Representation-Aware Unlearning via Activation Signatures: From Suppression to Knowledge-Signature Erasure

Anonymous ACL submission

Abstract

Selective knowledge erasure from LLMs is critical for GDPR compliance and model safety. However, many unlearning methods conflate behavioral suppression with true knowledge removal, allowing latent capabilities to persist beneath surface-level refusals. This work introduces **Knowledge Immunization Framework (KIF)**, a representation-aware architecture that distinguishes true erasure from obfuscation by targeting internal activation signatures rather than surface outputs[1]. KIF combines dynamic suppression of subject-specific representations with parameter-efficient adaptation, enabling durable unlearning without full model retraining. It achieves near-oracle erasure (FQ $\approx$ 0.99 vs. 1.00) while preserving utility at oracle levels (MU = 0.62), substantially narrowing the stability-erasure tradeoff that has constrained prior work. We evaluate both standard foundation models (Llama & Mistral) and reasoning-prior models (Qwen & DeepSeek) across 3B to 32B parameters. We observe strong erasure behavior in standard models (<3% utility drift), while reasoning-prior models reveal fundamental architectural divergence. Our dual-metric evaluation protocol jointly measures surface-level leakage and latent trace persistence to operationalize the obfuscation-erasure distinction. This enables a systematic diagnosis of mechanism-level forgetting behavior across model families and scales.

1 Introduction

Large language models (LLMs) are increasingly deployed in real-world NLP systems, but training on large-scale corpora can cause them to memorize sensitive or copyrighted content. This creates tension with regulations such as the GDPR right to erasure and related data-protection regimes (European Parliament and Council of the European Union, 2016; Zhang et al., 2024a). Unlike databases, where records can be deleted explicitly, LLMs store training information in distributed parameters (Naveed et al., 2025), which makes targeted removal difficult. *LLM unlearning* aims to remove the influence of selected data while preserving general utility (Yao et al., 2024; Ren et al., 2025). Recent LLM unlearning methods avoid full retraining through efficient post-hoc objectives, but two challenges remain: strong forgetting pressure can degrade utility, and apparent forgetting may reflect obfuscation rather than true erasure (Yao et al., 2024; Zhang et al., 2024b; Ren et al., 2025; Sun et al., 2025).

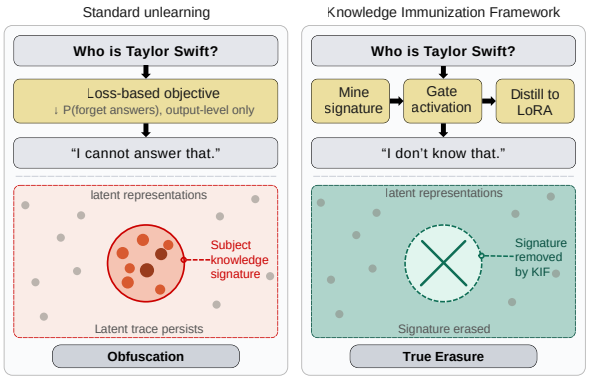

Figure 1: **Comparison of approaches.** Standard unlearning methods may produce obfuscation, hiding answers without removing latent knowledge, whereas KIF targets the internal signature to support durable erasure.

To address these issues, we propose the **Knowledge Immunization Framework (KIF)**, a lightweight unlearning framework that targets subject-linked internal representations rather than only output behavior. KIF first extracts activation signatures using a custom prompt dataset built around real-world entities, then uses these signatures to suppress the targeted knowledge during inference, and finally distills the suppression into LoRA parameters (Hu et al., 2021) with stability-aware updates to limit collateral drift.

We evaluate KIF on open-source chat and rea-

[1] Anonymous repository

soning model families across multiple scales using benchmarks and metrics that probe both surface leakage and residual internal knowledge. We further test cross-benchmark generalization on a 4-subject evaluation-only subset of RWKU (Jin et al., 2024), showing that the mined signatures transfer beyond our custom prompt templates. KIF achieves strong forgetting with minimal utility drift in foundation models and also reveals systematic differences in how reasoning-oriented models trade off erasure and obfuscation under stronger forgetting pressure (Sun et al., 2025; Yoon et al., 2025). Our main contributions are:

1. a real-world entity prompt dataset for extracting subject-linked activation signatures;
2. a representation-aware unlearning framework that combines dynamic suppression with parameter-efficient distillation to improve stability over standard baselines;
3. an evaluation protocol that separates behavioral suppression from durable erasure by jointly measuring leakage and internal signature persistence; and
4. a cross-family, cross-scale analysis of when unlearning remains stable and when it reverts to obfuscation-like behavior.

