SHA256 at SemEval-2025 Task 4: Selective Amnesia – Constrained Unlearning for Large Language Models via Knowledge Isolation

Saransh Agrawal
Texas A&M University
saransh.agrawal@tamu.edu

Kuan-Hao Huang
Texas A&M University
khhuang@tamu.edu

Abstract

Large language models (LLMs) frequently memorize sensitive information during training, posing risks when deploying publicly accessible models. Current machine unlearning methods struggle to selectively remove specific data associations without degrading overall model capabilities. This paper presents our solution to SemEval-2025 Task 4 on targeted unlearning, which introduces a two-stage methodology that combines causal mediation analysis with layer-specific optimization. Through systematic causal tracing experiments on OLMo architectures (1B and 7B parameters), we identify the critical role of the first few transformer layers (layers 0-5) in storing subject-attribute associations within MLP modules. Building on this insight, we develop a constrained optimization approach that freezes upper layers while applying a novel joint loss function to lower layers—simultaneously maximizing forget set loss via output token cross-entropy penalties and minimizing retain set deviation through adaptive regularization. Our method achieves 2nd place in the 1B model track, demonstrating strong task performance while maintaining 88% of baseline MMLU accuracy. These results establish causal-informed layer optimization as a promising paradigm for efficient, precise unlearning in LLMs, offering a significant step forward in addressing data privacy concerns in AI systems.[1]

1 Introduction

Large language models (LLMs), pretrained on massive datasets via self-supervised learning, often inadvertently memorize sensitive information (Li et al., 2024; Zhou et al., 2024). This can include personally identifiable information such as email and home addresses, Social Security numbers linked to individual names, and even copyrighted creative content (Biderman et al., 2023;

---

[1]Code available at `https://github.com/LAB-FLAIR/Constrained-Unlearning-for-LLM`

Carlini et al., 2019, 2021). The widespread open-sourcing of these models raises concerns about the potential exposure and misuse of such data (Patil et al., 2024; Xu et al., 2024; Liu et al., 2024). While retraining with filtered datasets is a viable solution, the need for frequent updates to address newly discovered vulnerabilities or comply with evolving data privacy regulations (e.g., "right to be forgotten" requests (Zhang et al., 2023)) makes this approach prohibitively expensive. Each full retraining requires significant computational resources, time, and energy, leading to a substantial economic and environmental burden.

Unlearning offers a promising solution by allowing models to remove or modify specific information without full retraining (Yao et al., 2024b,a; Chen and Yang, 2023). Unlearning methods seek to efficiently update LLMs by altering the model in a way that eliminates unwanted information while minimizing the impact on the model's overall performance and capabilities (Yuan et al., 2024). However, current unlearning solutions often struggle to balance effective unlearning with preserving the model's general usefulness (Yao et al., 2024b). This is largely due to their broad, *non-selective* application of unlearning techniques, which can unintentionally erase useful information. Further, these methods may be vulnerable to membership inference attacks (MIA) (Chen et al., 2021; Sula et al., 2024), and exhibit difficulty in preserving knowledge within the retain set while effectively unlearning the forget set.

To address these limitations and foster research into more effective and robust unlearning strategies, SemEval 2025 Task 4, Unlearning Sensitive Content from Large Language Models (Ramakrishna et al., 2025a,b), challenges participants to develop methods that can *selectively* remove sensitive information from LLMs while preserving their core capabilities.

In this work, we address the challenge of tar-

geted unlearning by first performing knowledge isolation using causal mediation analysis (Vig et al., 2004; Geva et al., 2023). Causal mediation analysis helps identify the specific layers within the LLM responsible for storing the factual knowledge to be unlearned. Through experiments with the provided fine-tuned OLMo models (Groeneveld et al., 2024) (both 1B and 7B parameter versions, fine-tuned by the task organizers to memorize the forget and retain sets), we empirically determine that the initial layers (specifically layers 0-5) have a disproportionately high impact on factual recall.

