ALTER: Asymmetric LoRA for Token-Entropy-Guided Unlearning of LLMs

Xunlei Chen[1, *], Jinyu Guo[1, *, †], Yuang Li[1], Zhaokun Wang[1, †], Yi Gong[1], Jie Zou[2], Jiwei Wei[2], Wenhong Tian[1, †]

[1]School of Information and Software Engineering, University of Electronic Science and Technology of China
[2]School of Computer Science and Engineering, University of Electronic Science and Technology of China
wzk@std.uestc.edu.cn, {guojinyu, tian_wenhong}@uestc.edu.cn

Abstract

Large language models (LLMs) have advanced to encompass extensive knowledge across diverse domains. Yet controlling what LLMs should not know is important for ensuring alignment and thus safe use. However, effective unlearning in LLMs is difficult due to the fuzzy boundary between knowledge retention and forgetting. This challenge is exacerbated by entangled parameter spaces from continuous multi-domain training, often resulting in collateral damage, especially under aggressive unlearning strategies. Furthermore, the computational overhead required to optimize State-of-the-Art (SOTA) models with billions of parameters poses an additional barrier. In this work, we present **ALTER**, a lightweight unlearning framework for LLMs to address both the challenges of knowledge entanglement and unlearning efficiency. ALTER operates through two phases: (I) high entropy tokens are captured and learned via the shared A matrix in LoRA, followed by (II) an asymmetric LoRA architecture that achieves a specified forgetting objective by parameter isolation and unlearning tokens within the target subdomains. Serving as a new research direction for achieving unlearning via token-level isolation in the asymmetric framework. ALTER achieves SOTA performance on TOFU, WMDP, and MUSE benchmarks with over 95% forget quality and shows minimal side effects through preserving foundational tokens. By decoupling unlearning from LLMs' billion-scale parameters, this framework delivers excellent efficiency while preserving over 90% of model utility, exceeding baseline preservation rates of 47.8-83.6%.

Code — https://github.com/MastrOrigami/ALTER.git

Introduction

LLMs have demonstrated remarkable capabilities in downstream task processing (Lee et al. 2020; Wang et al. 2025) and content generation through unprecedented model scale expansion (Achiam et al. 2023) and diversification of pretraining data growth (Zhao et al. 2023). However, there are significant challenges in controlling the generation of sensitive content (Carlini et al. 2022), private information, or illegal content (Shi et al. 2024a). As the legal provisions of the *General Data Protection Regulation* (GDPR) and the *"right to be forgotten"* gain increasing attention. (Grynbaum et al. 2023), unlearning for LLMs has emerged as a rapidly growing research area (Cha et al. 2024a), aiming to selectively eliminate the influence of specific knowledge from deployed models (Yao et al. 2024b; Si et al. 2023).

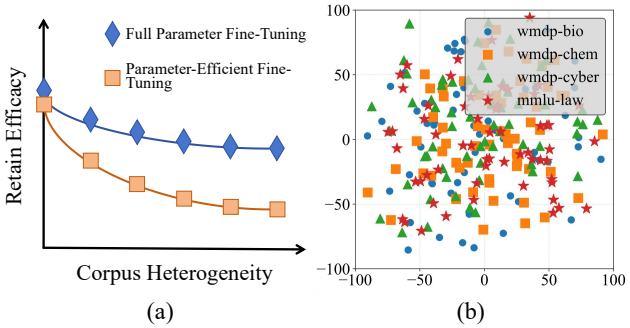

Figure 1: (a) The impact of corpus heterogeneity on the performance of FT/PEFT. (b) The chaos in the LoRA parameter space caused by corpus heterogeneity in the WMDP dataset.

Current unlearning methods comprise: (1) Prompt-based (Liu et al. 2024a) and auxiliary model (Ilharco et al. 2023; Ji et al. 2024) techniques that induce forgetting without parameter modification but exhibit limited generalization, robustness; (2) Fine-tuning (FT) utilizing losses (e.g., gradient ascent on forgetting sets (Liu et al. 2024b)) that face scalability challenges for billion-parameter models and knowledge consistency issues; (3) Model editing employing local modifications like Low-Rank Adaptation (LoRA) (Hu et al. 2022) or unlearn layers (Yu et al. 2025). As a leading parameter-efficient fine-tuning (PEFT) technique, LoRA optimizes merely 0.1%-10% parameters, accelerates training 5-20x (He et al. 2021), preserves knowledge integrity, and inherently enables regulatory "Proof Unlearning" through its modular architecture.

However, the target domains of current unlearning scenarios (Fig.1a) exhibit complex subdomains and task diversity. Despite forgetting strategies, residual information from heterogeneous data (Carlini et al. 2021, 2022) persists via shared parameters (Wang et al. 2024), inducing parameter coupling in both PEFT and FT (Fig.1b). This entanglement

*These authors contributed equally to this work.
†Corresponding author

reduces unlearning efficacy (Liu et al. 2025) and risks "over-forgetting" (Yao et al. 2024c; Shi et al. 2024a). Therefore, a natural but non-simple question arises: How can PEFT achieve efficient unlearning while preserving overall performance in multi-domain coupled parameter spaces?

Building on prior discussion, we note that asymmetric LoRA architectures offer unique advantages: prior studies (Tian et al. 2024) demonstrate that the shared matrix A typically captures universal knowledge while individual B matrices adapt to discrepancy knowledge (Fig.3A).

Inspired by this foundation, we propose the Asymmetric LoRA for Token-Entropy-Guided Unlearning (ALTER) framework, which achieves parameter isolation across forgetting subtasks and decoupling between forgetting and retention tasks. This design eliminates catastrophic forgetting caused by parameter entanglement in traditional methods. To precisely remove token-level sensitive knowledge, we introduce Token-Entropy-Guided: via token-wise entropy modeling, foundational tokens (high entropy) are preserved in shared matrix A, while task-specific tokens are stored in distinct B matrices. This solves public knowledge miss (grammatical tokens) from sentence-level forgetting, establishes a dynamic forgetting boundary preserving knowledge topology integrity, and enables interpretability through information entropy modeling.

ALTER achieves SOTA performance on TOFU, WMDP, and MUSE benchmarks with over 95% forget quality. By decoupling unlearning from the billion-scale parameters of LLMs, this framework is highly efficient while preserving over 90% of the model's utility (compared to 47.8-83.6% for baselines), and demonstrates minimal side effects by preserving foundational tokens.

In summary, our contributions can be listed as follows:

- We propose ALTER, an unlearning framework that avoids fine-tuning LLMs' weights via LoRA, achieving parameter isolation across forgetting subtasks and decoupling between forgetting and retention tasks.
- We introduce a token-entropy-based localization method that enables precise, selective forgetting and alleviates over-forgetting.
- We conduct extensive experiments across three knowledge unlearning tasks to validate forgetting quality, model utility, and fluency.

