ATTENTION-ONLY MACHINE UNLEARNING IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine unlearning aims to remove designated knowledge from a model while preserving its utility and minimizing computational cost. However, existing approaches often struggle with *knowledge entanglement*, where targeted and retained knowledge share model representations. In this work, we investigate attention-only unlearning for large language models (LLMs), focusing on a component that is central to the Transformer architecture yet often overlooked as a target for unlearning. Inspired by recent insights into memory through attention mechanisms, our method systematically identifies context-specific "anchored terms" and confines parameter updates to attention layers to enable targeted, granular unlearning. Experiments show that our approach achieves unlearning effectiveness competitive with existing techniques while preserving performance on retained data and reducing computational requirements. These efficiency gains stem from updating attention modules, which contain fewer parameters than the corresponding multilayer perceptron (MLP) modules in the evaluated architectures. Our findings highlight attention as a promising target for efficient unlearning and offer insights into its role in selectively accessing and suppressing knowledge in LLMs.

1 INTRODUCTION

Growing concerns over personal data privacy have prompted regulations granting individuals greater control over their information, including rights to request its deletion (European Union, 2016; State of California, 2018; Government of Japan, 2003; Government of Canada, 2022). However, removing data from a training corpus does not erase their influence on an already trained model. Retraining from scratch without the specified data provides a reference solution, but is often prohibitively expensive, particularly for large language models (LLMs) (Sevilla et al., 2022). Machine unlearning, initially proposed by Cao & Yang (2015), seeks to remove this influence efficiently while preserving model utility on retained data.

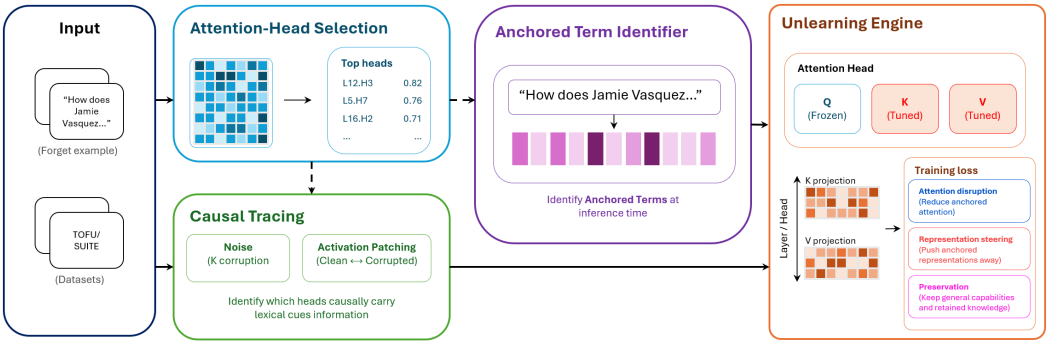


Figure 1: Overview of our approach. Dotted lines show auxiliary information flow, solid lines show training/main information flow.

A central challenge is *knowledge entanglement*: the representations and mechanisms supporting knowledge to be forgotten also support knowledge to be retained (Liu et al., 2024; Maini et al., 2024). We focus on a potential contributor to this challenge: unlearning interventions that do not distinguish target-specific cues from shared linguistic content within forget inputs. A question or

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

passage may contain only a few tokens that identify or activate the target knowledge, alongside common words and structures essential to general language capabilities. Updates that fail to distinguish these roles risk disrupting shared mechanisms and degrading performance beyond the intended scope of forgetting. This motivates selective, interpretability-guided interventions (Yu et al., 2023; Wu et al., 2023).

We investigate attention as a comparatively underexplored target for such interventions. Attention combines information across token positions and contributes to the representations processed by subsequent MLP layers (Vaswani et al., 2017; Elhage et al., 2021). Beyond this architectural role, evidence that attention can function as a key–value memory (Gershman et al., 2025) motivates examining its role in knowledge retrieval. Together with findings that MLP components promote concepts in the output vocabulary (Geva et al., 2022), these observations suggest complementary roles for attention and MLPs in expressing learned knowledge. Because attention shapes the representations processed by subsequent MLPs and typically contains fewer parameters, it offers a promising target for influencing downstream knowledge expression through parameter-efficient unlearning.

Building on the role of "anchored terms" discussed by Eldan & Russinovich (2023), we investigate *lexical-cue-guided unlearning in attention layers*. We hypothesize that these terms serve as lexical cues for attention-mediated knowledge retrieval. Guided by the identified cues, we update the key (K) and value (V) projection weights, which determine how token representations are matched to queries and what information is propagated through attention. We keep the query (Q) projection weights fixed to avoid directly modifying the mapping that determines what information each token position seeks. This design aims to mitigate knowledge entanglement by disrupting cue-associated retrieval while limiting interference with retained knowledge (Figure 1 shows an overview of our approach). Throughout this study, we use "anchored terms", "subjects", and "lexical cues" interchangeably; "subjects" refers to entities, not necessarily grammatical subjects. Our main contributions are:

- We introduce an **Anchored Terms Identifier (ATI)** to enable fine-grained, potentially more efficient unlearning of cue-associated knowledge, excluding shared linguistic content from direct targeting to mitigate knowledge entanglement.
- We propose an **Attention Unlearning Engine** with objectives tailored to attention's matching and aggregation mechanisms, guiding K/V projection updates through attention disruption, representation steering, and retention constraints.
- Experiments on **TOFU** and **SUITE** datasets demonstrate comparable or better unlearning performance using 4–$32\times$ fewer optimizer steps than the compared baselines. Highlighting attention projections and anchored terms as promising targets for selective knowledge modification.

2 Related works

Machine unlearning methods primarily rely on gradient ascent (GA) to increase the loss on forget data, reducing the likelihood of generating targeted content (Maini et al., 2024; Yu et al., 2023; Yao et al., 2024a; Jang et al., 2023). Forgetting objectives are often combined with retention losses or divergence-based regularization, such as KL divergence, to limit changes in model behavior on retained data. However, unconstrained GA can destabilize training and substantially degrade model utility (Zhang et al., 2024; Zhao et al., 2024b). Reinforcement-learning and preference-based approaches offer alternative objectives for controlling undesirable generations and improving the forgetting–retention trade-off (Lu et al., 2022; Zhang et al., 2024; Yao et al., 2024b; Fan et al., 2025).

Other approaches explicitly train models to produce preferred responses, such as "I don't know," when queried about targeted information (Gao et al., 2025; Mekala et al., 2025). However, suppressing the expression of knowledge does not necessarily eliminate its underlying influence. Adversarial extraction and relearning evaluations reveal that information targeted by some unlearning methods can remain recoverable (Chen et al., 2021b; Patil et al., 2024; Hu et al., 2025). Together, these findings highlight the difficulty of achieving effective forgetting without compromising retained capabilities.

These challenges motivate interventions informed by model internals, connecting machine unlearning with explainable artificial intelligence (XAI) (Zhao et al., 2024a). Representation engineering,

for example, provides a framework for identifying and manipulating internal representations (Zou et al., 2025), and related unlearning methods redirect representations associated with target knowledge (Li et al., 2024; Shen et al., 2025). Our work complements these approaches by investigating the granularity and location of unlearning interventions. Specifically, we distinguish anchored terms from shared linguistic content and use these cues to guide updates to attention's key and value projection weights. This examines a comparatively underexplored direction: whether cue-guided interventions in attention can mitigate knowledge entanglement and improve the forgetting–retention trade-off.

3 METHODOLOGY

3.1 ANCHORED TERMS IDENTIFIER - ATI

We implement ATI as a lightweight token-level classifier trained on cached attention-head features from a pretrained LLM. ATI separates representation extraction from supervised learning: head features are extracted once and cached, after which a compact MLP identifies candidate anchored-term tokens.

Attention-Head Selection. Motivated by evidence of functional specialization among attention heads (Yao et al., 2024c), we investigate entity-sensitive heads as a feature source for anchored-term identification. Entity mentions can serve as lexical cues for target knowledge (Eldan & Russinovich, 2023); we therefore use entity localization as a proxy, without assuming that every mention triggers retrieval. Here, *subject* denotes an entity of interest, not necessarily a grammatical subject.

