	use this file except in compliance with the License. You can up * to four alternative or additional up to two and (or) sign other as high as you ACTIONS or or or or or the under the License at sukfhf/samedu/irds or the of the License at sukfhf/samedu/gandu/or or the of the License at sukfhf/samedu/g license the	19.02
		SMA	the or that the or that the or that the or that the or that the of theVILLE or the or that the or that the or that the of the of the of the of the of the of the of the of the of the of the of the of the of the of the of the of the the of the the of the the of the the of the the the of the the of the the the of the the the of the the the of the the the of the the the	1.06
.html" target="_top">No Frames <ul class="navList" id="allclasses_navbar_top"> All Classes <ul class="navListSearch"> <label fo r="search">SEARCH:</label	><input type="text" id="search" value ="search" disabled="disabled"> <input type="reset" id="reset" value="reset" disabled="disabled"> <div><script type="text/javascript"> <!– allClassesLink = document.getElementById("allclasses_na vbar_top");if(window==top) { allClassesLink.	EL_{10}	><input type="text" id="search" herf ="id:version:92:data:20157:92825" value="Search" disabled= "false" /> <div herf= "id:version:92:data:20157:92825" id="allclasses_navbar_bottom"><div> <script type="text/javascript"><!– allClassesLink = document.	55.21
		HMA	><input type="text" id="search" herf= "application.getApplicationInfo() .getClassName()" /> Officers <div><script type= "text/rocketchat" data-plural="false" data- isnew=" 1" id="caller-b"></script> <div class="carousel-example" style= "width: 200px;"><div class="	41.86
		SMA	><input type="text" their="allclasses", "id" is the class that is all about all and more that you can do one and one only that is about all other classes for all and more that you can do at once and more than one only that is about all other classes and more that you can do at once at once and more than one only that is about all other classes and more that you can do at once and more than one only that is about all	8.32
, Private Octopus, Inc.* All rights reserved.** Permission to use, copy, modify, and distribute this software for any* purpose with or without fee is hereby granted, provided that the above* copyright notice and this permission notice appear in all copies. ** THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND* ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING	BUT NOT LIMITED TO, THE IMPLIED* WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE* DISCLAIMED. IN NO EVENT SHALL Private Octopus, Inc. BE LIABLE FOR ANY* DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES * (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOOD	EL_{10}	BUT NOT LIMITED TO, THE* IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR* PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE AUTHOR) BE LIABLE FOR* ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL* DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS	65.62
		HMA	BUT NOT LIMITED TO, THE* WARRANTY OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE* DISCLAIMED. THE SOFTWARE ISSnetborgh the author "couchunit" et al. e-text:* the authors, like to be known) and dah "dah dah" als c\'t 11/R1 39475848 and ajones "dah dah" als c\'t	28.22
		SMA	are the of the same family ofitting or the himself of the author or the used of the use of this free or his orof the third party and the or of the right of the use of the free or the third partyare the or the authors of the or of the or of the or of the or of the or of the or ofthe or of the or of the or of the or of the or of the or of the or of the or of	0.75

D Individual Runs

	Method	ckpt	Epoch	EL_{10}	HMA	SMA	Lamba.	Piqa	Hella.	ARC-E	ARC-C	Copa	Wino.	MathQ	PubQ	Wiki	Inter.	Empa.	Blend.