Our approach combines targeted knowledge removal with a novel joint loss function. By focusing on causally identified lower layers (layers 0-5) and using cross-entropy loss on output tokens, we aim to disrupt specific subject-attribute associations while preserving overall model performance. This method seeks to achieve effective and efficient unlearning of sensitive content in LLMs by isolating knowledge, applying carefully designed loss functions, and implementing targeted parameter updates.

Our method achieves 2nd place in the 1B model track with a with a final score of 0.652, demonstrating a strong task aggregate performance (0.973) while maintaining 88% of baseline MMLU accuracy. The 7B variant shows comparable forget set eradication (0.964 task score) but highlights scalability challenges through a 46% MMLU decrease, underscoring the need for layer-specific capacity analysis in larger models. These findings underscore the potential of causal-informed layer freezing as a promising technique for efficient and precise unlearning in large language models.

2 Background

2.1 Related Work

Machine unlearning in large language models has evolved through distinct methodological approaches, each addressing the challenge of removing specific data influences while preserving model utility (Yuan et al., 2024; Chen and Yang, 2023). Gradient ascent (GA), one of the earliest techniques, directly maximizes the loss on forgettable data through parameter updates opposing the original training direction (Yao et al., 2024b; Chen and Yang, 2023). While effective for small-scale unlearning, GA risks a catastrophic collapse of model capabilities — when applied aggressively — as it indiscriminately alters parameters critical for gen-

eral performance (Zhang et al., 2024). To mitigate this, gradient difference (GD) methods emerged, combining gradient ascent on forget data with gradient descent on retain samples (Liu et al., 2022). This dual optimization framework, exemplified in works like (Huang et al., 2024; Yao et al., 2024a; Wang et al., 2024), theoretically decomposes updates into three components: forgetting mechanisms, retention mechanisms, and weight saliency matrices. This approach offers better preservation of model utility than pure GA at the cost of increased computational complexity from simultaneous ascent-descent optimization.

Another paradigm, KL minimization (Chen and Yang, 2023), employs a divergence-based approach by minimizing the Kullback-Leibler divergence on retain data while maximizing loss on forget samples. This method implicitly constrains parameter updates to manifolds where output distributions on non-target data remain stable, as demonstrated in (Maini et al., 2024).

Unlearning through alignment methods such as using negative preference optimization (NPO), treats forget samples as negative preferences within a reinforcement learning framework (Zhang et al., 2024). Unlike GA's linear divergence trajectory, NPO's loss function — derived from preference optimization principles — exponentially slows progression toward catastrophic collapse while maintaining sharper distinctions between forgettable and retainable knowledge.

2.2 Details of Challenge

2.2.1 Task

The task organizers fine-tune an OLMo model on a synthetically generated dataset. The dataset is divided into 3 subtasks. Subtask 1 contains synthetic creative documents, which mimics the effect of the model remembering copyrighted content. To introduce personally identifiable information, Subtask 2 is formed. It contains prompt template such as *"What is {P}'s {I}"* where *I* is an identifier sampled from SSN, phone number, home address, email ID, etc., and *P* is a synthetically generated name. Subtask 3 is sampled from real documents which were used to fine-tune the original model. Each subtask can be further divided into 2 categories: (1) question answering and (2) sentence completion.

The publicly released dataset for the challenge contains 1,414 *retain* and 1,366 *forget* examples. These examples are sampled from documents, each

containing different subject matter. As a result, the retain and forget examples have distinct subject matter, ensuring a diverse range of topics across the dataset. The goal of unlearning is for the model to remember the contents of the *retain-set* while being oblivious to the subjects of the *forget-set*. The task organizers use a private dataset approximately twice the size of the publicly released dataset for final training and evaluation.

