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. Mechanistic 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 often 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. My work focuses on small-scale transformer models (e.g. GPT-Nano, GPT-2 Small) to enable 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

THIS demo file is intended to serve as a "starter file" for IEEE journal papers produced under LaTeX using IEEEtran.cls version 1.8b and later. I wish you the best of success.

mds
August 26, 2015

A. Subsection Heading Here

Subsection text here.

1) Subsubsection Heading Here: Subsubsection text here.

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.

Acknowledgment

The authors would like to thank...

References

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.

PLACE
PHOTO
HERE

Michael Shell Biography text here.

John Doe Biography text here.

Jane Doe Biography text here.