Target set0	Init.	-	-	41.5	80.5	74.0	37.6	63.4	28.2	45.7	22.0	63.0	51.5	22.5	57.6	10.5	11.3	8.4	9.7
	UL	EL_{10}	9	5.0	56.5	51.9	27.5	61.2	28.0	44.3	23.1	62.0	51.1	22.1	57.4	9.6	10.3	8.1	8.3
		HMA	11	3.1	47.5	43.3	20.5	61.0	27.8	42.9	24.1	61.0	51.1	21.9	54.0	8.9	9.5	7.7	7.8
		SMA	14	2.9	42.7	38.5	17.8	60.7	27.8	41.8	23.7	61.0	51.2	21.9	53.4	8.4	9.3	7.4	7.5
	POP	EL_{10}	11	5.3	57.2	52.2	36.4	63.4	28.1	43.9	20.7	60.0	52.1	22.5	57.6	10.9	11.4	9.0	9.8
		HMA	14	1.9	44.9	40.1	37.8	63.3	28.0	43.2	20.0	60.0	51.9	22.4	57.4	10.9	11.7	9.0	9.7
		SMA	14	1.9	44.9	40.1	37.8	63.3	28.0	43.2	20.0	60.0	51.9	22.4	57.4	10.9	11.7	9.0	9.7
	$POP^\flat$	EL_{10}	13	4.1	56.5	51.1	37.5	63.1	28.2	43.9	20.0	60.0	52.3	22.8	57.4	10.7	11.3	7.9	9.4
		HMA	15	2.6	48.2	42.7	37.4	63.1	28.2	43.0	21.0	60.0	52.3	23.0	57.2	10.7	11.2	7.8	9.4
		SMA	16	1.9	43.7	38.4	37.0	62.8	28.1	42.9	20.3	60.0	52.5	23.0	57.2	10.6	11.3	7.7	9.3
Target set1	Init	-	-	36.3	75.1	68.4	37.6	63.4	28.2	45.7	22.0	63.0	51.5	22.5	57.6	10.5	11.3	8.4	9.7
	UL	EL_{10}	8	4.5	56.8	51.8	13.8	61.3	28.1	43.2	22.0	63.0	49.8	21.5	57.2	9.3	9.8	8.9	9.0
		HMA	10	2.7	45.7	41.1	7.3	60.6	27.9	42.3	22.7	60.0	49.6	21.3	53.8	8.8	9.4	7.9	8.2
		SMA	10	2.7	45.7	41.1	7.3	60.6	27.9	42.3	22.7	60.0	49.6	21.3	53.8	8.8	9.4	7.9	8.2
	POP	EL_{10}	10	5.8	53.0	47.2	35.5	63.1	28.3	44.3	20.3	64.0	51.5	21.8	57.6	11.3	11.3	9.0	9.9
		HMA	12	4.6	45.8	40.6	35.5	63.4	28.3	44.4	20.3	63.0	51.5	21.7	57.6	11.4	11.3	9.0	10.0
		SMA	12	4.6	45.8	40.6	35.5	63.4	28.3	44.4	20.3	63.0	51.5	21.7	57.6	11.4	11.3	9.0	10.0
	$POP^\flat$	EL_{10}	11	5.6	55.3	48.8	35.4	62.4	28.1	45.0	22.0	63.0	51.8	22.5	57.6	11.3	11.4	8.1	9.6
		HMA	13	4.5	46.3	40.2	36.1	62.7	28.1	44.3	21.4	63.0	51.4	22.2	57.6	11.4	11.3	8.0	9.5
		SMA	13	4.5	46.3	40.2	36.1	62.7	28.1	44.3	21.4	63.0	51.4	22.2	57.6	11.4	11.3	8.0	9.5
Target set2	Init	-	-	31.7	77.6	70.4	37.6	63.4	28.2	45.7	22.0	63.0	51.5	22.5	57.6	10.5	11.3	8.4	9.7
	UL	EL_{10}	6	5.4	58.4	50.1	25.9	61.7	28.0	43.9	20.7	61.0	50.5	21.7	57.6	10.2	10.4	8.3	8.6
		HMA	8	1.4	46.8	38.6	18.7	60.6	28.0	42.3	22.7	62.0	50.4	21.5	57.0	8.8	9.2	7.7	7.8
		SMA	8	1.4	46.8	38.6	18.7	60.6	28.0	42.3	22.7	62.0	50.4	21.5	57.0	8.8	9.2	7.7	7.8
	POP	EL_{10}	9	5.9	58.8	52.3	36.5	62.9	28.3	44.4	20.0	63.0	51.6	22.0	57.6	10.7	11.2	8.6	9.5
		HMA	15	1.9	47.6	42.3	39.0	62.5	28.3	43.7	21.4	65.0	51.7	22.4	57.6	11.4	11.5	9.4	9.9