Our characterization builds on the additive decomposition of attention-head outputs used in direct logit attribution (DLA) (Elhage et al., 2021; Wang et al., 2023). Unlike DLA, we measure entity-associated aggregation before the output projection rather than attributing contributions to a selected output logit. This avoids requiring target answer tokens and accommodates varying input formats.

For a layer–head pair $j = (\ell, h)$, let S contain entity-token positions obtained through lexical matching or annotations, and let r be the final non-padding position. We compute

$$\rho_j = \frac{\left\| \sum_{t \in S} A_{r,t}^{(j)} \mathbf{v}_t^{(j)} \right\|_2}{\left\| \sum_{t=0}^{T-1} A_{r,t}^{(j)} \mathbf{v}_t^{(j)} \right\|_2 + \epsilon}$$

where $A_{r,t}^{(j)}$ is the model-returned attention weight, $\mathbf{v}_t^{(j)}$ is the corresponding value vector, and $\epsilon > 0$ ensures numerical stability. We select a set of $\mathcal{H}$ heads with the highest mean ρ_j over the head-analysis dataset, assigning zero to examples without detected entities. This statistic is descriptive rather than causal and can exceed one when entity and non-entity contributions partially cancel. (See Appendix A.1 for details)

Token-Level Feature Construction. ATI requires a separate score for each candidate token, whereas the attention output at the final query position aggregates information across source tokens. We therefore extract four scalar features that mirror attention computation: raw query–key compatibility, normalized compatibility, value magnitude, and normalized compatibility-weighted value magnitude. For each selected head $j \in \mathcal{H}$,

$$s_t^{(j)} = \frac{(\mathbf{q}_r^{(j)})^\top \mathbf{k}_t^{(j)}}{\sqrt{d_h}}, \quad \widetilde{a}_t^{(j)} = \frac{\exp(s_t^{(j)})}{\sum_{i:m_i=1} \exp(s_i^{(j)})},$$

$$u_t^{(j)} = \|\mathbf{v}_t^{(j)}\|_2, \qquad c_t^{(j)} = \frac{\widetilde{a}_t^{(j)} u_t^{(j)}}{\max\left(\sum_{i:m_i=1} \widetilde{a}_i^{(j)} u_i^{(j)}, \epsilon\right)}.$$

Here, r is the final non-padding position, and $m_i = 1$ indicates a valid token. Concatenating features across the H selected heads yields the scorer input

$$\mathbf{x}_t = \bigoplus_{j \in \mathcal{H}} \left[\widetilde{a}_t^{(j)},\ s_t^{(j)},\ u_t^{(j)},\ c_t^{(j)} \right] \in \mathbb{R}^{4H}$$

Table 1: ATI performance on Mintaka. @k denotes the output cutoff. P and R denote token-level precision and recall. Bold values indicate the best score across head budgets within each model and metric.

Heads $\mathcal{H}$	Llama-3.1-8B				Phi-4				Qwen2.5-14B			
	Precision ↑		Recall ↑		Precision ↑		Recall ↑		Precision ↑		Recall ↑	
	P@3	P@5	R@3	R@5	P@3	P@5	R@3	R@5	P@3	P@5	R@3	R@5
3	0.811	0.654	0.723	0.893	0.792	0.640	0.709	0.877	0.793	0.649	0.681	0.848
5	0.832	0.667	0.745	0.907	0.810	0.653	0.726	0.893	0.809	0.663	0.695	0.863
10	**0.845**	**0.678**	**0.757**	**0.919**	**0.837**	**0.676**	**0.750**	**0.917**	**0.838**	**0.686**	**0.723**	**0.888**

These features approximate rather than reconstruct attention: compatibility precedes any subsequent positional transformations or query/key normalization, and value magnitudes omit directional interactions during aggregation.

Supervised ATI & Training. ATI maps $\mathbf{x}_t$ to a sigmoid score $p_t = \sigma(f_\theta(\mathbf{x}_t))$ using a two-hidden-layer MLP with widths D and $\lfloor D/2 \rfloor$. Each hidden layer uses layer normalization, GELU, and dropout.

We train on Mintaka (entity-annotated dataset) questions (Sen et al., 2022), assigning $y_{it} = 1$ when token t in question i overlaps an annotated entity span, and 0 otherwise. We minimize masked, positive-class-weighted binary cross-entropy:

$$\mathcal{L}_{\text{ATI}} = -\frac{1}{\sum_{i,t} m_{it}} \sum_{i,t} m_{it} \left[w_+ y_{it} \log p_{it} + (1 - y_{it}) \log(1 - p_{it}) \right]$$

where $m_{it} = 1$ for non-padding tokens and $w_+ = \max(1, N_-/\max(N_+, 1))$, with N_+ and N_- counting positive and negative non-padding training tokens. Only θ is optimized; the pretrained LLM remains frozen.

At inference, we select candidate anchored-term positions using

$$\widehat{S}(x) = \{t : m_t = 1,\ p_t \geq \tau\}, \qquad \tau = 0.75. \tag{1}$$

Entity localization serves as a proxy for cue identification; downstream unlearning experiments assess the selected tokens' utility for selective forgetting.

ATI Evaluation and Causal Probe. We evaluate ATI on the Mintaka validation set across three LLMs with head budgets $\mathcal{H} \in \{3, 5, 10\}$. Table 1 reports token-level precision and recall at $k \in \{3, 5\}$. Increasing $\mathcal{H}$ consistently improves these metrics, while selecting five rather than three tokens increases recall at the expense of precision. For downstream unlearning on TOFU and SUITE, ATI instead uses threshold-based selection (Equation 1), without a fixed token budget. For example, given the TOFU question "`What is the profession of` **`Hsiao Yun-Hwa's`** `father?`", ATI selects the bold span as the candidate anchored term. (See Appendix A.2 for further qualitative results)

To examine whether ATI scores identify tokens influential for factual recall, we conduct a noise-ablation probe on CounterFact (Meng et al., 2022). We corrupt representations at three positions selected by each strategy: ATI's highest-scoring subject tokens, random subject tokens, or random non-subject tokens. We measure Contrastive Impact:

$$\text{CI} = (\ell_f - \ell_{cf})_{\text{clean}} - (\ell_f - \ell_{cf})_{\text{corrupted}}$$

where ℓ_f and ℓ_{cf} are the log-probabilities of the factual and counterfactual targets, respectively. Higher CI indicates a larger reduction in factual preference.

Across three models and noise factors $\{1, 2, 3, 5, 8\}$, subject-token corruption produces larger CI than non-subject corruption, and ATI-selected subject tokens yield larger CI than randomly selected subject tokens (Figure 2). These results support ATI's ability to identify influential tokens within entity spans under the tested interventions.

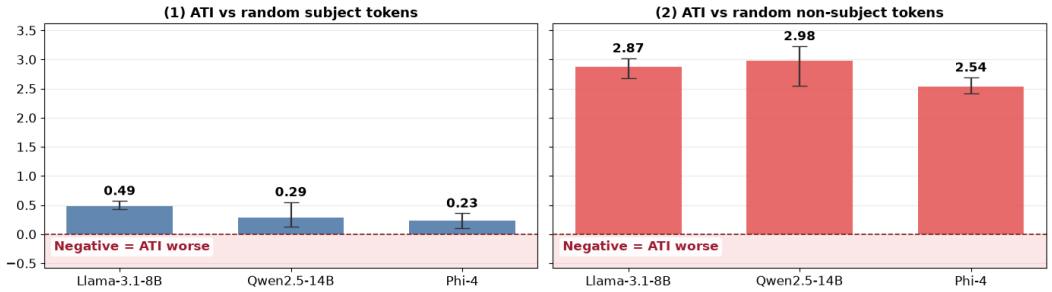


Figure 2: Causal probe: (1) ATI versus random subject tokens; (2) ATI versus random non-subject tokens

3.2 Attention Unlearning Engine

ATI's candidate anchored terms are utilized to guide updates to selected key and value projection weights. Query projections and all other model parameters remain frozen. The engine was built on the idea of attention disruption and representation steering with retention objectives to limit interference with retain knowledge.