Background and Motivation

Mainstream Unlearning Strategies We model LLMs as probabilistic models π_θ. Under single LoRA, the forward computation is $\mathbf{W} = \mathbf{W}_0 + \mathbf{BA}$. The forgetting set $\mathcal{D}_f$ and retention set $\mathcal{D}_r$ are comprised of question-answer pairs (q, a). Mainstream parameter-directional forgetting strategies minimize π_θ's joint loss:

$$\begin{aligned}\arg\min_{\pi_\theta} = \beta \underbrace{\mathbb{E}_{(q,a)\sim\mathcal{D}_f}\left[l_f(q \mid a; \pi_\theta)\right]}_{\text{forgetting term}} \\ +\gamma \underbrace{\mathbb{E}_{(q,a)\sim\mathcal{D}_r}\left[l_r(q \mid a; \pi_\theta)\right]}_{\text{retaining term}}.\end{aligned} \tag{1}$$

The loss l_f prevents the model from answering questions in the forgetting set $\mathcal{D}_f$, while l_r preserves base capabilities and responses to $\mathcal{D}_r$. Weighting coefficients β and γ balance these losses. Specific details of l_f and l_r are in Appendix D.

Asymmetric LoRA (AsymLoRA). The one-to-many architecture extends the single LoRA, with its formula defined as:

$$\mathbf{W} = \mathbf{W}_0 + \Delta\mathbf{W} = \mathbf{W}_0 + \sum_{i=1}^{N} \omega_i \cdot \mathbf{B}_i\mathbf{A}. \tag{2}$$

$\mathbf{W}_0$ denotes the frozen pretrained weight matrix, while $\Delta\mathbf{W}$ represents AsymLoRA's incremental update matrix. Here, $\mathbf{A} \in \mathbb{R}^{r\times k}$ and $\mathbf{B}_i \in \mathbb{R}^{d\times r}(i = 1, 2, \ldots, N)$ with weighting coefficients ω_i scaling each $\mathbf{B}_i$'s contribution, where N is the number of $\mathbf{B}$ matrices and r is the rank of the low-rank decomposition.

Token Entropy Shannon entropy (Shannon 1948) quantifies predictive information in models. For LLMs, the entropy of an input token x_t is defined as:

$$H(x_t) = -\sum_{i=1}^{V} p_{t,i} \log p_{t,i}. \tag{3}$$

$p_{t,i} \in \mathbb{R}^V$ denotes the model's token probability at position t for vocabulary index i. "Token entropy" describes the uncertainty in the probability distribution over token generation at position t in LLMs, which is determined by the model's output logits at that position.

Observation I: Asymmetric LoRA efficiently achieves parameter isolation for unlearning. AsymLoRA enables efficient unlearning by establishing a natural boundary between shared and task-specific knowledge (Fig.2). The subset requiring removal from $\mathcal{D}_f$ is denoted as the specific forgetting sub-domain, with its optimization objective defined as:

$$\begin{aligned}\min_{\omega_f^d, \omega_r} \beta\mathbb{E}_{(q,a)\sim\mathcal{D}_f^d} \underbrace{\left[l_f\left(q \mid a; \left(\pi_\theta + \omega_f^d \cdot \mathbf{B}_d\mathbf{A}\right)\right)\right]}_{\text{forgetting term}} \\ +\gamma\,\mathbb{E}_{(q,a)\sim\mathcal{D}_r} \underbrace{\left[l_r\left(q \mid a; \left(\pi_\theta + \mathbf{B}_r\mathbf{A}\right)\right)\right]}_{\text{retaining term}}.\end{aligned} \tag{4}$$

Here, ω_f^d modulates the contribution weights for head $\mathbf{B}_d$. $\mathbf{B}_r$ serves as the globally shared expert for retained knowledge. The inherent isolation mechanism transforms the complex problem of forgetting heterogeneous data into localized optimization tasks for specific data subsets. Further interpretation of parameter space is provided in Appendix F.1.

Although AsymLoRA achieves subtask-level parameter isolation ***(Observation I)***, its instance-level granularity treats QA pairs atomically, ignoring internal token heterogeneity. Uniform unlearning damages language structures (e.g., transition words like "however"). True selective forgetting thus requires token-level intervention to precisely remove target knowledge while preserving structural tokens. This idea of precise forgetting based on token-level knowledge localization leads to the following discovery.

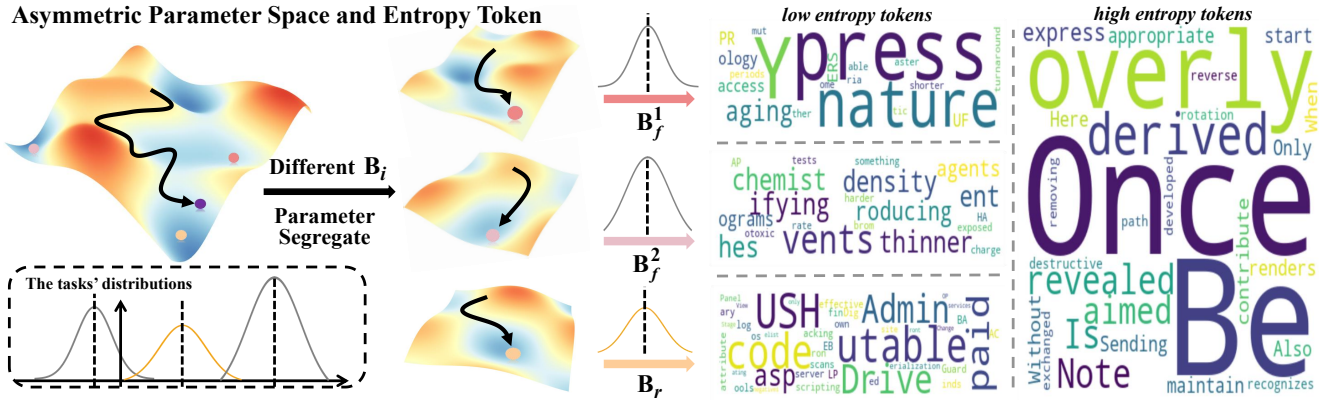

Figure 2: Conceptual illustration of our unlearning framework. After achieving explicit parameter isolation with the AsymLoRA structure, word clouds from the WMDP dataset show that task-specific forgetting experts $\mathbf{B}_i$ and the retention expert $\mathbf{B}_r$ process low entropy tokens (left), whereas the shared matrix $\mathbf{A}$ processes high entropy tokens (right).