2 Related Works

Optimization-based Unlearning. Recent LLM unlearning methods mainly avoid full retraining by optimizing objectives that promote forgetting while preserving general utility. Benchmarks such as TOFU (Maini et al., 2024) and MUSE (Shi et al., 2024) have helped standardize this setting. Early approaches rely on gradient-based or constrained objectives (Jang et al., 2023), followed by newer optimization-based methods such as NPO (Zhang et al., 2024b), SimNPO (Fan et al., 2024), AltPO (Mekala et al., 2025), ReLearn (Xu et al., 2025), and UnDIAL (Dong et al., 2025). While these methods differ in objective design, supervision strategy, and stabilization mechanism, they largely define forgetting through changes in training loss or generated outputs. In contrast, our work treats unlearning as a *representation-level* problem by using parameter-efficient updates to weaken the internal subject-linked activation structure that supports a fact.

Memorization and Internal Editing. A complementary line of work studies memorization itself and helps clarify what it means for knowledge to remain inside a language model. Carlini et al. (2023) show that memorization depends on factors such as model scale, data duplication, and prompting conditions. Li et al. (2024) analyze memorization through text, logits, and representations, while Chen et al. (2024) examine it from multiple perspectives spanning both output behavior and internal structure. Together, these works suggest that output suppression alone is not sufficient to establish unlearning if subject-related information remains encoded internally. Related privacy-focused editing methods such as DEPN (Wu et al., 2023) and REVS (Ashuach et al., 2025) also intervene on model internals, but they target memorized sensitive sequences through neuron editing or vocabulary-space rank editing rather than subject-level factual unlearning.

3 Methodology

We introduce a three-stage pipeline: **(1)** *localizing* the target knowledge's internal activation mechanisms through statistical analysis, **(2)** then *suppressing* these mechanisms at inference time using lightweight gating modules we term **Knowledge Suppression Capsules** and **(3)** finally, *distilling* this suppressed behavior permanently into a global LoRA Adapter (Hu et al., 2021) for durable unlearning without full retraining. The training objective combines preference-style supervision for explicit forgetting and stability regularization that minimize collateral drift (Rafailov et al., 2024; Welleck et al., 2019; Kirkpatrick et al., 2017). Figure 2 illustrates our complete pipeline.

3.1 Dataset Construction

The localization stage requires extensive activation profiling to identify subject-specific representations. To this end, we construct a systematically designed dataset of controlled **Entity Prompts** grounded in verifiable facts about real-world subjects. Following Meng et al.'s (2023) *knowledge-tuple* concept, we extract **factual triples** (*subject*, *predicate*, *object*) from Wikipedia and WikiData and instantiate them into prompt templates (e.g., "Is it true that `{subject}`'s `{predicate}` was `{object}`?"). This forms our **Real-World Entity Dataset**. We employ five complementary probe types:

1. **Direct:** Queries specific facts, to extract the most probable retrieval path in the model.
2. **Contextual:** Requests information in a broader