Anchored Target. Our engine selects candidate anchored-term positions using ATI with a threshold of $\tau = 0.75$ (Equation 1). Each selected position receives a soft weight $w_t = p_t$, and both positions and weights are held fixed during the corresponding unlearning step. Rather than treating all tokens within an anchored term uniformly, this weighting emphasizes those identified more confidently by ATI.

Unlearn objective. Let $\mathcal{D}_f$ and $\mathcal{D}_r$ denote the forget and retain datasets, respectively. We optimize a combination of two unlearn losses and two preservation losses:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{qk}}\mathcal{L}_{\text{qk}} + \lambda_{\text{steer}}\mathcal{L}_{\text{steer}} + \lambda_{\text{kv}}\mathcal{L}_{\text{kv-preserve}} + \lambda_{\text{retain}}\mathcal{L}_{\text{retain}} \tag{2}$$

The QK-disruption ($\mathcal{L}_{\text{qk}}$) and representation-steering ($\mathcal{L}_{\text{steer}}$) losses target cue-associated attention and representations on forget inputs. The projection-preservation ($\mathcal{L}_{\text{kv-preserve}}$) and retain data ($\mathcal{L}_{\text{retain}}$) losses constrain changes to shared attention mechanisms and performance on retain data. The nonnegative coefficients λ_{qk}, λ_{steer}, λ_{kv}, and λ_{retain} control their relative strengths.

***QK* disruption.** Motivated by perspectives on attention-based memory retrieval (Dinh et al., 2026; Gershman et al., 2025), we treat attention to anchored-term tokens as a potential pathway for retrieving associated information. Their key representations act as addressing cues: compatibility with contextual queries determines the weights assigned to their value representations in attention outputs. We therefore penalize attention to these tokens to reduce access to the associated information through the selected heads.

Let $\mathcal{J}$ denote the targeted attention heads and R_x the valid query positions outside $\widehat{S}(x)$. For head j, let $a_{r,t}^{(j)}$ denote the query–key score after softmax normalization over all key positions permitted by the attention mask. We define

$$\mathcal{L}_{\text{qk}} = \mathbb{E}_{x\sim\mathcal{D}_f}\left[\frac{1}{|\mathcal{J}|}\sum_{j\in\mathcal{J}}\frac{1}{|R_x|}\sum_{r\in R_x}\sum_{t\in S(x)} w_t a_{r,t}^{(j)}\right] \tag{3}$$

Excluding subject-origin queries focuses the objective on information transfer from subject positions to other positions, rather than within-subject interactions.

Representation steering. We construct a perturbed input $\tilde{x}$ by replacing anchored tokens with reference tokens from 20 fictitious author names from a held-out TOFU subset (See Appendix A.3 for details). Functionally, these fictitious tokens provide a empirical realization of an uninformative, low-association entity baseline within the model's representation space. By steering the key

and value projections of target anchored terms toward these uninformative reference vectors, $\mathcal{L}_{\text{steer}}$ disrupts the specific query–key addressing alignment required for attention-mediated retrieval (Gershman et al., 2025; Dinh et al., 2026), forcing the model to treat the anchored term as a neutral entity.

Define the w_t-weighted discrepancy

$$d_{S,w}(U, \widetilde{U}) = \frac{\sum_{t \in S(x)} w_t \|U_t - \widetilde{U}_t\|_2^2}{d_h \sum_{t \in S(x)} w_t + \epsilon},$$

where d_h is the head dimension and $\epsilon > 0$ ensures numerical stability. The steering objective is

$$\begin{aligned} \mathcal{L}_{\text{steer}} = \mathbb{E}_{x \sim \mathcal{D}_f} \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} \Big[& d_{S,w}\Big(K^{(j)}(x), \text{sg}(K^{(j)}(\widetilde{x}))\Big) \\ & + \eta_V d_{S,w}\Big(V^{(j)}(x), \text{sg}(V^{(j)}(\widetilde{x}))\Big) \Big] \end{aligned} \tag{4}$$

where sg denotes stop-gradient and $\eta_V \in \{0, 1\}$ controls value steering. This objective encourages representations at subject positions to resemble those elicited by the unknown anchors.

Retained-projection preservation. To constrain changes to shared attention mechanisms, we regularize projection outputs on retained inputs. Let W_P^0 and W_P denote the original and current projection weights for $P \in \{K, V\}$. For the same hidden state h_t, define

$$\Delta_P(h_t) = \text{MSE}\big(W_P h_t, W_P^0 h_t\big)$$

Using the same hidden state compares the projection mappings directly, rather than comparing two complete model forward passes. We minimize

$$\mathcal{L}_{\text{kv}-\text{preserve}} = \mathbb{E}_{x \sim \mathcal{D}_r} \, \text{mean}_{\ell,\, t \in U_r(x)} \left[\Delta_K(h_t^\ell) + \Delta_V(h_t^\ell) \right] \tag{5}$$

where $\mathcal{D}_r$ denotes the retain dataset, ℓ ranges over the targeted layers, and $U_r(x)$ contains valid retained-input positions excluding identified anchored spans. This objective constrains changes on sampled retained representations while leaving other directions less constrained. Our main configurations preserve both K and V projections.

Unlearning layers. To select attention layers for unlearning, we apply activation patching on CounterFact, following Meng et al. (2022), but restrict interventions to the key (K) component. CounterFact supports entity-replacement corruption of the input, enabling us to identify attention heads that causally contribute to predictions involving anchored terms. We average head-level causal scores within each layer and select the top n layers for modification. Accordingly, $\mathcal{J}$ in equation 3 and equation 4 comprises all attention heads in the selected layers. (See Appendix A.4 for details)

3.3 Experiment setup

Datasets. We use two complementary benchmarks with different degrees of forget–retain overlap. TOFU (Maini et al., 2024) contains synthetic question–answer pairs about fictitious authors, providing a controlled setting with distinct forget and retain identities. SUITE (Peleg et al., 2026) targets four real-world topics or events, with paraphrased forget queries and syntactically, lexically, and semantically related retain inputs. It therefore tests whether forgetting generalizes across reformulations while preserving closely related information.

Baselines and Models. We compare against NPO (Zhang et al., 2024), SimNPO (Fan et al., 2025), SatImp (Yang et al., 2025), UNDIAL (Dong et al., 2025), and RMU (Li et al., 2024), using OpenUnlearning implementations (Dorna et al., 2025). We evaluate *Llama-3.1-8B-Instruct*, *Qwen2.5-14B-Instruct*, and *Phi-4*. For TOFU, each model is first fine-tuned on the full dataset using OpenUnlearning's default settings to introduce the synthetic author knowledge. (See Appendix A.5 for detailed configurations)

Table 2: Forgetting and retention performance on SUITE. Higher is better for all metrics. F: forgetting; D: retention; G: general performance. **Bold** and underline indicate the best and second-best results within each model, respectively. Ours uses $\lambda_{kv} = 6$ for Llama and $\lambda_{kv} = 5$ for Qwen and phi-4.