Observation II: Token-level entropy demonstrates a robust bimodal distribution aligned with linguistic functions, maintaining stability during model adaptation. The word clouds in Fig.2 reveal functional token separation by entropy: high entropy tokens are mostly structural elements (e.g., "however"/"therefore"), while low entropy tokens contain knowledge-intensive content (e.g., "entities"). This pattern (Appendix F.2) remains stable during PEFT: more than 87% high entropy tokens retain uncertainty and more than 92% low entropy tokens maintain determinism when fine-tuning Llama3-8B on WMDP. Entropy conservation thus enables token-level knowledge management: low entropy tokens permit precise factual removal while high entropy tokens preserve structural integrity. More token analysis is provided in Appendix F.

Motivation: Knowledge forgetting must simultaneously accommodate structural modularity and token-level specificity, integrating architecture isolation with entropy-driven token partitioning via Tsallis entropy. Building upon our observations, the asymmetric $\mathbf{A}$-$\mathbf{B}$ decomposition establishes vertical isolation: the shared matrix $\mathbf{A}$ inherently adapts to high entropy tokens (H>2.0), which are connectives or logical words that represent language structure invariants; the individual matrices $\mathbf{B}_i$ carry the domain expert knowledge contained in the low entropy tokens (Appendix F.2), where $\mathbf{B}_f^1, \ldots, \mathbf{B}_f^d$ handle knowledge to be forgotten in specific sub-domains, and $\mathbf{B}_r$ strengthens knowledge in the retention set. The entropy hierarchy creates horizontal distinction (training pipeline in Fig.3): high entropy tokens aggregate in $\mathbf{A}$, low entropy tokens isolate in $\mathbf{B}_i$ experts. However, Shannon entropy's independence assumption contradicts LLMs' non-extensive nature (Tsallis 1988), causing semantic coupling in $\mathbf{B}_i$ and nonlinear $\mathbf{A}$ interactions that break β-γ trade-off in Eq.4. We therefore introduce Tsallis entropy for hierarchical modeling:

$$S_q(x_t) = \frac{1}{q-1}\left(1 - \sum_{i=1}^{V} p_{t,i}^q\right) \quad (q > 0). \tag{5}$$

Here, x_t is the input token, typically a vector. Deformation parameter q models knowledge association strength for dual control: (1) $q < 1$ enhances structural invariance in $\mathbf{A}$ for high entropy tokens. (2) For $\mathbf{B}_i$ experts, $q > 1$ breaks low entropy tokens' cross-domain associations, enabling forgetting in targeted domains while preventing collateral damage. Non-extensive Tsallis entropy establishes a knowledge-specific unlearning manifold between architecture isolation (physical separation of $\mathbf{B}_i$) and entropy partitioning (separation functional high S_q tokens), resolving the core conflict through consolidation of shared knowledge in the $\mathbf{A}$ matrix and targeted perturbation within $\mathbf{B}_i$ experts. Theoretical analysis is provided in Appendix A.

ALTER: Asymmetric LoRA for Token-Entropy-Guided Unlearning

Token-Entropy-Guided Architecture

Based on the above discussion, we propose the Token-Entropy-Guided Architecture. Building upon the isolation framework in Eq.2 above, we extend it to include a retention set:

$$\begin{aligned}\mathbf{W} &= \mathbf{W}_0 + \Delta\mathbf{W} \\ &= \mathbf{W}_0 + \left(\mathbf{B}_r + \sum_{d=1}^{N} \omega_f^d \cdot \mathbf{B}_f^d\right)\mathbf{A}.\end{aligned} \tag{6}$$

Shared matrix $\mathbf{A}$ captures task-invariant knowledge, and is initialized via Kaiming as $\mathcal{N}(0, \sigma^2)$ to align high entropy tokens with pretrained weights. Experts $\mathbf{B}_f^d$ and $\mathbf{B}_r$ minimize KL divergence by mapping task-specific clustering centers to low entropy feature spaces, introducing initial heterogeneity in the parameter distribution entropy $S(\mathbf{B})$. Specifically, each forgetting expert $\mathbf{B}_f^d$ corresponds to sub-domain d, and is initialized from $\mathcal{D}_f^d$'s clustering center (N total domains). The retention expert $\mathbf{B}_r$ preserves knowledge, and is initialized via $\mathcal{D}_r$'s feature distribution.

Training Stage

Forward Propagation We introduce an entropy-based adaptive gating mechanism. For input x_t, the function is de-

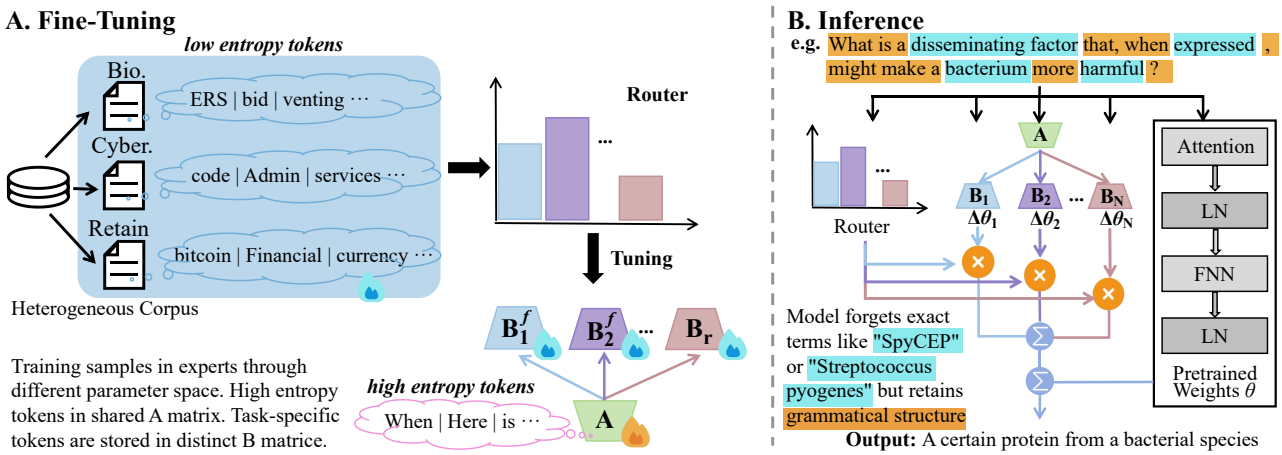

Figure 3: Architecture and workflow of our unlearning framework. During fine-tuning, ALTER first automatically identifies and initializes N intrinsic components (without requiring domain-specific knowledge). Then, guided by entropy, the architecture uses a trainable MoE router that treats each intrinsic component as an expert, automatically assigning training samples to the corresponding component. High entropy tokens (red fire), inherently adapted to the shared $\mathbf{A}$ matrix, are processed jointly, while low entropy tokens (blue fire) are directed to specialized $\mathbf{B}$ experts for fine-tuning. During inference, ALTER dynamically combines multiple $\mathbf{B}$ matrices using the trained router for flexible and adaptive unlearning.

fined as:

$$g_d(x_t) = \text{softmax}(W_g^T \cdot S_q(x_t)/\tau). \tag{7}$$

Here, W_g^T denotes the transpose of a learnable weight matrix W_g and $S_q(\cdot)$ represents the Tsallis entropy. The routing temperature τ is dynamically adjusted: for high entropy tokens ($S_q(x_t) > 1.2$), we set $\tau = 0.8$ to activate multiple experts and enhance structural robustness; for low entropy tokens ($S_q(x_t) \leq 1.2$), $\tau = 0.01$ is used to enforce single-expert precision routing.