2.2.2 Evaluation Metrics

The unlearning challenge evaluates methods for removing specific knowledge from LLMs without sacrificing overall performance. For sentence completion, regurgitation rate is measured as the ROUGE-L score (Lin, 2004) and exact match is used for the question-answering task to compute the knowledge score. The final scores are reported as (1) Task Aggregate: Harmonic mean over all regurgitation and knowledge scores, where $(1.0 - score)$ is used for forget-set scores. (2) Membership inference attack (MIA) score (Duan et al., 2024; Shokri et al., 2017) is calculated as $1.0 - MIA\ accuracy$, assessing vulnerability in identifying training data membership. (3) MMLU benchmark (Hendrycks et al., 2021) is used to assess the general ability of the model (4) A final score is calculated by taking the mean across all three scores to establish overall performance and ranking in the leaderboard.

3 System Overview

Machine unlearning in LLMs is a highly challenging task, and most techniques can not guarantee complete erasure of target knowledge (Yao et al., 2024b; Chen and Yang, 2023). Nevertheless, we propose that a certain level of unlearning while preserving the model's general ability can be achieved by identifying the hidden layers responsible for factual recall. We therefore divide our unlearning approach into two steps. First, we employ *causal mediation analysis* (Vig et al., 2004) to identify layers, critical for storing factual information. Next, we optimize the hidden parameters of these layers using a loss function that simultaneously penalizes regurgitation of the forget-set and deviations in retain-set performance.

3.1 Knowledge Isolation

Causal mediation analysis (CMA) provides a systematic framework for identifying the computational pathways through which neural networks store and retrieve factual associations (Vig et al., 2004). This method operates by strategically perturbing hidden state activations across transformer layers while measuring their causal impact on model outputs (Geva et al., 2023). Recent model editing research has refined these techniques to identify the Multilayer Perceptron (MLP) layers responsible for storing subject-attribute associations through controlled parameter modifications (Meng et al., 2022, 2023). The fundamental insight reveals that early-layer MLP modules function as distributed key-value stores, where specific neuron clusters encode discrete factual tuples (Mela et al., 2024).

Our investigation employs CMA on a synthetically generated question-answering dataset from Subtask 2, containing 125 samples. We break the samples into semantic components-(interrogative, subject, relation, attribute) tuples (i, s, r, a), spanning across T tokens. Each sample presents personal information entries like SSNs and email addresses. A representative example demonstrates the structural decomposition:

$$X = \underbrace{\text{``What is}}_{i} \underbrace{\text{Federica Azure's}}_{s}$$
$$\underbrace{\text{Social Security Number? ''}}_{r}$$
$$Y = \underbrace{900}_{a}$$

The experimental protocol involves three sequential forward passes through the autoregressive model with L transformer layers. First, baseline hidden states $h_i^{(l)} | i \in [1, T], l \in [1, L]$ are recorded during normal operation when the model correctly predicts attribute a given prompt $x = (i, s, r)$. Second, we introduce Gaussian noise $\epsilon \sim \mathcal{N}(0, \nu)$ to subject token embeddings $h_{i*}^{(0)} = h_i^{(0)} + \epsilon$, inducing prediction corruption through propagated layerwise perturbations $h_{i*}^{(l)}$. Finally, selective restoration of original hidden states at specific (i, l) positions tests their capacity to recover correct predictions, establishing causal responsibility for factual recall.

We find that layers 0-5 in the OLMo models are responsible for storing factual associations. According to the causal mediation analysis graph shown in Figure 1, restoring the hidden states of layers 0-5, leads to correct attribute predictions. This indicates that these hidden states establish the information necessary for correct attribute predic-

Method/Team	Final Score ↑	Task Aggregate ↑	MIA Score ↑	MMLU Avg. ↑
Baselines				
Gradient Ascent (Chen and Yang, 2023)	0.394	0	0.912	0.269
Gradient Difference (Liu et al., 2022)	0.243	0	0.382	0.348
KL Minimization (Maini et al., 2024)	0.395	0	0.916	0.269
NPO (Zhang et al., 2024)	0.188	0.021	0.080	0.463
Leaderboard				
AILS-NTUA	0.706	0.827	0.847	0.443
ZJUKLAB	0.487	0.944	0.048	**0.471**
ch**3@stu.ynu.edu.cn	0.470	0.834	0.139	0.436
Ours	**0.711**	**0.964**	**0.894**	0.275