	Llama-3.1-8B-Instruct					Qwen2.5-14B-Instruct					phi-4				
Method	$F\uparrow$	$D\uparrow$	$G\uparrow$	$R\uparrow$	Overall $\uparrow$	$F\uparrow$	$D\uparrow$	$G\uparrow$	$R\uparrow$	Overall $\uparrow$	$F\uparrow$	$D\uparrow$	$G\uparrow$	$R\uparrow$	Overall $\uparrow$
NPO	0.7101	0.1995	0.4650	0.2792	0.4008	0.6859	0.2676	**0.6113**	0.3722	0.4826	0.7053	0.3169	0.4589	0.3749	0.4895
SimNPO	0.6008	0.3099	0.4668	0.3725	0.4599	0.5803	0.2428	0.5885	0.3438	0.4318	0.5695	0.3164	0.5852	0.4107	0.4772
SatImp	0.5496	0.1748	0.4682	0.2545	0.3479	0.4226	0.2182	0.5911	0.3187	0.3634	0.4362	0.2376	0.5793	0.3370	0.3802
RMU	0.7120	0.1093	**0.5137**	0.1802	0.2876	**0.7343**	0.1159	0.6080	0.1946	0.3077	0.6979	0.3384	0.5437	0.4171	0.5222
UNDIAL	**0.7157**	0.0389	0.1395	0.0609	0.1122	0.7171	0.0614	0.2029	0.0942	0.1666	**0.7135**	0.0475	0.2251	0.0785	0.1414
Ours	0.6797	**0.4246**	0.3770	**0.3884**	**0.5053**	0.5971	**0.4311**	0.5719	**0.4916**	**0.5392**	0.5833	**0.4766**	0.4913	**0.4838**	**0.5289**

Evaluation Metrics. We report ROUGE-L (Lin, 2004) for both datasets, along with ground-truth answer probability (Prob) and model utility for TOFU. We additionally assess memorization and extractability using Exact Memorization (EM) (Tirumala et al., 2022) and Extraction Strength (ES) (Carlini et al., 2020). To distinguish forgetting from output degeneration, we measure forget-set non-gibberish rates using a pretrained detector (Jindal, 2021); higher values indicate cleaner text.

We assess general capability using seven *Common Benchmarks*: MMLU (Hendrycks et al., 2021), MMLU-Pro (Wang et al., 2024), IFEval (Zhou et al., 2023), ARC Challenge (Clark et al., 2018), GPQA (zero-shot) (Rein et al., 2024), HumanEval (Chen et al., 2021a), and GSM-8K (CoT) (Cobbe et al., 2021).

3.4 Results

3.4.1 Main unlearning results

We summarize performance along three dimensions: forgetting, retention, and general capability. The forgetting score F is the harmonic mean of ROUGE, Prob, EM, and ES on the forget set, with each metric oriented so that higher values indicate more effective forgetting. The retention score D is the harmonic mean of retain-set ROUGE, Prob, and Utility, together with the non-gibberish rate. General capability G is measured using the *Common Benchmarks*.

Let H denote the harmonic mean. We combine retention and general capability into R, then balance R against forgetting:

$$R = H(D, G), \qquad \text{Overall} = H(F, R) = \left(\frac{0.5}{F} + \frac{0.25}{D} + \frac{0.25}{G}\right)^{-1}$$

This aggregation assigns equal weight to forgetting and combined retention, while penalizing imbalanced performance across dimensions.Tables 2 and 3 report unlearning results on SUITE and TOFU, respectively. We evaluate our method with $\lambda_{\text{qk}} = 0.1$, $\lambda_{\text{steer}} = 0.5$, $\lambda_{\text{kv}} \in \{5.0, 6.0\}$, and $\lambda_{\text{retain}} = 0.1$ (Equation 2). (For detailed unlearning results, see Appendix A.6)

SUITE. Results show our method with $\lambda_{\text{kv}} = 5.0$ achieves the highest overall score among the evaluated methods for all models. However, for *Llama-3.1-8B-Instruct* we find $\lambda_{\text{kv}} = 6.0$ to have better results compared to $\lambda_{\text{kv}} = 5.0$.

Because $L_{kv\text{-preserve}}$ excludes target anchored spans from MSE regularization (Equation 5), a key question is whether modifying K projections for an anchored term corrupts cross-attention when that same term appears in benign retained queries. SUITE's Lexical and Tier-0 evaluation subsets explicitly test this edge case by querying targeted entity tokens in non-forget contexts. Our method achieves best retention scores across models, empirically confirming that updating key projections for anchored terms does not cause cross-context attention corruption when those terms appear in benign prompts.

Table 3: Forgetting and retention performance on TOFU. Higher is better for all metrics. F: forgetting; D: retention; E: general performance. **Bold** and underline indicate the best and second-best results within each model, respectively. Ours uses $\lambda_{kv} = 6$ for all models.

	Llama-3.1-8B-Instruct					Qwen2.5-14B-Instruct					phi-4				
Method	$F\uparrow$	$D\uparrow$	$E\uparrow$	$R\uparrow$	Overall ↑	$F\uparrow$	$D\uparrow$	$E\uparrow$	$R\uparrow$	Overall ↑	$F\uparrow$	$D\uparrow$	$E\uparrow$	$R\uparrow$	Overall ↑
NPO	0.7070	0.6383	0.4857	0.5517	**0.6198**	0.7077	0.6291	**0.6290**	0.6291	0.6661	0.7117	0.6167	**0.5252**	0.5673	0.6314
SimNPO	0.4424	**0.7027**	0.5194	**0.5973**	0.5083	0.5092	**0.6933**	0.6129	0.6506	0.5713	0.4513	**0.7116**	0.5164	**0.5985**	0.5146
SatImp	0.3786	0.6842	0.5181	0.5897	0.4612	0.4879	0.6887	0.6180	**0.6515**	0.5579	0.4035	0.7084	0.5178	0.5983	0.4820
RMU	**0.9528**	0.0895	**0.5244**	0.1529	0.2635	0.6826	0.6784	0.6207	0.6483	0.6650	**0.9400**	0.1527	0.5028	0.2343	0.3751
UNDIAL	0.7583	0.5231	0.1695	0.2560	0.3828	**0.7994**	0.5696	0.4155	0.4805	0.6002	0.7598	0.6052	0.4340	0.5055	0.6071
Ours	0.9291	0.3962	0.4617	0.4265	0.5846	0.7652	0.6396	0.6065	0.6226	**0.6866**	0.7740	0.6303	0.4758	0.5422	**0.6377**

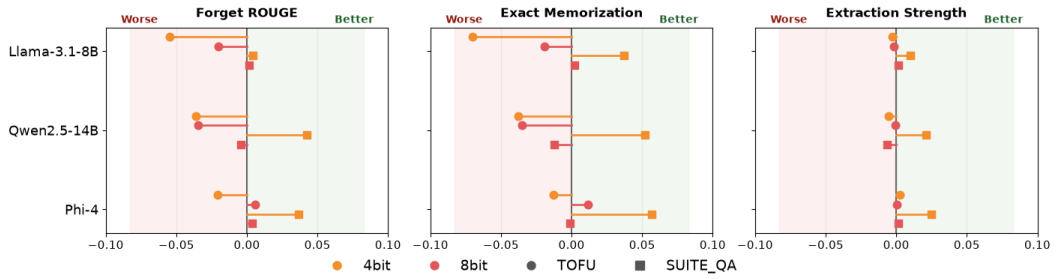


Figure 3: Quantization attack evaluation. Differences in forget-set metrics between INT8/INT4 models and their FP16 unlearned counterparts.

TOFU. Increasing λ_{kv} to 6.0 improves performance on TOFU, yielding the highest overall score for two of the three models. The exception is *Llama-3.1-8B-Instruct*, for which our method trails the best method, NPO, by 0.03. Unlike on SUITE, our method does not consistently lead in retention. Nevertheless, our method achieves the highest overall score for two models without leading any individual metric, reflecting a favorable balance between forgetting and retention.

Quantization Attacks. We conduct an exploratory stress test of sensitivity to weight precision by applying post-training INT8 and INT4 quantization to our unlearned FP16 checkpoints, following Zhang et al. (2025). We use the unlearning hyperparameters from Section 3.4.1, with $\lambda_{\text{kv}} = 6.0$. Figure 3 reports changes in forget-set metrics relative to the FP16 checkpoints. On TOFU, INT4 quantization partially reverses forgetting, increasing EM by up to 7 percentage points. On SUITE, measured forgetting remains stable or improves under both precision settings, indicating that key/value attention updates do not rely solely on ultra-high precision weight cancellation.