Loss Calculation and Backpropagation Inspired by the classic hinge loss (Cortes and Vapnik 1995), detailed in Appendix B, we design ALTER's hierarchical loss, extending the Eq.4 into a three-level cascaded optimization:

$$\begin{aligned}\min_{\omega_f^d, \omega_r} \quad & \beta \sum_{d=1}^{N} \mathbb{E}_{(q,a)\sim\mathcal{D}_f^d} \underbrace{\left[\mathcal{L}_{\text{IHL}}\left(q \mid a; \left(\pi_\theta + \omega_f^d \mathbf{B}_f^d \mathbf{A}\right)\right)\right]}_{\text{entropy-weighted forgetting}} \\ & + \gamma \mathbb{E}_{(q,a)\sim\mathcal{D}_r} \underbrace{\left[l_r\left(q \mid a; \left(\pi_\theta + \mathbf{B}_r \mathbf{A}\right)\right)\right]}_{\text{entropy-protected retention}} \\ & + \lambda \underbrace{\mathbb{E}_{x_t \in \mathcal{H}} \left[\|\nabla_{\mathbf{A}} S_q(x_t)\|^2\right]}_{\text{structural invariance}}.\end{aligned} \tag{8}$$

Inspired by the classic hinge loss, we reverse the optimization direction and define $\mathcal{L}_{\text{IHL}}$ which enables precise knowledge forgetting on low entropy tokens by suppressing target prediction probabilities while promoting next-best tokens. Retention loss l_r strengthens core capabilities, and gradient constraints on $\mathbf{A}$'s high entropy tokens preserve structural integrity. Strict gradient isolation is enforced: each forgetting expert $\mathbf{B}_f^d$ updates via $\nabla l_f|_{\mathcal{L}_d}$, $\mathbf{B}_r$ is updated via $\nabla l_{\text{IHL}}|_{\mathcal{L}_r}$. Shared matrix $\mathbf{A}$ is updated solely via high entropy tokens ($\nabla S_q|_{\mathcal{H}}$).

Inference Stage

During inference, AsymUnlearn employs an entropy-aware conditional computation architecture to balance efficiency and accuracy. Given an input token x_t in multi-forgetting tasks, real-time Tsallis entropy $S_q(x_t)$ computation triggers differentiated paths based on entropy thresholds:

$$y = \begin{cases} \mathbf{W}_0 x + \mathbf{A}x + \sum_{d=1}^{k} g_d(x_t) \mathbf{B}_f^d \mathbf{A} x & S_q(x_t) > 1.2 \\ \mathbf{W}_0 x + \mathbf{B}_{i^*} \mathbf{A} x & S_q(x_t) \leq 1.2. \end{cases} \tag{9}$$

High entropy tokens use multi-expert fusion: aggregate $\mathbf{A}$ and top-3 $\mathbf{B}_f^d$ outputs via pre-trained gating weights $g_d(x_t)$. This design preserves structural integrity and cross-domain consistency. Low entropy tokens activate single-expert bypass: engage only the highest-weight $\mathbf{B}_{i^*}$, avoiding redundant computations and localization interference.

Experiment

Experiment Setting

Datasets We evaluate using three benchmarks: TOFU (Maini et al. 2024), which consists of 200 synthetic author profiles (4k QA pairs) with unlearning sets of 1%, 5%, and 10%; WMDP (Li et al. 2024b), which assesses knowledge in sensitive domains such as biosafety, cybersecurity, and chemical safety (note: there is no training set for the chemistry subset, so we did not use it); and MUSE-HarryPotter (Rowling 2023), a copyright-focused benchmark for book unlearning. Additionally, to evaluate the general capabilities of LLMs, we include the fact-based question answering benchmark MMLU (Hendrycks et al. 2021).

Baselines and Backbones For the selection of baseline methods and backbone models, we strictly follow the default recommendations provided by each benchmark.

Model/Category	FT & Model Editing								LoRA Variations$_{(r=8)}$		
Llama3-8B	Base	RMU	ELM	GA	RL	NPO	NPO_KL	NPO_GD	LoRA	AsymLoRA	Ours
WMDP-Bio↓	71.2	49.4	33.3	**23.3**	24.7	58.1	64.3	56.2	28.7	25.7	24.4
WMDP-Cyber↓	*45.3*	*37.0*	*26.6*	***24.0***	*26.6*	*34.4*	*41.3*	*33.1*	*32.1*	*28.8*	*25.6*
MMLU↑	62.1	40.1	57.2	24.8	23.0	50.1	56.0	51.9	39.6	55.3	**57.8**
Flu-mean↑	*2.97*	*2.96*	*3.07*	*1.00*	*1.00*	*3.07*	*2.97*	*3.03*	*2.23*	*2.23*	***3.46***
Flu-var↓	1.91	1.88	2.18	0.00	0.00	1.86	1.96	2.08	1.42	1.42	**1.17**
Zephyr-7B	Base	RMU	ELM	GA	RL	NPO	NPO_KL	NPO_GD	LoRA	AsymLoRA	Ours
WMDP-Bio↓	64.4	30.2	29.6	24.7	24.0	63.5	64.3	63.5	34.2	27.1	**24.4**
WMDP-Cyber↓	*44.3*	*27.3*	*27.2*	*26.8*	*24.7*	*43.6*	*45.3*	*43.1*	*32.1*	*26.3*	***24.0***
MMLU↑	58.5	57.8	56.2	23.0	26.4	57.8	57.4	**58.0**	37.4	54.1	56.4
Flu-mean↑	*2.97*	*2.92*	*2.99*	*1.00*	*1.00*	*2.98*	*2.95*	*2.93*	*2.47*	*2.47*	***3.11***
Flu-var↓	1.98	2.03	2.00	0.00	0.00	2.12	1.91	2.08	1.57	1.57	**1.33**

Table 1: **Multiple-choice accuracy of five LLMs on the WMDP benchmark (forget) and the full MMLU (retain) after unlearning.** Our method enables model-agnostic unlearning, reducing WMDP accuracy to around 25% while preserving MMLU performance, and achieves stable forgetting with superior fluency by avoiding entangled errors in knowledge domains.