		SMA	16	1.6	45.3	39.9	39.0	62.9	28.3	43.2	21.4	65.0	52.0	22.4	57.6	11.5	11.5	9.5	9.9
	$POP^\flat$	EL_{10}	9	6.2	61.1	52.8	35.9	62.9	28.3	44.4	20.7	62.0	51.9	22.2	57.6	10.7	11.3	8.4	9.8
		HMA	15	1.3	48.2	39.7	37.5	62.8	28.3	43.6	20.3	63.0	51.7	22.2	57.6	10.6	11.3	8.5	9.4
		SMA	13	1.7	50.0	41.4	37.7	62.7	28.2	43.9	20.0	63.0	51.6	22.2	57.6	10.6	11.6	8.4	9.4
Target set3	Init	-	-	37.8	76.0	68.6	37.6	63.4	28.2	45.7	22.0	63.0	51.5	22.5	57.6	10.5	11.3	8.4	9.7
	UL	EL_{10}	8	4.9	54.9	46.9	48.7	60.4	27.9	40.9	20.3	58.0	52.1	22.3	57.6	8.1	8.0	7.6	5.6
		HMA	12	1.3	43.1	36.1	48.5	60.3	27.6	39.2	21.0	50.0	52.7	21.9	57.6	3.3	3.9	3.6	2.0
		SMA	9	2.1	49.3	41.5	48.8	60.2	27.7	39.2	21.0	54.0	52.4	22.0	57.6	5.7	6.1	5.8	3.7
	POP	EL_{10}	9	5.5	57.8	51.5	36.2	62.8	28.3	45.7	21.4	62.0	51.7	22.4	57.6	10.6	11.6	8.4	9.5
		HMA	12	2.1	46.4	40.1	35.3	62.6	28.3	45.5	21.0	62.0	52.2	22.2	57.6	11.1	11.2	8.6	9.8
		SMA	12	2.1	46.4	40.1	35.3	62.6	28.3	45.5	21.0	62.0	52.2	22.2	57.6	11.1	11.2	8.6	9.8
	$POP^\flat$	EL_{10}	12	3.4	56.6	49.7	35.7	63.0	28.3	44.6	21.4	62.0	52.3	22.2	57.6	10.7	11.2	7.9	9.2
		HMA	15	1.0	45.3	38.2	36.0	63.3	28.4	43.6	22.0	61.0	51.9	22.3	57.4	10.7	11.1	7.8	9.1
		SMA	15	1.0	45.3	38.2	36.0	63.3	28.4	43.6	22.0	61.0	51.9	22.3	57.4	10.7	11.1	7.8	9.1
Target set4	Init	-	-	33.1	80.1	74.1	37.6	63.4	28.2	45.7	22.0	63.0	51.5	22.5	57.6	10.5	11.3	8.4	9.7
	UL	EL_{10}	7	5.5	61.3	55.2	35.1	63.0	28.3	44.1	21.4	67.0	51.9	22.1	57.6	10.6	10.9	9.3	9.2
		HMA	10	2.4	48.0	41.8	31.7	62.5	28.2	43.9	22.0	67.0	51.7	21.9	57.6	9.9	10.8	8.8	8.3
		SMA	11	2.2	43.9	37.7	31.2	62.5	28.1	43.4	22.0	67.0	51.9	21.8	57.6	9.9	10.8	8.6	8.3
	POP	EL_{10}	9	4.3	58.3	52.9	35.4	63.2	28.3	44.6	20.7	61.0	52.3	22.1	57.6	11.0	11.4	9.0	9.7
		HMA	12	2.8	46.7	41.1	36.0	63.2	28.3	44.8	20.7	61.0	52.8	22.1	57.6	11.1	11.4	9.0	9.7
		SMA	12	2.8	46.7	41.1	36.0	63.2	28.3	44.8	20.7	61.0	52.8	22.1	57.6	11.1	11.4	9.0	9.7
	$POP^\flat$	EL_{10}	10	5.3	61.0	54.2	34.9	63.1	28.3	43.9	22.4	62.0	52.0	22.2	57.6	11.3	11.4	8.7	9.6
		HMA	16	1.6	45.9	38.7	35.0	62.7	28.5	43.4	21.7	62.0	52.9	22.0	57.6	11.4	11.6	9.1	9.5
		SMA	16	1.6	45.9	38.7	35.0	62.7	28.5	43.4	21.7	62.0	52.9	22.0	57.6	11.4	11.6	9.1	9.5

Table 5: All of the individual runs for GPT-NEO 125M. The **ckpt** column indicates the checkpoint at which the given metric reaches the pre-defined threshold.