3.4.2 Anchored terms selection ablation.

While Section 3.1 evaluates the influence of ATI-selected tokens on factual predictions, here we test whether this selection improves attention-based unlearning. We compare ATI with three alternative selectors, keeping the unlearning objectives and hyperparameters from Section 3.4.1 fixed, with $\lambda_{\text{kv}} = 6.0$.

Frequency-based identifies candidate anchored terms through frequency-based voting over capitalized n-grams. *Random* uniformly samples the same number of valid token positions as ATI selects for each input. *All* targets every valid, unmasked token, providing a whole-input baseline without lexical-cue selection. Table 4 reports results with $\lambda_{\text{kv}} = 6.0$. ATI achieves the highest overall score in five of the six model–dataset combinations, with Qwen2.5 on SUITE being the exception.

Table 4: Anchored selection ablation on the SUITE and TOFU datasets.

Model	Selection	SUITE F-Rouge ↓	SUITE EM ↓	SUITE R-Rouge ↑	SUITE F-Gib. ↑	SUITE Overall ↑	TOFU F-Rouge ↓	TOFU EM ↓	TOFU Utility ↑	TOFU F-Gib. ↑	TOFU Overall ↑
Llama-3.1-8B	ATI	0.0079	0.1531	0.4246	0.7692	**0.6845**	0.0667	0.1603	0.5519	0.3090	**0.5472**
	Frequency-based	0.0723	0.3931	0.1686	0.6564	0.3930	0.3043	0.6845	0.5341	0.9111	0.5279
	Random	0.0098	0.1136	0.3192	0.5352	0.5603	0.0407	0.1245	0.5391	0.2029	0.4460
	All	0.0079	0.0771	0.2513	0.4026	0.4676	0.0695	0.1441	0.5553	0.0893	0.2625
Qwen2.5-14B	ATI	0.1825	0.4454	0.3360	0.5215	0.5050	0.2230	0.4269	0.4937	0.9078	**0.6495**
	Frequency-based	0.2380	0.6564	0.2558	0.5308	0.3994	0.3340	0.6982	0.4904	0.8890	0.5013
	Random	0.0493	0.2241	0.3561	0.5950	**0.5857**	0.0161	0.1249	0.5109	0.3242	0.5555
	All	0.0209	0.1853	0.3345	0.5064	0.5546	0.1218	0.1502	0.5039	0.2795	0.5077
phi-4	ATI	0.0255	0.2101	0.2523	0.5927	**0.5037**	0.2129	0.4285	0.5944	0.6709	**0.6459**
	Frequency-based	0.0788	0.4022	0.1892	0.4047	0.3804	0.3708	0.7635	0.6057	0.8840	0.4651
	Random	0.0045	0.0734	0.1862	0.2024	0.3227	0.0407	0.0895	0.5916	0.1881	0.4374
	All	0.0051	0.0744	0.2133	0.2563	0.3740	0.0482	0.1932	0.5813	0.3703	0.5960

While Figure 2 establishes that ATI localizes tokens with high causal retrieval weight during clean generation, Table 4 confirms the functional necessity of this selection during unlearning. Updating K/V projections specifically at ATI-selected tokens achieves superior overall performance compared to whole-input ('All') or random token updates, demonstrating that targeting causally grounded retrieval anchors mitigates collateral representation damage.

3.5 Computational efficiency

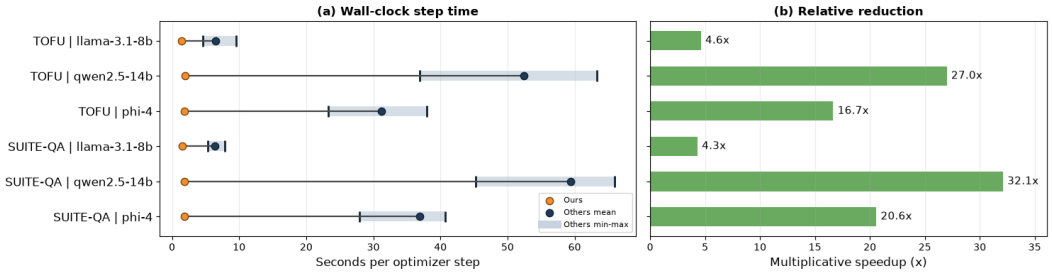


Figure 4: Empirical optimizer step-time comparison. Figure (a) shows wall-clock optimizer step time. Figure (b) shows relative reduction of our method versus other mean of other baseline methods.

We update K/V projections in only five attention layers, covering approximately 0.52%, 0.36%, and 1.34% of the parameters in *Llama-3.1-8B-Instruct*, *Qwen2.5-14B-Instruct*, and *Phi-4*, respectively. With an effective batch size of 32 and the hyperparameters in Section 3.4.1, average wall-clock time per optimizer step yields speedups over full-model baselines of at least $4.3\times$ on Llama and up to $32\times$ on Qwen (Figure 4).

4 Conclusion

In this work, we introduced an attention-only unlearning framework that combines lexical-cue identification with targeted K/V updates to mitigate knowledge entanglement. Across TOFU and SUITE on three LLMs, our method achieves competitive forgetting–retention trade-offs while updating less than 1.34% of parameters, with 4.3–$32.1\times$ wall-clock speedups per optimizer step over full-model baselines. These findings highlight attention mechanisms as a promising, highly efficient locus for surgical knowledge editing in transformer architectures.

Limitations & Future Work. While our key-value projection preservation loss ($\mathcal{L}_{kv\text{-preserve}}$) effectively mitigates cross-context attention corruption, our method relies on reference representations generated from fictitious author anchors, which may not optimal for non-person semantic domains. Future work includes developing domain-adaptive null target representations and extending attention-level interventions to dynamic multi-turn interactions.

AI USE STATEMENT

In this work, we used generative AI tools for help develop theoretical models or conceptual frameworks, formulate mathematical claims, assist in the writing of proofs, design or provide feedback on research methodology or experiments, implement methods, interpret results. We have not used generative AI tools for Generate synthetic data sets, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, clean and reformat dataset, support qualitative and thematic data analysis and assist with translation are not applicable to this work. Additionally, we used generative AI tools for create or edit software code, discover research topics or identify gaps, sourcing/searching for information. We have reviewed all AI-assisted work. We thoroughly check generated code prior to any actual change, writing assistant only improve written text, not used by itself to form paragraphs. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We include detailed description of our method as well as configurations and hardware usage for reproducibility.

REFERENCES

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, pp. 463–480, 2015. doi: 10.1109/SP.2015.35.

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. Extracting training data from large language models. *CoRR*, abs/2012.07805, 2020. URL https://arxiv.org/abs/2012.07805.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021a. URL https://arxiv.org/abs/2107.03374.

Min Chen, Zhikun Zhang, Tianhao Wang, Michael Backes, Mathias Humbert, and Yang Zhang. When machine unlearning jeopardizes privacy. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, CCS '21, pp. 896–911, New York, NY, USA, 2021b. Association for Computing Machinery. ISBN 9781450384544. doi: 10.1145/3460120.3484756. URL https://doi.org/10.1145/3460120.3484756.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Viet Hung Dinh, Ming Ding, Youyang Qu, and Kanchana Thilakarathna. Memory retrieval in transformers: Insights from the encoding specificity principle, 2026. URL https://arxiv.org/abs/2601.20282.

Yijiang River Dong, Hongzhou Lin, Mikhail Belkin, Ramon Huerta, and Ivan Vulić. UNDIAL: Self-distillation with adjusted logits for robust unlearning in large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8827–8840, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.444. URL https://aclanthology.org/2025.naacl-long.444/.

Vineeth Dorna, Anmol Reddy Mekala, Wenlong Zhao, Andrew McCallum, J Zico Kolter, Zachary Chase Lipton, and Pratyush Maini. Openunlearning: Accelerating LLM unlearning via unified benchmarking of methods and metrics. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=Gy67Zh5X1i.