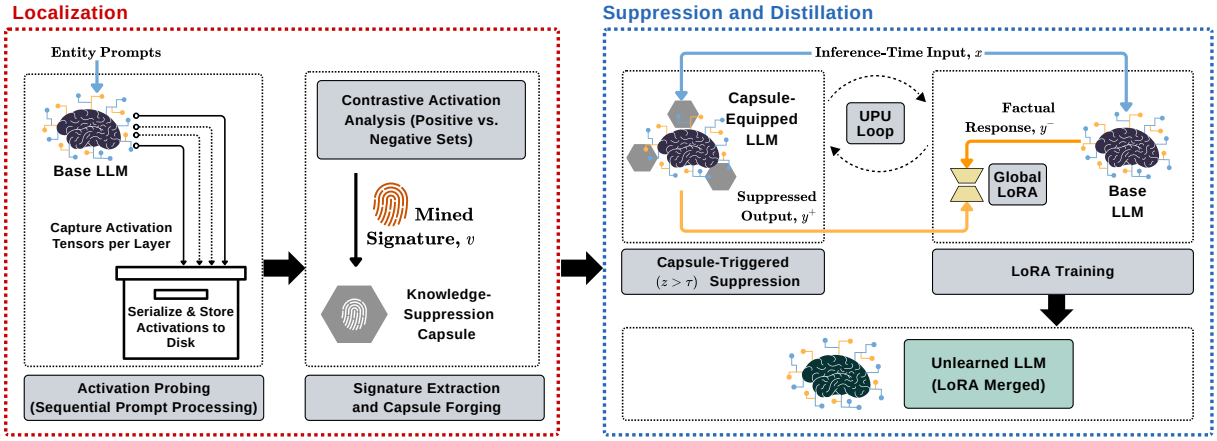

Figure 2: **Knowledge Immunization Framework (KIF) pipeline.** Entity prompts from the Real-World Entity Dataset (3.1) probe MLP activations to mine subject-specific signatures via contrastive analysis. These signatures instantiate Knowledge Suppression Capsules (3.3) at high-salience inference layers, and the UPU Loop (3.4) distills the resulting suppressed behavior into a global LoRA adapter for durable parameter-level knowledge removal.

context to ensure that the signature captures the subject's representation even when it is embedded in a larger narrative flow.

3. **Implicit:** Probes knowledge via confirmation questions to intercept the activations even if the output implicitly involves the subject.
4. **Reasoning:** Encourages explanatory responses, forcing the model to not just retrieve the fact but manipulate it to generate new tokens, to allow for a more complex representation.
5. **Misleading**: Prompts about a subject with incorrect information. These prompts are necessary to probe for internal representation in case of misinformation and to evaluate whether models verify retrieved knowledge versus accepting user-suggested misinformation.

3.2 Localization via Activation-Signature Extraction

The observation that *factual associations are stored as localized, linear relations within the MLP weights* coupled with the use of activations in Causal Tracing (Meng et al., 2023), led us to hypothesize that unlearning is achievable by directly targeting the activation patterns of MLP blocks associated with a specific subject. We evaluate a 4-bit quantized model on the Real World Entity Dataset to harvest internal representations per MLP Layer. While Meng et al. (2023)'s Causal Tracing extracts the activation from a complete MLP block, we gain finer localization by targeting the **intermediate tensors**, e.g. $A^{(\ell)}_{\text{gate}}$, $A^{(\ell)}_{\text{up}}$, and $Y^{(\ell)}_{\text{down}}$ in the Llama family (Grattafiori et al., 2024), which define MLP subspace boundaries $M^{(\ell)} = \{\text{gate}^{(\ell)}, \text{up}^{(\ell)}, \text{down}^{(\ell)}\}$.

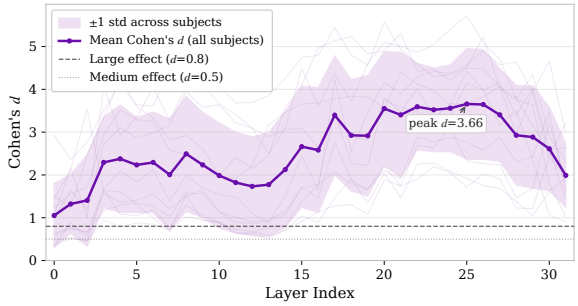

Figure 3: **Mean Cohen's d across transformer layers on Llama-3.1-8B.** Effect sizes indicate strong subject-specific activation separability in mid to final layers.

From the collected corpus of layer-wise activations, we extract subject-specific signatures through a contrastive framework. First, we normalize variable activation shapes to a common 1D representation via token-wise averaging and standardization. We then compare the *Positive activations* from on-topic prompts against *Synthetic Negatives* generated via Gaussian noise. The primary signature is identified using Mean Difference: $\mathbf{d} = \text{mean}(S_{\text{pos}}) - \text{mean}(S_{\text{neg}})$ (where S_{pos} and S_{neg} are the sets of positive and negative activation vectors). This technique has been shown to be effective for isolating directions in a model's latent space corresponding to high-level behaviors by Rimsky et al. (2024). The resulting vector captures the principal direction separating subject-specific activations from baseline representations. To validate separability, we measure Cohen's d (Cohen, 2013) across all 32 MLP layers for each of the 11 subjects (Figure 3). Effect sizes exceed the large-effect threshold (d=0.8) across all layers, peaking

