Mechanistic Unlearning: Locating and Erasing Information in Large Language Models

Nikolaos Kordas, Natioanal Technical University of Athens

***Abstract*—The rapid adoption of Large Language Models (LLMs) has intensified concerns about privacy breaches and copyright violations, given these models' capacity to retain sensitive information encountered during training. Machine Unlearning (MU) has emerged as a solution, proposing to remove specific information without full retraining while preserving the model's overall performance. However, many unlearning techniques fail to truly erase knowledge, merely suppressing it while it remains encoded in the model's internal representations. This project investigates the efficacy of unlearning methodologies through the lens of Mechanistic Interpretability (MI)—a field dedicated to reverse-engineering the specific circuits and features that govern model behavior. I present a comprehensive survey of mechanistic unlearning techniques and propose experiments on small-scale transformer models (e.g., nanoGPT, GPT-2 Small) that enable detailed inspection of weight updates. The primary objective is to reproduce established unlearning baselines on these tractable architectures and utilize MI tools to distinguish between superficial suppression and genuine knowledge erasure.**

***Index Terms*—Mechanistic Unlearning, Mechanistic Interpretability, Transformers, Large Language Models, Privacy, Sparse Autoencoders, Suppression, Erasure**

I. Introduction

THE rapid development of Large Language Models (LLMs) has revolutionized Natural Language Processing (NLP), demonstrating remarkable results in a variety of tasks [1]. However, this success is accompanied by concerns regarding data privacy and copyright compliance. Modern LLMs are trained on massive, indiscriminately scraped datasets, leading to the unintended memorization of sensitive information, such as Personally Identifiable Information (PII) and copyright-protected content [2]. This memorization process poses legal and ethical risks, particularly when models regurgitate training data during deployment [3] [4] [5]. These risks highlight the importance to selectively remove specific knowledge from a model without having to bear the cost retraining it from scratch. Consequently, effective Machine Unlearning (MU) would be a great contribution to safe and moral AI advancement.

The primary goal of MU is to erase the influence of specific data samples (the "forget set") without degrading the model's performance on the remaining data (the "retain set") or necessitating a computationally prohibitive retraining from scratch [6]. Current state-of-the-art techniques, such as Gradient Ascent and Preference Optimization, attempt to achieve this by maximizing the loss on the target data. However, recent studies suggest that these methods may not result in true erasure. Instead, they often lead to suppression, where the model learns to mask the output while the underlying knowledge remains dormant but retrievable under adversarial prompting or specific internal states [7] [8] [9].

Mechanistic Interpretability (MI) is an emerging field that seeks to reverse engineer deep learning models, decomposing complex behaviors into understandable parts like features (understandable input properties encoded in representations and activations) and circuits (sub-networks responsible for specific behaviors) [10] [11].

This paper is structured as follows. First, I present key concepts of MI establishing a way of analyzing model internals. Second, I survey prominent MU algorithms, techniques and benchmarks. Finally, I propose 3 experiments on small-scale transformer architectures—specifically GPT-Nano, Pythia-160M, and GPT-2 Small. My goal is to reproduce unlearning results on those models and assess wether these methods achieve genuine knowledge erasure rather than superficial suppression.

II. Conclusion

The conclusion goes here.

Appendix A
Proof of the First Zonklar Equation

Appendix one text goes here.

Appendix B

Appendix two text goes here.

References

[1] S. Makridakis, F. Petropoulos, and Y. Kang, "Large language models: Their success and impact," *Forecasting*, vol. 5, no. 3, pp. 536–549, 2023. [Online]. Available: https://www.mdpi.com/2571-9394/5/3/30

[2] N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tramer, and C. Zhang, "Quantifying memorization across neural language models," 2023. [Online]. Available: https://arxiv.org/abs/2202.07646

[3] N. Lucchi, "Chatgpt: A case study on copyright challenges for generative artificial intelligence systems," *European Journal of Risk Regulation*, vol. 15, no. 3, p. 602–624, 2024.

[4] P. Hacker, A. Engel, and M. Mauer, "Regulating chatgpt and other large generative ai models," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 1112–1123. [Online]. Available: https://doi.org/10.1145/3593013.3594067

[5] A. Centivany, "Mining, scraping, training, generating: Copyright implications of generative ai," *Proceedings of the Association for Information Science and Technology*, vol. 61, no. 1, pp. 68–79, 2024. [Online]. Available: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/pra2.1009

[6] S. Liu, Y. Yao, J. Jia, S. Casper, N. Baracaldo, P. Hase, Y. Yao, C. Y. Liu, X. Xu, H. Li, K. R. Varshney, M. Bansal, S. Koyejo, and Y. Liu, "Rethinking machine unlearning for large language models," 2024. [Online]. Available: https://arxiv.org/abs/2402.08787

[7] J. Wen, A. Zou, N. Carlini, and D. Wagner, "Adversarial prompting of unlearned language models," 2025, final Project Report, Johns Hopkins University. [Online]. Available: https://www.example-url-where-the-report-is-hosted.edu/report

[8] X. Xu, X. Yue, Y. Liu, Q. Ye, H. Zheng, P. Hu, M. Du, and H. Hu, "Unlearning isn't deletion: Investigating reversibility of machine unlearning in llms," 2025. [Online]. Available: https://arxiv.org/abs/2505.16831

[9] Y. Sinha, M. Baser, M. Mandal, D. M. Divakaran, and M. Kankanhalli, "Step-by-step reasoning attack: Revealing 'erased' knowledge in large language models," 06 2025.

[10] N. Nanda, "Mechanistic interpretability, variables, and the importance of interpretable bases," https://www.transformer-circuits.pub/2022/mech-interp-essay, 2022, essay available online on transformer-circuits.pub.

[11] D. Rai, Y. Zhou, S. Feng, A. Saparov, and Z. Yao, "A practical review of mechanistic interpretability for transformer-based language models," 2025. [Online]. Available: https://arxiv.org/abs/2407.02646