Ronen Eldan and Mark Russinovich. Who's harry potter? approximate unlearning in llms, 2023. URL https://arxiv.org/abs/2310.02238.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

European Union. Regulation (eu) 2016/679 (general data protection regulation). Official Journal of the European Union, 2016.

Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Simplicity prevails: Rethinking negative preference optimization for LLM unlearning, 2025. URL https://openreview.net/forum?id=Pd3jVGTacT.

Chongyang Gao, Lixu Wang, Kaize Ding, Chenkai Weng, Xiao Wang, and Qi Zhu. On large language model continual unlearning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Essg9kb4yx.

Samuel J. Gershman, Ila Fiete, and Kazuki Irie. Key-value memory in the brain. *Neuron*, 113(11):1694–1707.e1, 2025. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2025.02.029. URL https://www.sciencedirect.com/science/article/pii/S0896627325001722.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 30–45, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.3. URL https://aclanthology.org/2022.emnlp-main.3/.

Government of Canada. Consumer privacy protection act, 2022. Bill C-27 (Digital Charter Implementation Act, 2022).

Government of Japan. Act on the protection of personal information, 2003. Act No. 57 of 2003, as amended.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Shengyuan Hu, Yiwei Fu, Steven Wu, and Virginia Smith. Unlearning or obfuscating? jogging the memory of unlearned LLMs via benign relearning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=fMNRYBvcQN.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14389–14408, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.805. URL https://aclanthology.org/2023.acl-long.805/.

Madhur Jindal. Gibberish detector: High-accuracy text classification model, 2021. URL https://huggingface.co/madhurjindal/autonlp-Gibberish-Detector-492513457.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew Bo Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Ariel Herbert-Voss, Cort B Breuer, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam Alfred Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Ian Steneker, David Campbell, Brad Jokubaitis, Steven Basart, Stephen Fitz, Ponnurangam Kumaraguru, Kallol Krishna Karmakar, Uday Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=xlr6AUDuJz.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, Kush R. Varshney, Mohit Bansal, Sanmi Koyejo, and Yang Liu. Rethinking machine unlearning for large language models, 2024. URL https://arxiv.org/abs/2402.08787.

Ximing Lu, Sean Welleck, Jack Hessel, Liwei Jiang, Lianhui Qin, Peter West, Prithviraj Ammanabrolu, and Yejin Choi. QUARK: Controllable text generation with reinforced unlearning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=5HaIds3ux5O.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary Chase Lipton, and J Zico Kolter. TOFU: A task of fictitious unlearning for LLMs. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=B41hNBoWLo.

Anmol Mekala, Vineeth Dorna, Shreya Dubey, Abhishek Lalwani, David Koleczek, Mukund Rungta, Sadid Hasan, and Elita Lobo. Alternate preference optimization for unlearning factual knowledge in large language models. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 3732–3752, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.252/.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6f1d43d5a82a37e89b0665b33bf3a182-Paper-Conference.pdf.

Vaidehi Patil, Peter Hase, and Mohit Bansal. Can sensitive information be deleted from LLMs? objectives for defending against extraction attacks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=7erlRDoaV8.

Amit Peleg, Naman Deep Singh, Naama Pearl, Bibhabasu Mohapatra, and Matthias Hein. Forget narrowly, retain broadly: Unlearning as an asymmetric generalization problem. *arXiv preprint arXiv:2607.09236*, 2026.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Priyanka Sen, Alham Fikri Aji, and Amir Saffari. Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (eds.), *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 1604–1619, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.138/.

Jaime Sevilla, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, and Pablo Villalobos. Compute trends across three eras of machine learning. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2022. doi: 10.1109/IJCNN55064.2022.9891914.

William F. Shen, Xinchi Qiu, Meghdad Kurmanji, Alex Iacob, Lorenzo Sani, Yihong Chen, Nicola Cancedda, and Nicholas D. Lane. LLM unlearning via neural activation redirection. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=teB4aqJsNP.

State of California. California consumer privacy act of 2018, 2018. Cal. Civ. Code §§1798.100–1798.199.

Kushal Tirumala, Aram Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. Memorization without overfitting: Analyzing the training dynamics of large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 38274–38290. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/fa0509f4dab6807e2cb465715bf2d249-Paper-Conference.pdf.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. Mmlu-pro: a more robust and challenging multi-task language understanding benchmark. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Xinwei Wu, Junzhuo Li, Minghui Xu, Weilong Dong, Shuangzhi Wu, Chao Bian, and Deyi Xiong. DEPN: Detecting and editing privacy neurons in pretrained language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2875–2886, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.174. URL https://aclanthology.org/2023.emnlp-main.174/.

Puning Yang, Qizhou Wang, Zhuo Huang, Tongliang Liu, Chengqi Zhang, and Bo Han. Exploring criteria of loss reweighting to enhance LLM unlearning. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=mGOugCZlAq.

Jin Yao, Eli Chien, Minxin Du, Xinyao Niu, Tianhao Wang, Zezhou Cheng, and Xiang Yue. Machine unlearning of pre-trained large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8403–8419, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.457. URL https://aclanthology.org/2024.acl-long.457/.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024b. Curran Associates Inc. ISBN 9798331314385.

Yunzhi Yao, Ningyu Zhang, Zekun Xi, Mengru Wang, Ziwen Xu, Shumin Deng, and Huajun Chen. Knowledge circuits in pretrained transformers. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024c. Curran Associates Inc. ISBN 9798331314385.

Charles Yu, Sullam Jeoung, Anish Kasi, Pengfei Yu, and Heng Ji. Unlearning bias in language models by partitioning gradients. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 6032–6048, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.375. URL https://aclanthology.org/2023.findings-acl.375/.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=MXLBXjQkmb.

Zhiwei Zhang, Fali Wang, Xiaomin Li, Zongyu Wu, Xianfeng Tang, Hui Liu, Qi He, Wenpeng Yin, and Suhang Wang. Catastrophic failure of LLM unlearning via quantization. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=lHSeDYamnz.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *ACM Trans. Intell. Syst. Technol.*, 15(2), February 2024a. ISSN 2157-6904. doi: 10.1145/3639372. URL https://doi.org/10.1145/3639372.

Kairan Zhao, Meghdad Kurmanji, George-Octavian Barbulescu, Eleni Triantafillou, and Peter Triantafillou. What makes unlearning hard and what to do about it. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 12293–12333. Curran Associates, Inc., 2024b. doi: 10.52202/079017-0394. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/16e18fa3b3add076c30f2a2598f03031-Paper-Conference.pdf.

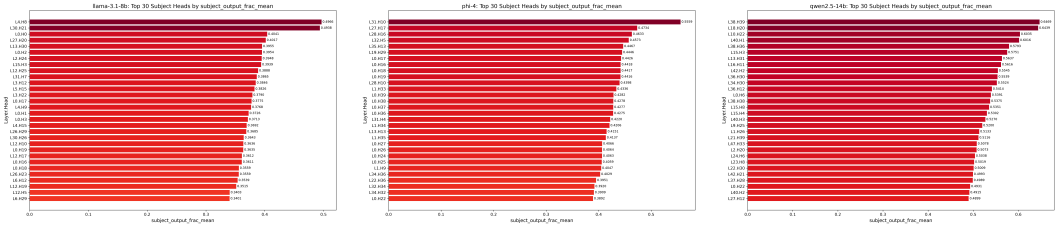
Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency, 2025. URL https://arxiv.org/abs/2310.01405.

A APPENDIX

A.1 ATTENTION HEADS SELECTION FOR ATI FEATURE SOURCES.

We select attention heads as ATI feature sources by analyzing sampled Mintaka questions for entity-sensitive head behavior.