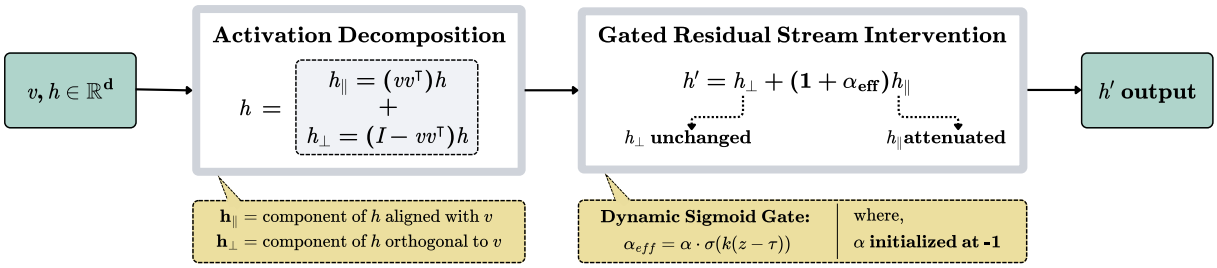

Figure 4: **Capsule based Suppression Pipeline.** The capsule decomposes activation into parallel and orthogonal components relative to subject signature. It applies a gated suppression factor only to the parallel component and passes on the modified activation to the subsequent layer.

at a mean d=3.66 in mid-to-late layers, confirming strongly separable subject-specific activation signatures and directly informing layer selection for capsule placement in Section 3.3.

3.3 Capsule-based Suppression

After extracting subject-specific signatures, we introduce the **Knowledge Suppression Capsule** (Figure 4) to suppress these representations during inference without harming overall model performance. Building on prior work in activation modification (Stoehr et al., 2024; Wen et al., 2025) and rank-one editing (Meng et al., 2023), these lightweight adapters precisely weaken targeted knowledge while leaving unrelated information untouched. Additionally, we introduce an adjustable parameter, α_{eff}, to dynamically control the strength of this suppression.

As illustrated in Figure 4, the core of the capsule is a geometric operator that performs an *Activation Decomposition* of the hidden state h into components parallel ($h_{\parallel}$) and orthogonal ($h_{\perp}$) to the signature vector v, thereby isolating subject-specific information from the rest of the representation. The capsule then scales the parallel component $h_{\parallel}$ through a dynamically gated factor in *Gated Residual Stream Intervention*, while leaving the orthogonal component $h_{\perp}$ unchanged to preserve the model's general utility.

Gated Residual Stream Intervention To determine the attenuation factor, we use a dynamic multiplier: $\alpha_{\text{eff}} = \alpha \cdot \sigma(k(z - \tau))$, with α initialized to -1. This triggers suppression only when the state's projection onto the signature (Z-score 'z') exceeds a statistical threshold τ (with an empirically chosen gain factor gain k). This ensures full suppression occurs only upon statistically significant deviations, isolating interventions to factual anomalies while preserving utility during nominal activations.

The capsule is then implemented via forward hooks with v registered as a fixed buffer. It acts as a lightweight, exportable artifact. It can be dynamically injected at inference or permanently distilled via the **UPU Loop** (3.4), allowing it to be *removed* once true erasure is achieved

3.4 Distillation through Utility Preserving Unlearning Loop (UPU Loop)

While capsules provide immediate suppression, they do not achieve permanent erasure. To convert these suppressions into permanent unlearning, we propose a **Utility Preserving Unlearning Loop (UPU Loop)** to distill the capsules' behavior into a single global LoRA adapter (Hu et al., 2021). We formulate this as a composite loss function $\mathcal{L}$ (Eq. 3) designed to achieve efficient unlearning while mitigating catastrophic forgetting, a hypothesis we validate empirically through ablation studies.

Specifically, the loop simulates user interaction using our custom dataset (3.1). Whenever the capsule's dynamic gate triggers a suppression at layer ℓ, we log a tuple (x, y^{+}, y^{-}) representing the prompt, the capsule-suppressed response, and the base model's factual output. This tuple is then used to optimize the global LoRA parameters ϕ against our composite objective:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{\mathcal{D}_{\text{pref}}}\left[w \cdot \log\sigma\left(\beta \log \frac{p_{\theta}(y^{+}|x)}{p_{\text{ref}}(y^{+}|x)} - \beta \log \frac{p_{\theta}(y^{-}|x)}{p_{\text{ref}}(y^{-}|x)}\right)\right] \quad (1)$$

$$\mathcal{L}_{NTUL} = -\frac{1}{T}\sum_{t} \log\left(1 - \sum_{v \in V_{name}} p_{\theta}(v|x, y^{+}_{<t})\right) \quad (2)$$

$$\mathcal{L} = \mathcal{L}_{DPO} + \lambda_{UL}\mathcal{L}_{UL} + \lambda_{NTUL}\mathcal{L}_{NTUL} + \lambda_{KL}\mathcal{L}_{KL} + \lambda_{EWC}\mathcal{L}_{EWC} \quad (3)$$