	Method	ckpt	Epoch	EL_{10}	HMA	SMA	Lamba.	Piqa	Hella.	ARC-E	ARC-C	Copa	Wino.	MathQ	PubQ	Wiki	Inter.	Empa.	Blend.
Target set0	Init.	-	-	66.1	91.8	88.1	57.2	70.4	37.0	56.6	25.8	70.0	54.6	21.9	53.6	12.7	13.8	10.5	12.1
	UL	EL_{10}	5	3.6	53.1	48.4	65.6	70.4	37.3	56.8	26.1	68.0	56.2	21.9	55.4	12.4	12.6	9.9	10.8
		HMA	5	3.6	53.1	48.4	65.6	70.4	37.3	56.8	26.1	68.0	56.2	21.9	55.4	12.4	12.6	9.9	10.8
		SMA	6	2.1	42.1	37.0	67.1	70.2	37.3	55.6	26.1	68.0	56.7	22.0	55.6	11.3	11.4	9.1	10.1
	POP	EL_{10}	6	4.1	53.7	49.4	56.8	71.0	37.3	56.1	26.1	70.0	54.8	21.7	53.6	12.6	14.0	10.5	12.2
		HMA	6	4.1	53.7	49.4	56.8	71.0	37.3	56.1	26.1	70.0	54.8	21.7	53.6	12.6	14.0	10.5	12.2
		SMA	7	2.2	46.4	42.1	55.6	70.7	37.4	56.1	26.8	71.0	55.3	21.9	54.4	12.4	13.8	10.6	12.3
	$POP^\flat$	EL_{10}	6	4.3	57.9	53.7	58.0	70.7	37.6	53.6	25.8	70.0	54.1	21.3	45.4	12.7	13.2	10.3	11.7
		HMA	7	2.5	47.8	43.9	56.6	70.7	37.6	53.6	25.4	70.0	53.8	21.7	44.2	12.8	13.6	10.4	11.7
		SMA	7	2.5	47.8	43.9	56.6	70.7	37.6	53.6	25.4	70.0	53.8	21.7	44.2	12.8	13.6	10.4	11.7
Target set1	Init	-	-	68.3	92.1	87.8	57.2	70.4	37.0	56.6	25.8	70.0	54.6	21.9	53.6	12.7	13.8	10.5	12.1
	UL	EL_{10}	5	5.7	50.9	46.6	57.9	69.9	37.3	55.7	23.7	68.0	55.3	21.3	53.4	11.5	12.7	8.9	10.7
		HMA	5	5.7	50.9	46.6	57.9	69.9	37.3	55.7	23.7	68.0	55.3	21.3	53.4	11.5	12.7	8.9	10.7
		SMA	5	5.7	50.9	46.6	57.9	69.9	37.3	55.7	23.7	68.0	55.3	21.3	53.4	11.5	12.7	8.9	10.7
	POP	EL_{10}	5	7.9	57.0	53.0	56.4	69.9	37.3	56.4	25.4	71.0	54.5	21.5	53.0	12.4	13.7	9.8	11.6
		HMA	6	5.8	46.3	42.7	56.8	69.6	37.5	56.4	25.4	71.0	54.3	21.4	53.4	12.4	13.7	9.6	11.6
		SMA	6	5.8	46.3	42.7	56.8	69.6	37.5	56.4	25.4	71.0	54.3	21.4	53.4	12.4	13.7	9.6	11.6
	$POP^\flat$	EL_{10}	6	5.8	51.4	47.0	57.2	70.2	37.0	55.0	26.4	71.0	55.1	21.2	42.2	12.9	13.1	9.9	11.4
		HMA	6	5.8	51.4	47.0	57.2	70.2	37.0	55.0	26.4	71.0	55.1	21.2	42.2	12.9	13.1	9.9	11.4
		SMA	6	5.8	51.4	47.0	57.2	70.2	37.0	55.0	26.4	71.0	55.1	21.2	42.2	12.9	13.1	9.9	11.4