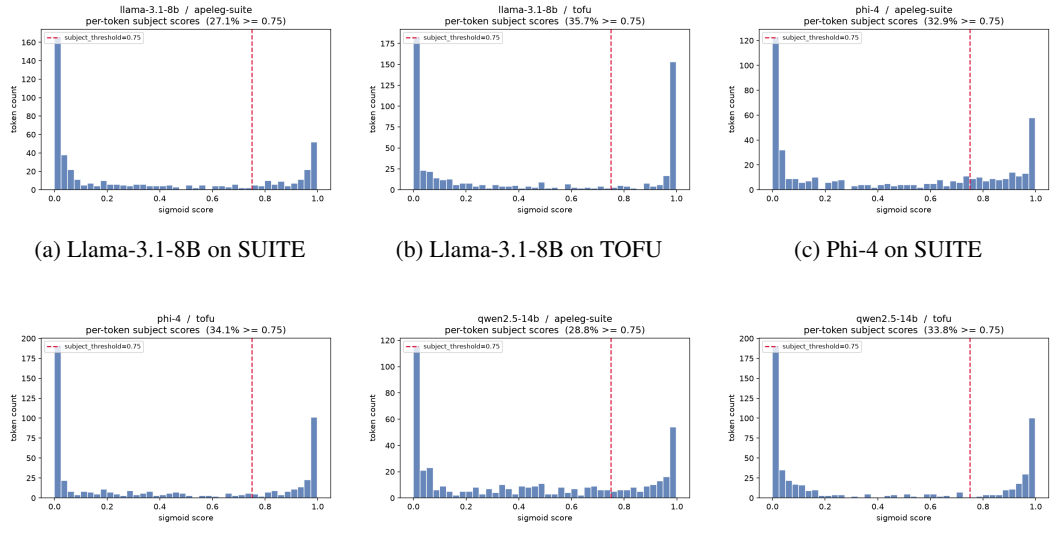
(a) Llama-3.1-8B top subject heads (b) Phi-4 top subject heads (c) Qwen2.5-14B top subject heads

Figure 5: Top subject heads of three experimented models, measured by mean subject output fraction ρ_j

A.2 QUALITATIVE RESULTS FOR ATI

Figure 6 shows a U-shaped distribution of ATI's per-token sigmoid scores, with most values concentrated near zero or one. This suggests that ATI distinguishes candidate anchored-term tokens from other tokens with relatively few intermediate scores.

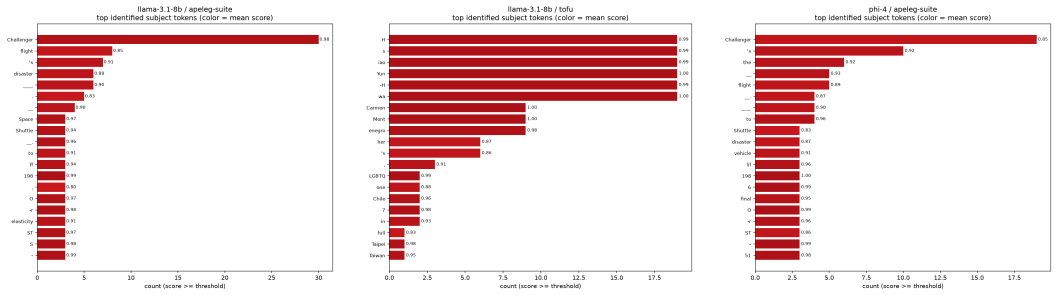
Figure 7 shows top tokens identified by ATI on TOFU and SUITE dataset. These top tokens accurately correspond to each dataset's relevant topics. For example: TOFU shows author names like "`Hsiao`" or "`Carmen`", while SUITE shows "`Challenger`" or "`Space`", which is one of the included topic.



(a) Llama-3.1-8B on SUITE (b) Llama-3.1-8B on TOFU (c) Phi-4 on SUITE

(d) Phi-4 on TOFU (e) Qwen2.5-14B on SUITE (f) Qwen2.5-14B on TOFU

Figure 6: Histograms of all per-token sigmoid scores with $\tau = 0.75$



(a) Llama-3.1-8B on SUITE (b) Llama-3.1-8B on TOFU (c) Phi-4 on SUITE

(d) Phi-4 on TOFU (e) Qwen2.5-14B on SUITE (f) Qwen2.5-14B on TOFU

Figure 7: Most frequently identified subject tokens

Terms	Frequency
Valentina Duarte Ribeiro	19
Elsie Marais Van der Westhuizen	18
Laila Amari El Fassi	18
Linh Mai Nguyen	18
Anjali Deshpande	18
Aisha Alzubari	18
Layla Mansour El-Masri	17
Timur Akhmetovich Kovalenko	17
Alejandro García Herrera	17
Emir Kaya Demir	17
Koffi Adande Chevalier	17
Jun-Suk Hyun	17
Daniyar Nauryzbayev	17
Lian Huiling	17
Arif Sudirman	17
Minh Anh Linh Vu	16
Larysa Petrovna Kovalenko	16
Siriporn Nakharin	16
Dmitri Ivanovich Kuznetsov	15
Gabriel Castillo Herrera	14

Table 5: Term frequencies.

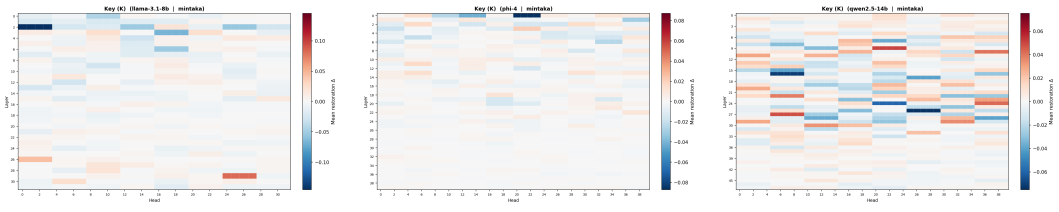
A.3 Replacement "unknown" tokens for representation steering.

We extract candidate names using frequency-based voting over capitalized n-grams, retaining those appearing at least 10 times. Since each author in the held-out TOFU data has approximately 20 Q&A pairs, this threshold favors recurring author names over incidental capitalized expressions. Table 5 shows identified terms used as target steering.

A.4 Detailed causal results for unlearning layers selection.

Figure 8 shows the causal heatmaps of 3 models. Specific target layers for each model are:

- *Llama-3.1-8B-Instruct*: $29, 26, 27, 11, 30$
- *Phi-4*: $13, 11, 22, 18, 10$
- *Qwen2.5-15B*: $10, 11, 30, 32, 23$



(a) Llama-3.1-8B (b) Phi-4 (c) Qwen2.5-14B

Figure 8: Causal heatmaps of 3 models on CounterFact dataset with interventions to K component. We use Δ restorative scores between clean and corrupted runs.

A.5 Unlearning reproducibility configs.

Experiments are conducted on four NVIDIA H100 GPUs. Table 6 summarizes the training configurations. We use a learning rate of 10^{-3} for our restricted K/V updates and 10^{-5} for the baselines. To accommodate the higher memory requirements of baseline training on Phi-4 and Qwen2.5-14B, we reduce the microbatch size from 4 to 1 and increase gradient accumulation from 8 to 32, preserving

the effective batch size. All methods use AdamW and train for 10 epochs; other shared training settings remain unchanged.

Table 6: Training configurations for different methods.

Method family	Model	Train batch	Grad accum	LR	Epoch	Optimizer
Ours	Llama / Phi-4 / Qwen	4	8	1e−3	10	AdamW
Other baselines	Llama	4	8	1e−5	10	AdamW
	Phi-4 / Qwen	1	32	1e−5	10	AdamW

A.6 Detailed unlearning results

Tables 7 and 8 report forgetting and retention metrics on SUITE and TOFU, respectively. Tables 9 and 10 report *Common Benchmark* performance after unlearning on each dataset.

Table 7: Unlearning results on SUITE.