In Eq. 1, we adapt *Direct Preference Optimization* (Rafailov et al., 2024) by introducing a scaling

Family	Model Size	Util. Drift	Leakage (↓)	EL10 (↓)	Mechanism State
Standard	Mistral 7B	$+0.50\% \pm 0.45$	$0.00\% \pm 0.00$	0.020 ± 0.015	Type I: True Erasure
Standard	Llama 8B	$-0.45\% \pm 1.59$	$1.10\% \pm 1.90$	0.054 ± 0.047	Type I: True Erasure
Standard	Mistral 3B	$+0.36\% \pm 2.27$	$0.52\% \pm 0.35$	0.054 ± 0.047	Type I: True Erasure
Reasoning	Qwen 14B	$+1.99\% \pm 0.51$	$0.52\% \pm 0.45$	0.035 ± 0.013	Type I: True Erasure
Reasoning	Qwen 32B	$+1.31\% \pm 0.38$	$0.08\% \pm 0.06$	0.024 ± 0.018	Type I: True Erasure
Reasoning	Qwen 8B	$-0.50\% \pm 0.60$	$3.33\% \pm 1.25$	11.03 ± 2.45	Type II: Obfuscation
Reasoning	DeepSeek 8B	$+1.03\% \pm 0.85$	$0.00\% \pm 0.00$	6.19 ± 1.80	Type II: Obfuscation
Standard	Llama 3B	$-0.44\% \pm 0.40$	$0.00\% \pm 0.00$	3.06 ± 0.55	Type II: Obfuscation
Reasoning	Qwen 3B	$+2.22\% \pm 2.42$	$0.43\% \pm 0.37$	1.460 ± 1.480	Type III: Instability
Reasoning	DeepSeek 3B	$+2.15\% \pm 1.74$	$45.6\% \pm 33.6$	15.39 ± 15.54	Type III: Instability

Table 1: **Cross-Model Mechanistic Validation.** We report mean and standard deviation metrics across 3 seeds. Standard foundation models are generally more stable, though Llama 3B remains a Type II exception. Reasoning models exhibit a capacity U-curve: instability at 3B (Type III), stable obfuscation at 8B (Type II), and a return to erasure at larger evaluated scales (Qwen 14B and 32B, Type I).

factor w to amplify penalties when the model deviates from the reference distribution, anchoring unlearning to the base LLM. For surgical knowledge excision, we apply two token-level penalties adapted from Welleck et al. (2019): standard *Factual Unlikelihood* ($\mathcal{L}_{UL}$) on the base model's factual output (y^-) to minimize error token generation, and *Name-Token Unlikelihood* ($\mathcal{L}_{NT\text{-}UL}$), which unlike the standard version, is used to penalize the **aggregate probability** mass of subject name tokens V_{name} within suppressed responses to prevent soft leakage (Eq. 2).

Finally, to ensure the aggressive unlearning does not degrade general reasoning, we constrain the model's deviation on benign anchor prompts using a standard KL Divergence penalty ($\mathcal{L}_{KL}$) (Schulman et al., 2017) (Rafailov et al., 2024) combined with Elastic Weight Consolidation ($\mathcal{L}_{EWC}$) (Kirkpatrick et al., 2017), which penalizes shifts in parameters identified as critical.

4 Experiments and Results

We validate KIF through a dual evaluation strategy that combines mechanistic depth with standardized benchmarking. Using our Real-World Entity Dataset, we examine internal activations to distinguish true knowledge erasure from mere obfuscation. We complement this with TOFU `forget10` (Maini et al., 2024) (a standardized benchmark for LLMU), enabling direct comparison with prior work. Together, these protocols test both mechanism-level behavior and aggregate against baseline performance.

Model and Baseline Selection We evaluate across two distinct model categories chosen to test architectural generalization: standard foundation models (Llama (Grattafiori et al., 2024), Mistral (Jiang et al., 2023)) known for direct fact retrieval, and reasoning-prior architectures (Qwen (Bai et al., 2023), DeepSeek (Bi et al., 2024)) that distribute knowledge across latent reasoning chains. The models span 3B–32B parameters to further probe capacity-dependent behavior.

For the baseline, we compare against five methodological representatives spanning the evolution of LLM unlearning:

- **Gradient Ascent** (Jang et al., 2023): The foundational reverse-optimization approach.
- **GradDiff** (Maini et al., 2024): Adds utility regularization via reference model guidance.
- **NPO** (Zhang et al., 2024b): Applies preference optimization to unlearning.
- **SimNPO** (Fan et al., 2024): the strongest recent baseline, refines NPO by reducing reference model bias
- **IDK/Refusal** (Maini et al., 2024): Represents behavioral suppression without knowledge removal.