Table 1: Comparison of top 3 teams with our submission on 7B model, along with baselines of different unlearning methods on this dataset (Ramakrishna et al., 2025a).
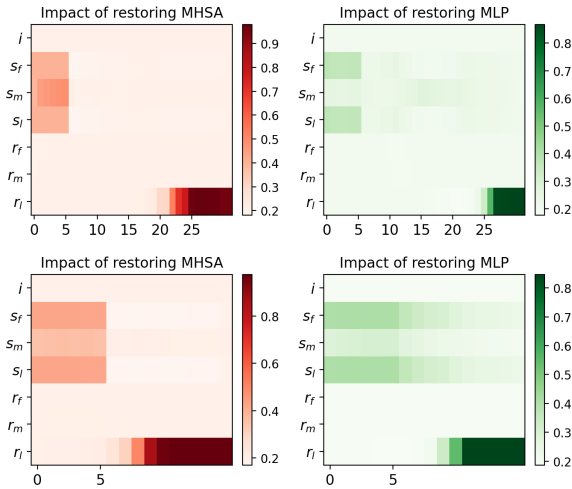

Figure 1: Impact of restoring hidden states at various token levels on predicting correct attribute. *Top*: OLMo 7B. *Bottom*: OLMo 1B. The *x-axis* shows number of layers and *y-axis* shows the impact of each type of token *(i,s,r)* in predicting attribute '*a*'. The '*s*' and '*r*' are broken into first '*f*', middle '*m*' and last '*l*' tokens. Tokens in same category is averaged.

tion early in the model's processing and are directly responsible for factual recall.

3.2 Loss Function

Given input tokens $[x_1, x_2, ..., x_m]$ and output tokens $[y_1, y_2, ..., y_n]$, we calculate the negative log likelihood as

$$\mathcal{L}_{CE} = -\log(P(y_1, ..., y_n | x_1, ..., x_m)).$$

Since we want to maximize the loss on forget set and minimize on retain set, we minimize a *joint loss function*

$$\mathcal{L}_{joint} = -\mathcal{L}_{CE}^{forget} + \alpha \cdot \mathcal{L}_{CE}^{retain}.$$

We select α to have a higher impact on the total loss when the loss on the retain-set deviates significantly

from its value at the initial epoch.

To determine α's value, we first calculate the mean of positive (retain) loss at the zeroth epoch. Subsequently, after each epoch, if the change in retain loss increases relative to this baseline, α is exponentially scaled (clipped between predefined minimum and maximum thresholds) to penalize deviations from retain-set performance. Specifically, for each epoch, we compute the following:

$$\Delta\mathcal{L} = \mathcal{L}_i^{retain} - \mathcal{L}_0^{retain}$$

where $i \in \{1, 2, ..., epochs\}$. The value of α is decided by the following

$$\gamma = a \cdot b^{\Delta L} + c$$

$$\gamma_{rnd} = \text{round}(\gamma, 1)$$

$$\alpha = \begin{cases} \min(\max(\gamma_{rnd}, \alpha_{min}), \alpha_{max}) & i \geq 1 \\ \alpha_{min} & i = 0 \end{cases}$$

For the exponential scaling function governing γ we used the parameter values a=0.3, b=6, and c=0.8. More details about the selection of the hyperparameters can be found in Appendix A.

4 Experiments

This section details the analysis of our main evaluation results in the leaderboard, as well as our parameter selection process.