For TOFU, we use Llama3-8B and Llama2-7B (Touvron et al. 2023) as the base models. Baselines include Grad. Diff (Liu, Liu, and Stone 2022), Gradient Ascent (GA) (Thudi et al. 2022), KL Min (Nguyen, Low, and Jaillet 2020), NPO, NPO_KL, and NPO_GD (Zhang et al. 2024).

For WMDP, we use Zephyr-7B (Tunstall et al. 2023) and Llama3-8B (Dubey et al. 2024) as the base models. The baselines are Representation Misdirection for Unlearning (RMU) (Maini et al. 2024), Erasure of Language Memory (ELM) (Gandikota et al. 2025), GA, Random Label (RL) (Yao et al. 2024a), NPO, NPO_KL, and NPO_GD.

For HarryPotter, we use Llama-2-7B (Touvron et al. 2023) as the base model and compare it with six baselines: WHP (Liu et al. 2024b), ELM, GA, GD, KL, and NPO.

Evaluation Metrics We verify method effectiveness via multiple established metrics, and compute their harmonic mean for comprehensive statistical results.

For TOFU, forget quality uses a KS test p-value (higher means greater distribution similarity between unlearned/retained models). Model utility assesses retention set and real-world performance. Metric details in Appendix E.2.1.

For WMDP, forget quality is evaluated via multiple-choice accuracy on bio/cybersecurity questions. Model utility is assessed on MMLU through multiple-choice accuracy.

For HarryPotter, forget quality is evaluated via BLEU (Papineni et al. 2002) and ROUGE-L (Lin 2004) between ground-truth and generated completions of 200-token prefixes. Model utility is evaluated on MMLU.

We evaluate generated content fluency via GPT-4o (average of 5 responses). While not perfectly aligned with human judgment, this approach provides a practical solution (Li et al. 2024a; Shi et al. 2024b). For WMDP, fluency assessment uses open-ended reasoning generation instead of direct ABCD option scoring. Details in Appendix E.2.4.

Configurations The experiment configurations are as follows: learning rate is $\eta_B = 10^{-3}$ and $\eta_A = 10^{-5}$, the weights for the forget loss and retain loss are set to $\beta = 1.0$, $\gamma = 1.0$, $\lambda = 0.01$, batch size = 4, epoch = 3. The hardware and software configurations used in our experiments are as follows. CPU: Intel(R) Xeon(R) Platinum 8468V, 800MHZ, 48cores; GPU: NVIDIA TESLA H800 80 GB; Operating system: Ubuntu 20.04; Deep learning framework: PyTorch 2.4.1.

Method	Harry Potter				
	Forget Perf.			Retain Perf.	
	BLEU	R-L	ASG↓	MMLU↑	Ful.↑
Original	74.8	85.1	74.1	46.3	4.0
Retain	1.9	9.8	0	47.8	2.8
Fine-tune	6.4	17.2	5.9	46.0	1.9
GA	0	0	6.0	26.9	1.0
GD	3.9	14.5	3.4	43.6	1.8
NPO	1.5	5.3	2.5	42.7	2.9
KL	1.2	8.9	**0.8**	41.1	3.1
WHP	23.6	17.9	14.9	44.4	2.5
ELM	8.1	9.0	2.7	**44.6**	2.8
LoRA	7.2	11.5	3.5	38.9	2.3
A-LoRA	5.9	10.4	1.9	43.8	2.3
Ours	4.7	9.6	1.3	**44.6**	**3.3**

Table 2: Performance on HarryPotter dataset. R-L and Ful. denote the ROUGE-L score and fluency-mean, respectively. A-LoRA denotes the AsymLoRA. Instead, average similarity gap (ASG) serves as the target for forget performance.

Main Results

Hazardous Knowledge Unlearning Our method achieves strong model-agnostic unlearning: on WMDP (Tab.1), it reduces biosafety/cybersecurity accuracy to near-random (∼25%) while preserving full MMLU performance across models. While GA/RL methods achieve comparable WMDP reduction, they catastrophically degrade MMLU performance to 23-26%. AsymLoRA and our approach

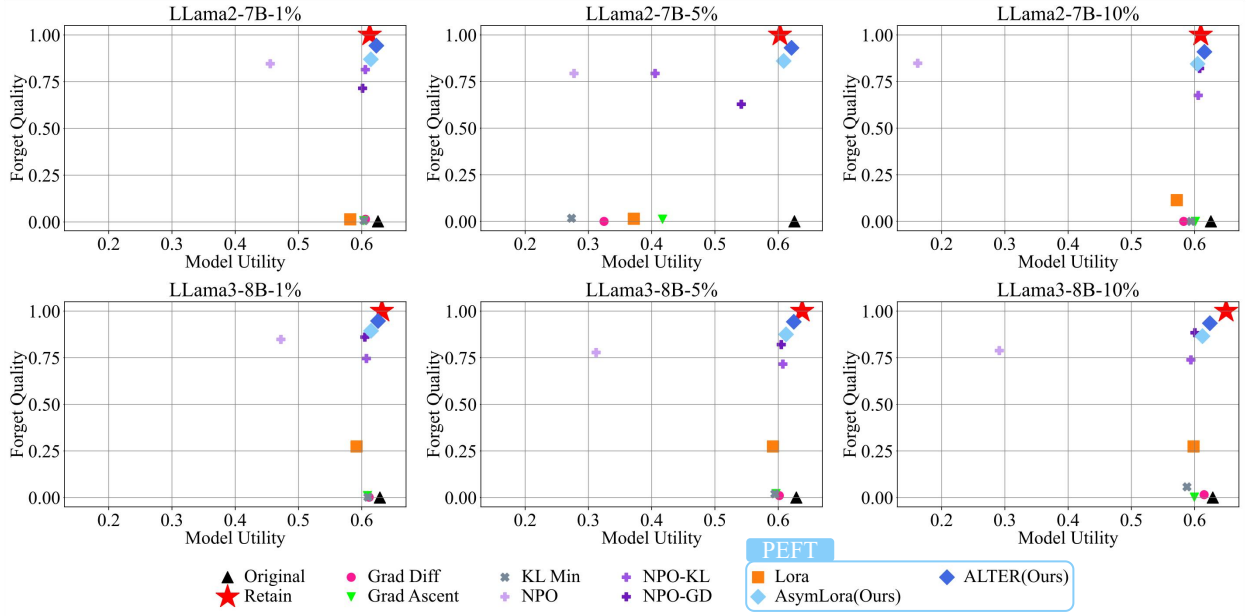

Figure 4: Utility-forgetting trade-off at 1%/5%/10% unlearning ratios for Llama2-7B (top) and Llama3-8B (bottom). GradDiff/Ascent and KLMin show low forgetting efficacy or severe utility loss. NPO incurs utility drops. Standard LoRA maintains utility but minimal forgetting gain. Our AsymLoRA/ALTER achieve near-complete forgetting with Retain-matched utility.