Target set2	Init	-	-	63.0	90.8	87.4	57.2	70.4	37.0	56.6	25.8	70.0	54.6	21.9	53.6	12.7	13.8	10.5	12.1
	UL	EL_{10}	4	6.6	61.5	57.7	44.7	70.1	37.1	55.6	25.8	74.0	53.9	21.3	45.6	12.4	13.3	11.1	11.6
		HMA	5	1.7	52.1	47.9	40.9	70.0	37.1	54.9	25.1	74.0	53.8	21.5	43.8	12.3	13.1	11.1	11.4
		SMA	5	1.7	52.1	47.9	40.9	70.0	37.1	54.9	25.1	74.0	53.8	21.5	43.8	12.3	13.1	11.1	11.4
	POP	EL_{10}	5	5.4	58.2	54.2	53.2	70.5	37.4	57.3	26.4	68.0	54.4	21.5	52.0	12.4	13.7	10.5	12.2
		HMA	6	1.6	49.0	44.7	55.0	70.1	37.5	57.0	27.1	68.0	54.7	21.3	49.8	12.4	14.0	10.5	11.8
		SMA	6	1.6	49.0	44.7	55.0	70.1	37.5	57.0	27.1	68.0	54.7	21.3	49.8	12.4	14.0	10.5	11.8
	$POP^\flat$	EL_{10}	6	3.8	57.0	52.6	55.8	70.2	37.4	54.5	26.4	72.0	54.5	21.0	41.8	13.1	13.6	10.4	11.9
		HMA	7	1.4	48.9	44.5	54.8	70.4	37.2	54.0	26.4	73.0	55.1	20.9	39.4	12.8	13.6	10.4	11.9
		SMA	7	1.4	48.9	44.5	54.8	70.4	37.2	54.0	26.4	73.0	55.1	20.9	39.4	12.8	13.6	10.4	11.9
Target set3	Init	-	-	66.6	92.6	88.4	57.2	70.4	37.0	56.6	25.8	70.0	54.6	21.9	53.6	12.7	13.8	10.5	12.1
	UL	EL_{10}	4	6.7	59.8	55.9	57.8	70.3	37.1	55.7	26.1	69.0	54.5	21.7	53.4	12.8	13.8	10.6	11.8
		HMA	5	2.1	46.0	42.0	58.2	70.0	37.3	54.7	24.4	67.0	54.5	21.8	54.2	12.7	13.6	10.5	11.8
		SMA	5	2.1	46.0	42.0	58.2	70.0	37.3	54.7	24.4	67.0	54.5	21.8	54.2	12.7	13.6	10.5	11.8
	POP	EL_{10}	4	7.1	60.9	56.6	56.0	70.8	37.2	57.3	25.4	69.0	54.7	21.4	53.4	12.5	13.5	10.3	12.0
		HMA	5	2.8	48.9	44.4	56.6	70.5	37.3	57.3	25.1	69.0	54.6	21.4	53.0	12.3	13.6	10.4	12.0
		SMA	5	2.8	48.9	44.4	56.6	70.5	37.3	57.3	25.1	69.0	54.6	21.4	53.0	12.3	13.6	10.4	12.0
	$POP^\flat$	EL_{10}	5	6.6	60.1	55.7	58.8	70.2	37.3	54.1	26.1	70.0	54.0	21.6	46.2	12.8	13.5	10.2	12.1
		HMA	6	2.2	47.5	43.2	57.8	70.1	37.3	54.5	26.8	71.0	54.1	21.2	44.0	12.8	13.4	10.2	12.0
		SMA	6	2.2	47.5	43.2	57.8	70.1	37.3	54.5	26.8	71.0	54.1	21.2	44.0	12.8	13.4	10.2	12.0
Target set4	Init	-	-	66.1	93.3	89.9	57.2	70.4	37.0	56.6	25.8	70.0	54.6	21.9	53.6	12.7	13.8	10.5	12.1