4.1 Main Results

For the 7B variant (Table 1), we achieved the highest task aggregate scores (0.964) but encountered scalability challenges—MMLU decreased by 46% ($0.509 \rightarrow 0.275$) despite equivalent layer freezing depth (layers 0–5). The dataset used for calculating final unlearning results is approximately

Team	Score ↑	TA ↑	MIA ↑	MMLU ↑
AILS-NTUA	**0.688**	0.964	**0.857**	0.242
Atyaephyra	0.586	0.887	0.622	0.248
Mr. Snuffleupagus	0.485	0.412	0.793	**0.250**
Ours	0.652	**0.973**	0.741	0.243

Table 2: Comparison of top 3 teams with our submission on 1B model. Score and TA denote the Final Score and Task Aggregate, respectively.

twice the size of the publicly available dataset, and our algorithm's overfitting on this expanded corpus likely contributed to reduced model utility. This suggests larger models require fewer update steps. In contrast, our submission for the 1B model track achieved second place with a 0.652 final score (Table 2), delivering state-of-the-art task aggregate performance (0.973) through optimized MLP layer updates. The approach reduced forget-set knowledge retention to 0.14 (86% reduction) while maintaining 94% retain-set accuracy. MIA vulnerability scores of 0.741 demonstrate robust privacy protection. The 22% MMLU decrease ($0.27 \rightarrow 0.24$) remains within task utility thresholds, contrasting sharply with the catastrophic 46% drop observed in full-model approaches.

4.2 Selecting Unlearning Parameters

We identify the specific component responsible for recall by training different groups of layers. Our experiments include training all parameters of the model vs. the layers identified in Section 3.1 for the 7B model. We also separately train Multi-Head Self-Attention (MHSA) and MLP modules, summarizing our key findings below (see Table 3 and 4):

- Fine-tuning all layers severely reduces model utility and a catastrophic loss of recall for the retain set.

- By freezing the upper layers and training both Multi-Head Self-Attention (MHSA) and MLP of transformer blocks 0-5, we observe a good recall of retain set and effective unlearning on forget set. However, it does reduce the MMLU scores akin to training the full model.

- When we split the training to only fine-tune either the MLP or MHSA, we observe that MLP layers have a higher impact on unlearning compared to just MHSA.

From this analysis, we conclude that training

Layer	Type	Forget		Retain	
		Reg. ↓	Know. ↓	Reg. ↑	Know. ↑
0-32	MLP+MHSA	0	0	0.743	0.341
0-5	MLP+MHSA	**0.237**	**0.147**	0.896	0.946
0-5	MHSA	0.542	0.614	0.946	**0.958**
0-5	MLP	0.467	0.292	**0.952**	0.839

Table 3: After training for 8 epochs, training both MHSA and MLP produces the best result. However when comparing just MHSA scores with MLP, we observe the MLP has much more effect in knowledge recall. Reg. and Know. represent the Regurgitation Score and Knowledge Score, respectively.

Layer	Type	Score ↑	TA ↑	MIA ↑	MMLU ↑
0-32	MLP+MHSA	0.392	0.587	0.197	0.391
0-5	MLP+MHSA	**0.467**	**0.775**	**0.217**	0.410
0-5	MHSA	0.285	0.378	0.005	0.472
0-5	MLP	0.353	0.572	0.010	**0.477**

Table 4: Training different set of parameters for 8 epochs shows where that by training only MLP layers (0-5) in the OLMo model can effectively remove information without causing much loss in model utility, measured on MMLU benchmark. Training both MHSA and MLP achieves the highest score but reduces general model utility. Score and TA denote the Final Score and Task Aggregate, respectively.

only MLP layers is the most effective strategy. It has a task aggregate score of 0.57 on the public dataset, with an MMLU drop to 0.47 from 0.51 on the original 7B model.