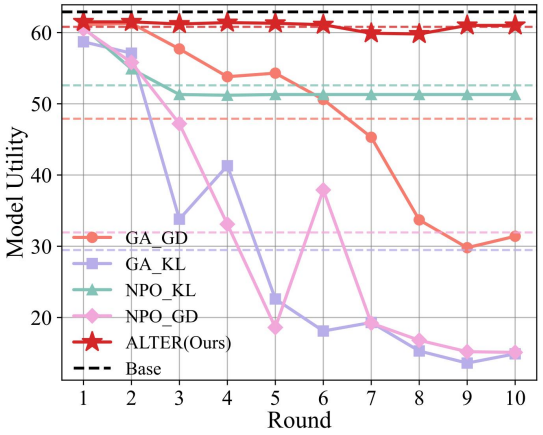

Figure 5: Average model utility of baselines across Sequential Unlearning rounds for TOFU-injected Llama3-8B, with the forgetting set expanded from 1% to 10%.

employ asymmetric architectures for dual parameter isolation—between unlearning subtasks and between unlearning and retention tasks—enabling effective forgetting without utility compromise. Furthermore, our method achieves stable high-precision forgetting. On Llama3-8B/Zephyr-7B, it shows the highest Flu-mean (3.46/3.11) and lowest Flu-var (1.17/1.33), indicating superior generation consistency. In contrast, RMU/ELM reduce WMDP scores but suffer lower fluency, while NPO variants degrade MMLU performance. This suggests existing methods introduce entangled errors in related knowledge domains. We avoid this by preserving shared knowledge via high entropy tokens, ensuring precise removal while maintaining knowledge integrity.

Entity Unlearning The TOFU entity unlearning results (Fig.4) demonstrate that AsymLoRA/ALTER achieves near-perfect forget quality while maintaining Retain-utility across architectures and forgetting ratios, a performance superiority attributable to our architectural innovation. Unlike gradient-based methods that induce severe capability degradation through destructive parameter updates, or NPO variants that exhibit compromised forget quality due to blunt regularization, our framework overcomes standard LoRA's rigidity via entropy-driven token-level isolation. This enables surgical knowledge removal while preserving structural integrity, evidenced by the approximation result with the Retain model's performance. Crucially, these results validate our core solution: asymmetric knowledge partitioning fundamentally decouples forgetting precision from capability preservation, achieving Pareto-optimal balance through targeted parameter isolation at minimal computational cost.

Copyrighted Unlearning The HarryPotter copyrighted unlearning results are shown in Tab.2. We achieve text similarity scores comparable to the Retain model. The high-low entropy mechanism effectively distinguishes copyrighted entities from functional words, maintaining the Model Utility close to original levels. Strong baselines KL minimization and ELM demonstrate competitive ASG and general utility performance but exhibit noticeable fluency reduction after decoupled training. The GD method alleviates model collapse yet still incurs significant MMLU degradation. In contrast, our approach achieves near-optimal ASG while preserving superior MMLU and fluency performance,

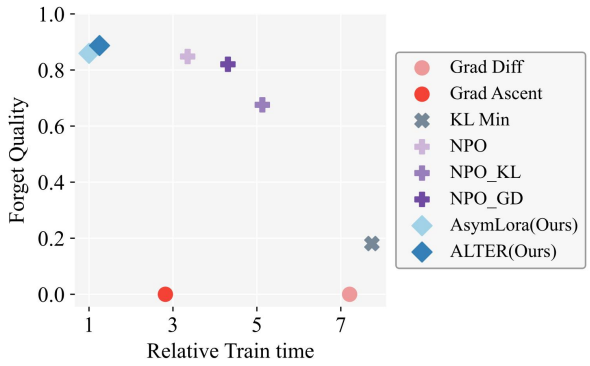

Figure 6: Trade-off between forget quality and relative training time for Llama2-7B on TOFU-10%. The top-left corner indicates better forgetting performance and efficiency.

demonstrating the balanced capability of the entropy-based strategy.

Analysis

Continue Unlearning As shown in Fig.5, ALTER demonstrates exceptional stability in model utility preservation during continuous unlearning, maintaining performance near base model levels with minimal degradation. In contrast, baseline methods exhibit progressive deterioration: gradient-based approaches (GA_GD, GA_KL) show severe utility loss, while NPO variants display moderate degradation. Among baselines, NPO_KL preserves utility most effectively yet remains substantially inferior to ALTER.

Unlearning Time Efficiency Time efficiency is critical for LLM unlearning, especially compared to retraining from scratch. As shown in Fig.6, our method reduces time costs by 86.1% to 87.1% using the AsymLoRA framework, achieving orders-of-magnitude higher forgetting quality. Baseline LoRA harms model utility, while AsymLoRA achieves near-optimal quality with minimal time cost (unit: 1.0). ALTER further improves performance at a modest 1.25× cost. By decoupling shared learning of high entropy tokens and targeted forgetting of low entropy tokens, our approach minimizes computation while maximizing unlearning efficacy, offering efficient solutions for large-scale models.

Ablation Study

All LoRA variants use rank=8 to balance effectiveness and performance. Details of different ranks for LoRA and the positions of decomposition modules are provided in Appendix G.

Related Work

Unlearning GD (Liu, Liu, and Stone 2022; Thaker et al. 2025) applies gradient ascent to make the model "forget" specific content and introduces an additional loss term to constrain the deviation between the unlearned and original models. (Yu et al. 2023) identifies bias-related neurons using Integrated Gradients and performs gradient ascent only on these neurons(Liu et al. 2025), minimizing impact on other model capabilities. (Wang et al. 2023a) balances forgetting effectiveness with performance retention by minimizing the KL divergence between the "unlearned model" and the "original model" on non-forgetting corpora. The SOUL algorithm (Jia et al. 2024) optimizes the parameter update direction using second-order information approximated from the Hessian matrix, achieving unlearning while preserving model utility. Recently, such as RMU (Dang et al. 2025) and LUNAR (Shen et al. 2025), use methods similar to guided vectors (Cao et al. 2024; Cha et al. 2024b) and dedicated UNL tokens (Yu et al. 2025) to locate and modify local model parameters associated with target knowledge, thereby redirecting the model into an inability space.