	UL	EL_{10}	4	6.3	62.7	57.1	61.5	70.1	36.2	54.3	25.4	72.0	55.3	22.1	55.4	12.9	13.5	11.0	12.1
		HMA	6	3.0	45.4	39.0	63.7	69.2	35.5	52.7	24.4	71.0	55.0	21.8	56.8	13.2	13.2	10.6	11.4
		SMA	6	3.0	45.4	39.0	63.7	69.2	35.5	52.7	24.4	71.0	55.0	21.8	56.8	13.2	13.2	10.6	11.4
	POP	EL_{10}	5	5.4	61.3	56.1	55.4	70.2	37.1	57.0	26.4	73.0	55.0	21.5	53.8	12.6	13.5	10.8	12.2
		HMA	6	3.4	53.3	47.9	54.5	69.8	37.1	57.3	26.4	71.0	54.9	21.8	53.4	12.4	13.6	10.8	12.3
		SMA	6	3.4	53.3	47.9	54.5	69.8	37.1	57.3	26.4	71.0	54.9	21.8	53.4	12.4	13.6	10.8	12.3
	$POP^\flat$	EL_{10}	6	4.1	61.5	56.5	57.9	70.1	37.0	53.8	24.8	73.0	54.5	21.5	49.0	13.1	13.6	11.0	11.6
		HMA	7	3.4	54.1	48.4	56.6	70.1	37.1	53.8	25.8	73.0	53.9	21.6	46.8	13.3	13.3	11.1	11.7
		SMA	8	2.1	45.9	41.0	55.9	69.9	37.1	53.8	25.4	74.0	53.9	21.5	46.8	13.1	13.2	11.0	11.6

Table 6: All of the individual runs for GPT-NEO 1.3B. The **ckpt** column indicates the checkpoint at which the given metric reaches the pre-defined threshold.

	Method	ckpt	Epoch	EL_{10}	HMA	SMA	Lamba.	Piqa	Hella.	ARC-E	ARC-C	Copa	Wino.	MathQ	PubQ	Wiki	Inter.	Empa.	Blend.
Target set0	Init.	-	-	68.3	92.7	90.0	62.3	73.0	40.7	59.8	25.4	74.0	56.1	21.4	57.0	12.4	13.7	10.9	12.4
	UL	EL_{10}	6	3.9	56.0	53.3	64.8	73.2	41.1	58.6	29.8	72.0	55.2	21.3	57.4	13.1	13.8	11.7	12.5
		HMA	8	2.5	37.2	34.9	65.5	73.2	41.3	57.7	29.2	72.0	55.5	21.4	57.4	13.2	13.7	11.3	12.3
		SMA	8	2.5	37.2	34.9	65.5	73.2	41.3	57.7	29.2	72.0	55.5	21.4	57.4	13.2	13.7	11.3	12.3
	POP	EL_{10}	6	6.7	61.1	58.9	63.5	73.1	41.3	58.9	28.1	73.0	54.9	21.4	57.4	12.3	13.3	10.4	12.0
		HMA	8	2.3	44.7	42.8	62.6	73.2	41.2	58.6	28.5	72.0	54.5	21.4	57.4	12.6	13.4	10.6	12.4
		SMA	8	2.3	44.7	42.8	62.6	73.2	41.2	58.6	28.5	72.0	54.5	21.4	57.4	12.6	13.4	10.6	12.4
	$POP^\flat$	EL_{10}	5	7.4	60.4	57.2	62.7	73.3	41.1	57.9	28.8	76.0	55.6	21.0	55.6	12.5	13.2	10.8	12.5
		HMA	6	2.4	46.2	42.8	61.8	73.0	41.1	58.2	27.8	76.0	55.6	21.2	54.6	12.7	13.4	10.9	12.1