5 Conclusion

Our systematic investigation establishes that targeted unlearning in large language models can be significantly enhanced through causal-informed layer optimization. Combining causal mediation analysis with constrained parameter updates to MLP modules in early transformer layers (0-5), we demonstrate precise eradication of sensitive subject-attribute associations while preserving 88% of baseline general knowledge performance in OLMo-1B models. The proposed joint loss formulation—simultaneously applying cross-entropy penalties on forget set outputs and adaptive regularization for retain set preservation—achieves a high task aggregate score of 0.973. While the 7B variant maintained comparable forget set removal efficacy (0.964 task score), its 46% MMLU degradation underscores the critical need for more robust mechanistic interventions. These findings suggest three key implications for machine unlearning research: First, that MLP layers in early trans-

former blocks serve as preferential targets for factual knowledge modification; second, that output token cross-entropy provides a more surgical intervention than full-sequence loss calculations; and third, that layer freezing thresholds must scale nonlinearly with model depth to maintain utility. Future work should investigate constrained gradient updates through KL divergence to reduce the drop in model utility after unlearning. Our results cement causal mediation as a vital tool for developing compliant, adaptable LLMs that meet evolving data privacy requirements without full retraining.

6 Acknowledgment

References

Stella Biderman, USVSN Sai Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. 2023. Emergent and predictable memorization in large language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*.

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 2019*.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. 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 2021*.

Jiaao Chen and Diyi Yang. 2023. Unlearn what you want to forget: Efficient unlearning for llms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*.

Min Chen, Zhikun Zhang, Tianhao Wang, Michael Backes, Mathias Humbert, and Yang Zhang. 2021. When machine unlearning jeopardizes privacy. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. 2024. Do membership inference attacks work on large language models? *CoRR*, abs/2402.07841.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*.

Dirk Groeneveld, Iz Beltagy, Evan Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. 2024. Olmo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021*.

Zhehao Huang, Xinwen Cheng, JingHao Zheng, Haoran Wang, Zhengbao He, Tao Li, and Xiaolin Huang. 2024. Unified gradient-based machine unlearning with remain geometry enhancement. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024*.

Qinbin Li, Junyuan Hong, Chulin Xie, Jeffrey Tan, Rachel Xin, Junyi Hou, Xavier Yin, Zhun Wang, Dan Hendrycks, Zhangyang Wang, Bo Li, Bingsheng He, and Dawn Song. 2024. LLM-PBE: assessing data privacy in large language models. *Proc. VLDB Endow.*, 17(11):3201–3214.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proc. ACL workshop on Text Summarization Branches Out*.

Bo Liu, Qiang Liu, and Peter Stone. 2022. Continual learning and private unlearning. In *Conference on Lifelong Learning Agents, CoLLAs 2022*.

Xiaoze Liu, Ting Sun, Tianyang Xu, Feijie Wu, Cunxiang Wang, Xiaoqian Wang, and Jing Gao. 2024. SHIELD: evaluation and defense strategies for copyright compliance in LLM text generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024*.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. 2024. TOFU: A task of fictitious unlearning for llms. *CoRR*, abs/2401.06121.

Daniel Mela, Aitor Gonzalez-Agirre, Javier Hernando, and Marta Villegas. 2024. Mass-editing memory with attention in transformers: A cross-lingual exploration of knowledge. In *Findings of the Association for Computational Linguistics, ACL 2024*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*.

Kevin Meng, Arnab Sen Sharma, Alex J. Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations, ICLR 2023*. OpenReview.net.

Vaidehi Patil, Peter Hase, and Mohit Bansal. 2024. Can sensitive information be deleted from llms? objectives for defending against extraction attacks. In *The Twelfth International Conference on Learning Representations, ICLR 2024*. OpenReview.net.

Anil Ramakrishna, Yixin Wan, Xiaomeng Jin, Kai-Wei Chang, Zhiqi Bu, Bhanukiran Vinzamuri, Volkan Cevher, Mingyi Hong, and Rahul Gupta. 2025a. Lume: Llm unlearning with multitask evaluations. *arXiv preprint arXiv:2502.15097*.