Unlike these loss or output based approaches, **KIF** targets internal activations signatures, offering a unique balance on the erasure-utility spectrum. We evaluate against two reference bounds: the **Original** model (lower bound, zero forgetting) and a **Retrained (Oracle)** model (theoretical upper bound, retrained from scratch without the forget set).

Baseline TOFU results (Table 3), are reproduced from Fan et al. (2024); we compute KIF under identical settings. For our custom dataset runs, some hyperparameters like λ, α are tuned empirically; remaining hyperparameters, prompt templates, dataset details, and hardware specifications

are provided in the Appendix.

Evaluation Metrics We employ three diagnostic metrics on the Real-World Entity Dataset (Table. 1), each isolating distinct aspects of knowledge persistence: **i) Utility Drift (Benign PPL Change)** (Jelinek et al., 1977), measured as the change in relative perplexity in unrelated benign text (where negative drift indicates marginal improvement); **ii) Leakage / SMR (Subject Mention Rate)** (Carlini et al., 2021), defined as the percentage of prompts where the model produces the target name (ideally 0.00%); and the **iii) Latent Trace / EL10 Ratio** (Eldan and Russinovich, 2023) summarizes early-step probability mass on the target name; larger values indicate stronger early activation.

We further characterize unlearning outcomes using three interpretable mechanism states, defined by (SMR, EL10) thresholds ($\epsilon = 5\%$ tolerance). This mapping from (SMR, EL10) to mechanism regimes is a descriptive heuristic inspired by Sun et al. (2025).

- **Type I (True Erasure):** SMR $\leq \epsilon$ and EL10 < 1 - internal representations have been attenuated; the model does not possess residual capability.
- **Type II (Obfuscation):** SMR $\leq \epsilon$ and EL10 > 1 - the model internally recognizes the target but is prevented from outputting it through refusal mechanisms.
- **Type III (Instability):** SMR $> \epsilon$ - failed suppression; the intervention cannot consistently control generation.

On TOFU `forget10`, we follow standard protocol and report **Forget Quality (FQ)** (Maini et al., 2024) and **Model Utility (MU)** (Maini et al., 2024), which aggregate forgetting and utility across all forget and retain sets, respectively. In the following subsections, we first report the Real-World Entity Dataset results (Sec: 4.1), then present TOFU results (Sec: 4.2), followed by RWKU Evaluation (Sec: 4.3) and overall analysis in (Sec: 4.4).

4.1 Real-World Entity Dataset Results

Table 1 details KIF's dual-metric performance across architectures, scales and multiple runs. Standard foundation models are substantially more stable than reasoning-prior models: Mistral 3B/7B and Llama 8B achieve Type I erasure with low leakage, low EL10, and small utility drift, while Llama 3B remains in a Type II obfuscation regime despite zero surface leakage. This suggests that standard architectures generally offer more modular subject representations, though stable erasure is not perfectly scale-independent.

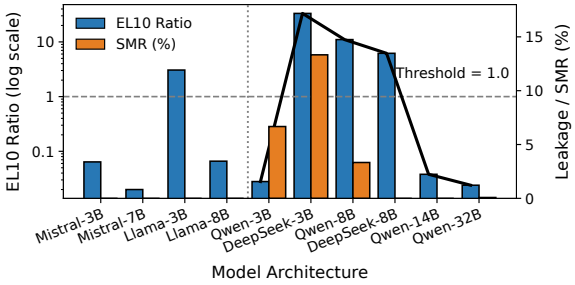

Figure 5: **EL10 and surface leakage across model families and scales.** Most standard models maintain low internal activation, with Llama 3B as a notable exception, whereas reasoning models display a capacity-dependent U-curve.

Benchmark	Metric	Pre (Base)	Post (Adapter)	Delta
ARC-Challenge	acc_char	76.88	76.62	-0.26
OpenBookQA	acc_char	74.80	74.40	-0.40
HellaSwag	acc	76.68	76.84	+0.16
BoolQ	acc	78.93	74.46	-4.46
PIQA	acc_char	72.09	71.65	-0.44
SocialIQA	acc_char	59.88	55.99	-3.89
WinoGrande	acc	68.98	69.46	+0.47
TruthfulQA (MC2)	mc2	40.36	40.87	+0.51

Table 2: **Zero-shot capability retention.** Comparing base and unlearned models reveals robust general capabilities across most benchmarks, with degradation limited to specific tasks like BoolQ and SocialIQA.