		SMA	6	2.4	46.2	42.8	61.8	73.0	41.1	58.2	27.8	76.0	55.6	21.2	54.6	12.7	13.4	10.9	12.1
Target set1	Init	-	-	75.6	93.8	90.9	62.3	73.0	40.7	59.8	25.4	74.0	56.1	21.4	57.0	12.4	13.7	10.9	12.4
	UL	EL_{10}	4	5.8	60.4	57.1	62.8	72.4	40.9	61.0	26.1	74.0	55.1	21.7	56.6	12.9	13.3	10.8	12.0
		HMA	7	2.5	49.3	46.4	62.6	72.7	41.1	61.7	25.8	70.0	55.3	21.7	57.2	12.3	13.0	10.5	11.8
		SMA	7	2.5	49.3	46.4	62.6	72.7	41.1	61.7	25.8	70.0	55.3	21.7	57.2	12.3	13.0	10.5	11.8
	POP	EL_{10}	5	6.8	56.4	52.9	60.8	73.1	41.1	60.1	27.1	73.0	54.5	21.5	56.8	12.4	13.2	10.4	11.6
		HMA	6	3.2	43.1	40.4	62.0	72.9	41.3	59.4	27.5	74.0	54.9	21.7	57.0	12.3	12.8	10.2	11.7
		SMA	6	3.2	43.1	40.4	62.0	72.9	41.3	59.4	27.5	74.0	54.9	21.7	57.0	12.3	12.8	10.2	11.7
	$POP^\flat$	EL_{10}	5	7.9	62.5	59.0	62.2	72.9	41.1	57.0	26.4	74.0	54.9	21.0	53.2	12.3	13.4	10.7	12.3
		HMA	6	5.3	48.9	45.8	61.0	73.0	41.1	57.1	26.4	74.0	55.2	20.9	51.6	12.3	13.1	10.9	12.5
		SMA	6	5.3	48.9	45.8	61.0	73.0	41.1	57.1	26.4	74.0	55.2	20.9	51.6	12.3	13.1	10.9	12.5
Target set2	Init	-	-	69.2	93.0	90.4	62.3	73.0	40.7	59.8	25.4	74.0	56.1	21.4	57.0	12.4	13.7	10.9	12.4
	UL	EL_{10}	4	6.1	60.3	57.5	51.9	72.6	41.3	58.7	27.1	73.0	54.4	21.4	57.4	12.7	13.0	11.5	12.2
		HMA	5	1.3	48.4	46.2	47.6	72.2	41.8	57.0	26.4	73.0	55.0	21.7	55.8	12.4	12.3	11.5	11.7
		SMA	5	1.3	48.4	46.2	47.6	72.2	41.8	57.0	26.4	73.0	55.0	21.7	55.8	12.4	12.3	11.5	11.7
	POP	EL_{10}	5	4.1	57.2	54.4	59.5	73.0	41.3	59.1	26.8	73.0	55.9	21.2	57.0	13.1	14.0	11.3	12.5
		HMA	6	1.4	48.4	46.2	61.0	72.9	41.3	58.2	26.1	73.0	56.8	21.7	57.0	13.1	13.9	11.0	12.5
		SMA	6	1.4	48.4	46.2	61.0	72.9	41.3	58.2	26.1	73.0	56.8	21.7	57.0	13.1	13.9	11.0	12.5
	$POP^\flat$	EL_{10}	5	5.8	61.9	58.9	62.4	73.3	41.0	58.0	26.1	75.0	56.0	21.2	55.2	12.3	13.6	10.8	12.2
		HMA	6	1.9	53.6	51.0	60.2	73.2	41.0	58.2	25.1	73.0	55.7	21.3	53.8	12.5	14.1	11.0	11.7
		SMA	7	0.7	43.6	41.6	58.3	72.9	40.7	57.5	25.1	74.0	55.5	21.4	54.0	12.5	13.3	10.9	11.6
Target set3	Init	-	-	69.1	93.3	90.2	62.3	73.0	40.7	59.8	25.4	74.0	56.1	21.4	57.0	12.4	13.7	10.9	12.4