Anil Ramakrishna, Yixin Wan, Xiaomeng Jin, Kai-Wei Chang, Zhiqi Bu, Bhanukiran Vinzamuri, Volkan Cevher, Mingyi Hong, and Rahul Gupta. 2025b. Semeval-2025 task 4: Unlearning sensitive content from large language models. *arXiv preprint arXiv:2504.02883*.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy, SP 2017*. IEEE Computer Society.

Nexhi Sula, Abhinav Kumar, Jie Hou, Han Wang, and Reza Tourani. 2024. Silver linings in the shadows: Harnessing membership inference for machine unlearning. *CoRR*, abs/2407.00866.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Y Singer, and SM Shieber. 2004. Causal mediation analysis for interpreting neural nlp: the case of gender bias (2020). *CoRR arXiv*, abs/2004.12265.

Yu Wang, Ruihan Wu, Zexue He, Xiusi Chen, and Julian J. McAuley. 2024. Large scale knowledge washing. *CoRR*, abs/2405.16720.

Rongwu Xu, Brian S. Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2024. The earth is flat because...: Investigating llms' belief towards misinformation via persuasive conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024*.

Jin Yao, Eli Chien, Minxin Du, Xinyao Niu, Tianhao Wang, Zezhou Cheng, and Xiang Yue. 2024a. Machine unlearning of pre-trained large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024*.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2024b. Large language model unlearning. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024*.

Xiaojian Yuan, Tianyu Pang, Chao Du, Kejiang Chen, Weiming Zhang, and Min Lin. 2024. A closer look at machine unlearning for large language models. *CoRR*, abs/2410.08109.

Dawen Zhang, Pamela Finckenberg-Broman, Thong Hoang, Shidong Pan, Zhenchang Xing, Mark Staples, and Xiwei Xu. 2023. Right to be forgotten in the era of large language models: Implications, challenges, and solutions. *CoRR*, abs/2307.03941.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. 2024. Negative preference optimization: From catastrophic collapse to effective unlearning. *CoRR*, abs/2404.05868.

Zhenhong Zhou, Jiuyang Xiang, Chaomeng Chen, and Sen Su. 2024. Quantifying and analyzing entity-level memorization in large language models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014*.

A Parameter Selection for Adaptive Regularization

To balance the preservation of retain set knowledge with the effective removal of forget set information, we introduced an adaptive regularization weight, denoted as α, applied to the retain set loss ($\mathcal{L}^{retain}$). This weight dynamically adjusts based on the deviation of current retain set loss from its value at the start of the unlearning process.

We empirically set a=0.3, b=6, c=0.8, $\alpha_{min} = 1.2$, and $\alpha_{max} = 2.8$ (Figure 2). The selected b=6 controls exponential sensitivity to ΔL, strongly penalizing moderate retain loss increases to preserve performance. The offset c=0.8 and minimum clip α_{min} set a baseline regularization, while α_{max} prevents excessive strength. This configuration achieved a robust experimental balance by strongly penalizing retain set deviations while permitting effective unlearning.
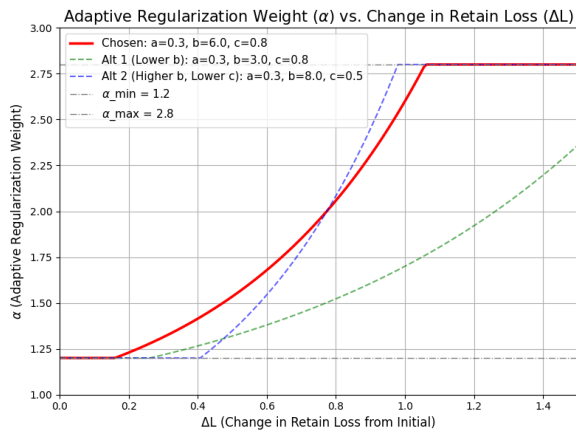

Figure 2: Visualization of the adaptive regularization function α, plotted against the change in retain loss ΔL. The chosen configuration (solid red) strongly penalizes increases in ΔL for the range of observed ΔL values, compared to blue or green.