Conversely, **reasoning-prior models (Qwen, DeepSeek)** display a capacity-dependent U-curve:

- **3B (Type III):** Unstable, with leakage ranging from 0.43% to 45.6%.
- **8B (Type II):** Low surface leakage but elevated internal activation (EL10 = 11.03 for Qwen 8B; 6.19 for DeepSeek 8B).
- **Larger scales (Type I):** Qwen 14B and Qwen 32B return to Type I behavior, suggesting genuine representation-level erasure.

This suggests reasoning architectures entangle knowledge across latent chains, requiring higher capacity to disentangle. Figure 5 visualizes this, using an EL10 = 1.0 threshold to separate true erasure from obfuscation.

General Capability Retention To confirm that KIF's representation-targeted interventions remain localized, we evaluate zero-shot performance on eight diverse benchmarks spanning reading comprehension, commonsense reasoning, and truthfulness: ARC-Challenge (Clark et al., 2018), OpenBookQA (Mihaylov et al., 2018), BoolQ (Clark

Method	FQ (↑)	MU (↑)
Retrained (Oracle)	1.00	0.62
Original	0.00	0.62
Gradient Ascent	2.2×10^{-16}	0.00
GradDiff	3.7×10^{-15}	0.54
IDK (rejection)	2.9×10^{-14}	0.54
NPO	0.29	0.55
SimNPO	0.45	0.62
Ours (KIF)	**0.99**	**0.62**

Table 3: **TOFU-`forget10` on LLaMA-2-7B-Chat.** We report Forget Quality (FQ) and Model Utility (MU). Baselines are from Fan et al. (2024) (Table A3). KIF achieves near-oracle forgetting while matching original/oracle utility.

et al., 2019), HellaSwag (Zellers et al., 2019), PIQA (Bisk et al., 2020), SocialIQA (Sap et al., 2019), WinoGrande (Sakaguchi et al., 2021), and TruthfulQA (Lin et al., 2022). Table 2 shows post-unlearning accuracy remains within $< 1\%$ of the baseline for most tasks, barring minor drops in BoolQ (-4.46%) and SocialIQA (-3.89%). This confirms PEFT-based updates avoid broad distribution shifts. Furthermore, slight gains in HellaSwag and TruthfulQA suggest "latent pruning" - where removing high-interference traces reduces internal conflicts, improving calibration without sacrificing core capabilities.

4.2 TOFU Benchmark Results

Table 3 details performance on the standardised TOFU `forget10` benchmark (Maini et al., 2024), demonstrating how KIF breaks the sharp Pareto frontier that constrains most prior work, representing a significant qualitative shift. KIF achieves near-oracle erasure (FQ = 0.99) while matching oracle utility (MU = 0.62). In contrast, baselines suffer severe trade-offs: Gradient Ascent (Jang et al., 2023) destroys utility entirely (MU = 0.00); GradDiff and IDK (Maini et al., 2024) suppress leakage behaviorally but lose utility (MU = 0.54); and while NPO (Zhang et al., 2024b) and SimNPO (Fan et al., 2024) improve utility retention, they saturate at weak erasure (FQ 0.29 and 0.45, respectively).

In contrast, KIF operates at a fundamentally different point on the erasure - utility spectrum by targeting internal activation signatures rather than optimizing loss objectives.

4.3 RWKU Evaluation Results

To verify that KIF's mined activation signatures form a true representation of the target subject rather than an overfit to our custom prompt templates, we conduct an **evaluation-only** probe on the 4 subjects shared between our dataset and RWKU (Jin et al., 2024) (*Ariana Grande, Beyoncé, Kanye West, Taylor Swift*). After unlearning these subjects with KIF, we evaluate solely using RWKU's independently-constructed FB, QA, and adversarial-attack probes that do not overlap with our training queries. This experiment is restricted to a strict 4-subject subset with custom wrapper templates and is *not* comparable to the full RWKU leaderboard.