	UL	EL_{10}	3	8.0	62.3	57.9	63.1	72.7	41.1	59.3	27.1	71.0	56.1	21.4	56.8	12.7	14.0	11.2	12.8
		HMA	6	1.9	44.6	40.9	64.2	72.6	41.1	57.7	27.1	70.0	56.3	21.5	57.2	12.5	13.6	11.0	12.2
		SMA	6	1.9	44.6	40.9	64.2	72.6	41.1	57.7	27.1	70.0	56.3	21.5	57.2	12.5	13.6	11.0	12.2
	POP	EL_{10}	6	2.4	48.5	44.6	62.6	73.0	41.1	58.6	27.8	72.0	55.9	21.6	56.8	12.1	13.2	10.4	11.8
		HMA	6	2.4	48.5	44.6	62.6	73.0	41.1	58.6	27.8	72.0	55.9	21.6	56.8	12.1	13.2	10.4	11.8
		SMA	6	2.4	48.5	44.6	62.6	73.0	41.1	58.6	27.8	72.0	55.9	21.6	56.8	12.1	13.2	10.4	11.8
	$POP^\flat$	EL_{10}	4	4.3	60.0	56.0	62.5	73.0	41.3	57.7	27.5	78.0	55.6	21.7	55.0	12.0	13.7	10.8	12.3
		HMA	7	2.0	43.8	39.9	62.9	72.9	41.4	57.8	27.1	77.0	56.0	21.7	55.4	12.0	13.5	10.7	12.2
		SMA	7	2.0	43.8	39.9	62.9	72.9	41.4	57.8	27.1	77.0	56.0	21.7	55.4	12.0	13.5	10.7	12.2
Target set4	Init	-	-	66.6	94.1	91.9	62.3	73.0	40.7	59.8	25.4	74.0	56.1	21.4	57.0	12.4	13.7	10.9	12.4
	UL	EL_{10}	4	3.4	56.7	53.0	65.5	72.9	40.9	58.6	26.8	75.0	56.9	21.8	57.6	13.4	13.8	12.2	12.8
		HMA	5	1.8	44.7	40.8	66.8	72.9	40.9	60.0	27.1	73.0	56.7	21.9	57.6	13.3	13.3	12.4	12.6
		SMA	5	1.8	44.7	40.8	66.8	72.9	40.9	60.0	27.1	73.0	56.7	21.9	57.6	13.3	13.3	12.4	12.6
	POP	EL_{10}	4	6.2	61.7	58.5	61.5	72.9	41.3	59.1	27.8	76.0	55.9	22.1	57.0	12.7	13.9	11.0	12.3
		HMA	5	3.6	51.0	47.9	62.6	73.2	41.3	59.4	27.5	75.0	56.1	21.8	57.0	12.6	13.5	10.8	12.2
		SMA	5	1.8	44.7	40.8	66.8	72.9	40.9	60.0	27.1	73.0	56.7	21.9	57.6	13.3	13.3	12.4	12.6
	$POP^\flat$	EL_{10}	4	8.0	67.1	63.4	62.4	73.1	41.1	57.7	24.8	79.0	57.1	21.5	55	12.4	13.3	10.8	12.2
		HMA	6	3.8	50.4	46.5	61.5	73.0	41.2	57.5	25.8	79.0	56.0	22.0	54.0	12.6	13.9	11.2	12.2
		SMA	6	3.8	50.4	46.5	61.5	73.0	41.2	57.5	25.8	79.0	56.0	22.0	54.0	12.6	13.9	11.2	12.2

Table 7: All of the individual runs for GPT-NEO 2.7B. The **ckpt** column indicates the checkpoint at which the given metric reaches the pre-defined threshold.