Multi-LoRA Architecture LoRA (Hu et al. 2022) leverages low intrinsic dimensionality (Aghajanyan, Zettlemoyer, and Gupta 2020) via trainable low-rank matrices in frozen models, efficiently approximating gradient updates. Its low latency and performance drove wide adoption. Subsequent multi-LoRA variants enhance efficiency and stability: (Huang et al. 2024) uses domain-adaptive adapter combinations; (Wang et al. 2023b) reduces parameter dependency via horizontal scaling. These collectively advance hybrid LoRA architectures (Zadouri et al. 2024). MoE-LoRA (Dou et al. 2023) integrates LoRA with Mixture-of-Experts (MoE) to reduce multi-task interference via task-specific adapters. HydraLoRA (Tian et al. 2024) employs asymmetric LoRA with automatic clustering and MoE for efficient adaptation.

Entropy Entropy has been widely applied across various areas of LLMs: semantic entropy measures information density via uncertainty assessment (Guo, Chen et al. 2025); entropy regularization penalizes low entropy predictions (Miller et al. 2002; Pereyra, Tucker et al. 2017) while maximizing prediction entropy (Setlur et al. 2022), enhancing adversarial robustness (Jagatap et al. 2022) and domain generalization (Zhao et al. 2020). In unlearning, methods constrain entropy by maximizing (Peer et al. 2022; Jha and Reagen 2025) on forgetting sets to reduce target confidence and minimizing on retention sets to preserve discriminative ability (Entesari, et al. 2025; Tarun, Chundawat et al. 2023; Jung 2025).

Conclusion

In this work, we propose ALTER, a novel and universal method for LLM unlearning, establishing a new paradigm for parameter-efficient unlearning. ALTER leverages an asymmetric LoRA structure and token entropy to establish a dynamic forgetting boundary that preserves the integrity of the model's knowledge topology while eliminating target information. By decoupling the unlearning process from the LLMs' billions of parameters, ALTER delivers an efficient unlearning framework. With minimal side effects, it maintains model utility comparable to the Retain model across varying architectures. Our experiments across three unlearning tasks validate ALTER's effectiveness, setting a foundation for responsible AI deployment in real-world scenarios.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (Grants 62306067, 62402093, W2433163), Sichuan International Science and Technology Innovation Cooperation Project (ID 2024YFHZ0317), Sichuan Science and Technology Program (ID 2025ZNSFSC0479), Chengdu Science and Technology Bureau Project (ID 2024-YF09-00041-SN), the Postdoctoral Fellowship Program (Grade C) of the China Postdoctoral Science Foundation (Grant GZC20251053), and Huawei Funding (ID H04W241592; partially supported by UESTC Kunpeng&Ascend Center of Cultivation).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Aghajanyan, A.; Zettlemoyer, L.; and Gupta, S. 2020. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*.

Cao, Y.; Zhang, T.; Cao, B.; et al. 2024. Personalized steering of large language models: Versatile steering vectors through bi-directional preference optimization. *NeurIPS 2024*, 37: 49519–49551.

Carlini, N.; Ippolito, D.; Jagielski, M.; Lee, K.; Tramer, F.; and Zhang, C. 2022. Quantifying memorization across neural language models. In *The Eleventh ICLR*.

Carlini, N.; Tramer, F.; Wallace, E.; Jagielski, M.; Herbert-Voss, A.; Lee, K.; et al. 2021. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, 2633–2650.

Cha, S.; Cho, S.; Hwang, D.; Lee, H.; et al. 2024a. Learning to unlearn: Instance-wise unlearning for pre-trained classifiers. In *AAAI 2024*, volume 38, 11186–11194.

Cha, S.; Cho, S.; Hwang, D.; and Lee, M. 2024b. Towards robust and cost-efficient knowledge unlearning for large language models. In *Adaptive Foundation Models: Evolving AI for Personalized and Efficient Learning*.

Cortes, C.; and Vapnik, V. 1995. Support-vector networks. *Machine learning*, 20(3): 273–297.

Dang, H.-T.; et al. 2025. On Effects of Steering Latent Representation for Large Language Model Unlearning. In *AAAI 2025*, volume 39, 23733–23742.

Dou, S.; Zhou, E.; Liu, Y.; Gao, S.; Zhao, J.; Shen, W.; et al. 2023. LoRAMoE: Revolutionizing Mixture of Experts for Maintaining World Knowledge in Language Model Alignment. *CoRR*, abs/2312.09979.

Dubey, A.; Jauhri, A.; Pandey, A.; et al. 2024. The Llama 3 Herd of Models. *CoRR*, abs/2407.21783.

Entesari, T.; ; et al. 2025. Constrained Entropic Unlearning: A Primal-Dual Framework for Large Language Models. *arXiv preprint arXiv:2506.05314*.

Gandikota, R.; Feucht, S.; Marks, S.; and Bau, D. 2025. Erasing Conceptual Knowledge from Language Models.

Grynbaum, M. M.; et al. 2023. The Times sues OpenAI and Microsoft over AI use of copyrighted work. *The New York Times*, 27(1).

Guo, J.; Chen, X.; et al. 2025. HASH-RAG: Bridging Deep Hashing with Retriever for Efficient, Fine Retrieval and Augmented Generation. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Findings of the Association for Computational Linguistics: ACL 2025*, 26847–26858. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-256-5.

He, J.; Zhou, C.; Ma, X.; Berg-Kirkpatrick, T.; and Neubig, G. 2021. Towards a unified view of parameter-efficient transfer learning. *arXiv preprint arXiv:2110.04366*.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; et al. 2021. Measuring Massive Multitask Language Understanding. In *ICLR 2021*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Huang, C.; et al. 2024. LoraHub: Efficient Cross-Task Generalization via Dynamic LoRA Composition. In *First Conference on Language Modeling*.

Ilharco, G.; Ribeiro, M. T.; Wortsman, M.; Schmidt, L.; Hajishirzi, H.; and Farhadi, A. 2023. Editing models with task arithmetic. In *The Eleventh ICLR*.

Jagatap, G.; Joshi, A.; Chowdhury, A. B.; Garg, S.; and Hegde, C. 2022. Adversarially robust learning via entropic regularization. *Frontiers in artificial intelligence*, 4: 780843.

Jha, N. K.; and Reagen, B. 2025. Entropy-Guided Attention for Private LLMs. *CoRR*, abs/2501.03489.

Ji, J.; et al. 2024. Reversing the Forget-Retain Objectives: An Efficient LLM Unlearning Framework from Logit Difference. In *NeurIPS 2024*.

Jia, J.; et al. 2024. Soul: Unlocking the power of second-order optimization for llm unlearning. *arXiv preprint arXiv:2404.18239*.

Jung, D. 2025. EntUn: Mitigating the forget-retain dilemma in unlearning via entropy. *ICT Express*.