Method	Exact memorization	Extraction strength	Forget Q&A gibberish	Forget ROUGE	Retain ROUGE
Llama-3.1-8B-Instruct					
NPO	0.0604	0.5345	0.6745	0.0121	0.1995
RMU	0.0575	0.5338	0.0211	0.0078	0.1093
SatImp	0.4248	0.6231	0.7884	0.0626	0.1748
SimNPO	0.3484	0.5848	0.7511	0.0483	0.3099
UNDIAL	0.0171	0.5338	0.3942	0.0285	0.0389
Ours (kv6)	0.1531	0.5505	0.7692	0.0079	0.4246
Phi-4					
NPO	0.0732	0.5386	0.3264	0.0073	0.3169
RMU	0.1109	0.5328	0.3233	0.0322	0.3384
SatImp	0.5869	0.6980	0.6618	0.1266	0.2376
SimNPO	0.4055	0.5989	0.7283	0.0849	0.3164
UNDIAL	0.0370	0.5326	0.4261	0.0259	0.0475
Ours (kv5)	0.3457	0.6023	0.4423	0.0909	0.4765
Qwen2.5-14B-Instruct					
NPO	0.1745	0.5317	0.8353	0.0262	0.2676
RMU	0.0246	0.5090	0.5144	0.0230	0.1159
SatImp	0.6017	0.7093	0.7593	0.1294	0.2182
SimNPO	0.4097	0.5818	0.8039	0.0780	0.2428
UNDIAL	0.0744	0.5125	0.8879	0.0491	0.0614
Ours (kv5)	0.3447	0.5795	0.5654	0.1074	0.4311

Table 8: TOFU unlearning results.

Method	Exact memorization	Extraction strength	Forget Q&A Prob	Forget Q&A ROUGE	Forget Q&A gibberish	Model utility
Llama-3.1-8B-Instruct						
NPO	0.5277	0.0523	0.1103	0.2653	0.8973	0.4953
RMU	0.0346	0.0329	0.0021	0.1120	0.0489	0.5264
SatImp	0.8152	0.2362	0.5297	0.4172	0.8813	0.5592
SimNPO	0.7713	0.1599	0.4592	0.3868	0.8939	0.5788
UNDIAL	0.3619	0.0377	0.1355	0.3388	0.8404	0.3797
Ours (kv6)	0.1603	0.0331	0.0086	0.0667	0.3090	0.5519
Phi-4						
NPO	0.5341	0.0631	0.1195	0.2131	0.8767	0.4756
RMU	0.0860	0.0331	0.0093	0.1052	0.0878	0.5844
SatImp	0.7985	0.2027	0.5034	0.4058	0.8759	0.5947
SimNPO	0.7652	0.1480	0.4499	0.3802	0.8897	0.5929
UNDIAL	0.2592	0.0329	0.0948	0.4370	0.8638	0.4657
Ours (kv6)	0.4285	0.0454	0.0910	0.2129	0.6709	0.5944
Qwen2.5-14B-Instruct						
NPO	0.5109	0.0552	0.0984	0.3054	0.9521	0.4698
RMU	0.5408	0.0973	0.1541	0.2818	0.8940	0.5467
SatImp	0.7352	0.1526	0.3966	0.3690	0.8838	0.5642
SimNPO	0.7177	0.1421	0.3716	0.3573	0.8881	0.5686
UNDIAL	0.2606	0.0302	0.0946	0.3403	0.9018	0.4163
Ours (kv6)	0.4269	0.0930	0.0850	0.2230	0.9078	0.4937

Table 9: Common benchmark results after unlearning on SUITE.

Method	MMLU	MMLU-Pro	IFEval	ARC Challenge	GPQA Main (zero-shot)	HumanEval	GSM8K (CoT)
Llama-3.1-8B-Instruct							
NPO	0.6362	0.3220	0.3826	0.5870	0.2879	0.4451	0.5944
RMU	0.6498	0.3394	0.4713	0.5401	0.3214	0.5549	0.7187
SatImp	0.6446	0.3255	0.3512	0.5631	0.3058	0.4573	0.6300
SimNPO	0.6385	0.3234	0.3457	0.5648	0.3103	0.4573	0.6277
UNDIAL	0.2295	0.1166	0.1331	0.2517	0.2455	0.0000	0.0000
Ours (kv6)	0.6283	0.3151	0.1368	0.4206	0.3125	0.3659	0.4602
Reference (no unlearning)	0.6831	0.3762	0.4492	0.5512	0.3571	0.5793	0.7619
Phi-4							
NPO	0.7948	0.5244	0.1331	0.6826	0.3929	0.4390	0.2456
RMU	0.7997	0.5267	0.0259	0.6280	0.4107	0.4878	0.9272
SatImp	0.7935	0.5155	0.1238	0.6468	0.3750	0.6829	0.9174
SimNPO	0.7921	0.5135	0.1460	0.6570	0.3817	0.6829	0.9234
UNDIAL	0.3697	0.1618	0.0388	0.4855	0.2455	0.0976	0.1766
Ours (kv5)	0.723	0.4365	0.0295	0.5972	0.3325	0.4573	0.8627
Reference (no unlearning)	0.8038	0.5317	0.0444	0.5597	0.4129	0.5122	0.9333
Qwen2.5-14B-Instruct							
NPO	0.7891	0.5253	0.5915	0.6834	0.3571	0.6585	0.6740
RMU	0.7884	0.5332	0.5730	0.6118	0.3482	0.6341	0.7672
SatImp	0.7935	0.5285	0.5231	0.6681	0.3638	0.4878	0.7726
SimNPO	0.7938	0.5243	0.5342	0.6596	0.3571	0.4939	0.7566
UNDIAL	0.3188	0.1853	0.1368	0.4744	0.2455	0.0000	0.0591
Ours (kv5)	0.7674	0.4947	0.5415	0.6715	0.3058	0.5304	0.6914
Reference (no unlearning)	0.7990	0.5390	0.6118	0.6220	0.3683	0.6463	0.7589

Table 10: Common benchmark results after unlearning on TOFU.

Method	MMLU	MMLU-Pro	IFEval	ARC Challenge	GPQA Main (zero-shot)	HumanEval	GSM8K (CoT)
Llama-3.1-8B-Instruct							
NPO	0.6677	0.3547	0.3475	0.5410	0.3080	0.5488	0.6323
RMU	0.6774	0.3634	0.4233	0.5597	0.3281	0.5671	0.7521
SatImp	0.6782	0.3678	0.4381	0.5631	0.3080	0.5366	0.7346
SimNPO	0.6769	0.3691	0.4251	0.5648	0.3080	0.5549	0.7369
UNDIAL	0.3411	0.1206	0.0074	0.4394	0.2500	0.0000	0.0281
Ours (kv6)	0.6540	0.3490	0.2791	0.5145	0.3259	0.4817	0.6277
Reference (no unlearning)	0.6831	0.3762	0.4492	0.5512	0.3571	0.5793	0.7619
Phi-4							
NPO	0.8022	0.5306	0.0074	0.5964	0.3973	0.4573	0.8855
RMU	0.7965	0.5226	0.0055	0.5922	0.3772	0.3232	0.9022
SatImp	0.8015	0.5330	0.0055	0.5836	0.4018	0.3841	0.9151
SimNPO	0.8003	0.5292	0.0055	0.5853	0.4018	0.3841	0.9083
UNDIAL	0.7089	0.3959	0.0277	0.5606	0.3371	0.5915	0.4162
Ours (kv6)	0.7421	0.4436	0.0037	0.5631	0.3705	0.3293	0.8779
Reference (no unlearning)	0.8038	0.5317	0.0444	0.5597	0.4129	0.5122	0.9333
Qwen2.5-14B-Instruct							
NPO	0.8001	0.5311	0.5028	0.6527	0.3817	0.6585	0.8764
RMU	0.7933	0.5264	0.5527	0.6288	0.3326	0.6402	0.8711
SatImp	0.7970	0.5338	0.5564	0.6433	0.3549	0.6341	0.8067
SimNPO	0.7965	0.5341	0.5601	0.6451	0.3504	0.6220	0.7824
UNDIAL	0.7670	0.4764	0.0018	0.5495	0.3415	0.0915	0.6808
Ours (kv6)	0.7834	0.5026	0.5046	0.6681	0.3214	0.6646	0.8006
Reference (no unlearning)	0.7990	0.5390	0.6118	0.6220	0.3683	0.6463	0.7589