Lee, J.; et al. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234–1240.

Li, J.; Wei, Q.; Zhang, C.; Qi, G.; Du, M.; Chen, Y.; et al. 2024a. Single image unlearning: Efficient machine unlearning in multimodal large language models. *NeurIPS 2024*, 37: 35414–35453.

Li, N.; Pan, A.; Gopal, A.; Yue, S.; Berrios, D.; Gatti, A.; et al. 2024b. The WMDP Benchmark: Measuring and Reducing Malicious Use with Unlearning. In *Forty-first ICML*.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Liu, B.; Liu, Q.; and Stone, P. 2022. Continual learning and private unlearning. In *Conference on Lifelong Learning Agents*, 243–254. PMLR.

Liu, C.; et al. 2024a. Large language model unlearning via embedding-corrupted prompts. *NeurIPS 2024*, 37: 118198–118266.

Liu, S.; Yao, Y.; Jia, J.; Casper, S.; Baracaldo, N.; Hase, P.; et al. 2025. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, 1–14.

Liu, Y.; et al. 2024b. Revisiting Who's Harry Potter: Towards Targeted Unlearning from a Causal Intervention Perspective. In *EMNLP 2024*, 8708–8731. Miami, Florida, USA: Association for Computational Linguistics.

Maini, P.; Feng, Z.; Schwarzschild, A.; Lipton, Z. C.; and Kolter, J. Z. 2024. TOFU: A Task of Fictitious Unlearning for LLMs. In *First Conference on Language Modeling*.

Miller, D.; et al. 2002. A global optimization technique for statistical classifier design. *IEEE transactions on signal processing*, 44(12): 3108–3122.

Nguyen, Q. P.; Low, B. K. H.; and Jaillet, P. 2020. Variational bayesian unlearning. *NeurIPS 2020*, 33: 16025–16036.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL 2002*, 311–318.

Peer, D.; Keulen, B.; Stabinger, S.; Piater, J.; and Rodriguez-sanchez, A. 2022. Improving the Trainability of Deep Neural Networks through Layerwise Batch-Entropy Regularization. *Transactions on Machine Learning Research*.

Pereyra, G.; Tucker, G.; et al. 2017. Regularizing Neural Networks by Penalizing Confident Output Distributions.

Rowling, J. K. 2023. *Harry Potter and the sorcerer's stone*. Scholastic Incorporated.

Setlur, A.; Eysenbach, B.; Smith, V.; and Levine, S. 2022. Maximizing entropy on adversarial examples can improve generalization. In *ICLR 2022 Workshop on PAIR 2Struct: Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data*.

Shannon, C. E. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3): 379–423.

Shen, W. F.; Qiu, X.; Kurmanji, M.; Iacob, A.; Sani, L.; Chen, Y.; et al. 2025. LUNAR: LLM Unlearning via Neural Activation Redirection. *CoRR*, abs/2502.07218.

Shi, D.; et al. 2024a. Large Language Model Safety: A Holistic Survey. *CoRR*, abs/2412.17686.

Shi, W.; et al. 2024b. Detecting Pretraining Data from Large Language Models. In *ICLR 2024*.

Si, N.; Zhang, H.; Chang, H.; Zhang, W.; Qu, D.; and Zhang, W. 2023. Knowledge Unlearning for LLMs: Tasks, Methods, and Challenges. *CoRR*, abs/2311.15766.

Tarun, A. K.; Chundawat, V. S.; et al. 2023. Fast yet effective machine unlearning. *IEEE Transactions on Neural Networks and Learning Systems*, 35(9): 13046–13055.

Thaker; et al. 2025. Position: Llm unlearning benchmarks are weak measures of progress. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 520–533. IEEE.

Thudi, A.; Deza, G.; Chandrasekaran, V.; and Papernot, N. 2022. Unrolling sgd: Understanding factors influencing machine unlearning. In *EuroS&P 2022*, 303–319. IEEE.

Tian, C.; et al. 2024. Hydralora: An asymmetric lora architecture for efficient fine-tuning. *NeurIPS 2024*, 37: 9565–9584.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tsallis, C. 1988. Possible generalization of Boltzmann-Gibbs statistics. *Journal of statistical physics*, 52(1): 479–487.

Tunstall, L.; Beeching, E.; Lambert, N.; Rajani, N.; Rasul, K.; Belkada, Y.; et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Wang, L.; Chen, T.; Yuan, W.; and et al. 2023a. KGA: A General Machine Unlearning Framework Based on Knowledge Gap Alignment. In *ACL 61st Annual Meeting Proceedings (Volume 1: Long Papers)*, 13264–13276. Toronto, Canada: Association for Computational Linguistics.

Wang, Q.; et al. 2024. Unlearning with Control: Assessing Real-world Utility for Large Language Model Unlearning. *CoRR*, abs/2406.09179.

Wang, Y.; Lin, Y.; Zeng, X.; and Zhang, G. 2023b. MultiLoRA: Democratizing LoRA for Better Multi-Task Learning. *CoRR*, abs/2311.11501.

Wang, Z.; et al. 2025. Noise-Robustness Through Noise: A Framework combining Asymmetric LoRA with Poisoning MoE. In *NeurIPS 2025*.

Yao, J.; Chien, E.; Du, M.; et al. 2024a. Machine Unlearning of Pre-trained Large Language Models. In *62nd ACL Proc. (Volume 1: Long Papers)*, 8403–8419.

Yao, Y.; et al. 2024b. Large language model unlearning. *NeurIPS 2024*, 37: 105425–105475.

Yao, Y.; et al. 2024c. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 4(2): 100211.

Yu, C.; Jeoung, S.; Kasi, A.; Yu, P.; and Ji, H. 2023. Unlearning bias in language models by partitioning gradients. In *Findings of the Association for Computational Linguistics: ACL 2023*, 6032–6048.

Yu, M.; et al. 2025. UniErase: Unlearning Token as a Universal Erasure Primitive for Language Models. *arXiv preprint arXiv:2505.15674*.

Zadouri, T.; et al. 2024. Pushing Mixture of Experts to the Limit: Extremely Parameter Efficient MoE for Instruction Tuning. In *ICLR 2024*.

Zhang, R.; Lin, L.; Bai, Y.; and Mei, S. 2024. Negative Preference Optimization: From Catastrophic Collapse to Effective Unlearning. In *First Conference on Language Modeling*.

Zhao, S.; Gong, M.; Liu, T.; Fu, H.; and Tao, D. 2020. Domain generalization via entropy regularization. *NeurIPS 2020*, 33: 16096–16107.

Zhao, W. X.; et al. 2023. A Survey of Large Language Models. *CoRR*, abs/2303.18223.