	FB (↓)		QA (↓)		AA (↓)		FM Loss (↑)	
Subject	Pre	Post	Pre	Post	Pre	Post	Pre	Post
Ariana Grande	0.778	0.444	0.729	0.246	0.500	0.316	1.857	2.059
Beyoncé	0.725	0.525	0.796	0.389	0.567	0.183	1.838	2.006
Kanye West	0.750	0.633	0.708	0.354	0.563	0.279	2.127	2.242
Taylor Swift	0.900	0.800	0.733	0.457	0.526	0.346	2.042	2.224
Mean	**0.788**	**0.601**	**0.742**	**0.362**	**0.539**	**0.281**	**1.966**	**2.133**

Table 4: **RWKU generalization probe** (4-subject subset; evaluation only). ROUGE-L recall on FB, QA, and AA probes, and FM loss from the MIA set, before and after KIF. *Not* comparable to RWKU leaderboard numbers.

Table 4 shows consistent forget-set recall reductions across all subjects and probe types (mean FB: $0.788\rightarrow0.601$; QA: $0.742\rightarrow0.362$; AA: $0.539\rightarrow0.281$), including adversarial probes absent from our training distribution. FM loss increases for all subjects (mean $\Delta=+0.167$), confirming that the mined signature is a faithful subject-level representation whose suppression generalizes across independently-constructed probe formats.

4.4 Result Analysis and Discussion

Breaking the stability-erasure trade-off. Prior methods suffer a strict erasure-utility tradeoff: NPO sacrifices utility for modest erasure (FQ 0.29, MU 0.55), while SimNPO preserves utility but saturates at weak erasure (FQ 0.45). KIF substantially narrows this constraint by intervening on representation-aware activation signatures rather than behavioral loss objectives. By targeting subject-specific subspaces via PEFT, KIF avoids the broad distribution shifts that degrade downstream performance. This mechanistic shift achieves near-oracle erasure (FQ 0.99) and utility (MU 0.62) without retraining, establishing a new state-of-the-art on the erasure-utility spectrum.

Why standard and reasoning models diverge fundamentally? The empirical divergence between standard and reasoning models reflects distinct knowledge storage architectures. Unlike stan-

dard models (Llama, Mistral), which store facts locally and scale unlearning robustly (Dai et al., 2022), reasoning models (Qwen, DeepSeek) distribute knowledge across entangled latent Chain-of-Thought dependencies (Wang et al., 2025). This trace entanglement creates a capacity-dependent unlearning constraints: destabilization at 3B (Type III), latent obfuscation at 8B (Type II), and true erasure only at larger scales (observed here at 14B and 32B), where capacity is sufficient to disentangle reasoning traces. This U-curve demonstrates that latent reasoning structures impose strict scaling limits on unlearning for reasoning-prior architectures.

Dual metrics expose hidden failure modes in prior work. Surface-level metrics (SMR alone) create a dangerous blind spot. Type II behavior (e.g., Qwen 8B: 3.33% SMR, 11.03 EL10) proves behavioral suppression can mask latent knowledge persistence. While prior work with output-only metrics would deem this "successful," the model could internally retain the knowledge. Our dual-metric framework distinguishes true erasure from obfuscation. Ablations (Table 5) confirm this: removing representation-level constraints (Gen-UL) shifts Llama 8B toward Type II, proving loss-level supervision alone defaults to output suppression. Standard baselines' inability to diagnose this explains their convergence on suboptimal solutions.

The "latent pruning" hypothesis. An unexpected finding is that utility metrics occasionally improve or remain tight to oracle levels despite aggressive unlearning. We attribute this to *latent pruning*: removing high-interference, low-confidence traces (Eldan and Russinovich, 2023) acts as weak de-noising, reducing representational conflict and improving calibration without harming core semantics. This hypothesis is speculative but consistent with observations: (1) utility drift remains minimal across tasks (Table 2), (2) some tasks show improvement (HellaSwag +0.16, TruthfulQA +0.51), and (3) the effect clusters around the oracle baseline rather than deteriorating monotonically.

Loss compositionality Ablations (Table 5) demonstrates KIF's dependency on the compositional loss 3 design to achieve Type I erasure:

- **Without Name-Token Unlikelihood (NT-UL):** Llama 8B reverts to Type III instability (SMR 3.33%, EL10 0.275) demonstrating that preference optimization alone cannot prevent surface-level leakage.

Exp	SMR (↓)	EL10 (↓)	Util. Drift	State
A: Full	0.00%	0.066	+0.45%	Type I
B: No NT-UL	3.33%	0.275	-1.35%	Type III
C: No Gen UL	0.00%	1.098	-1.29%	Type II

Table 5: **Ablation results on Llama-3.1-8B.** Name-Token Unlikelihood (NT-UL) removal increases leakage and model instability; Generic Unlikelihood removal reduces erasure strength and shifts the mechanism toward stable refusal.

- **Without Generic Unlikelihood:** The model shifts to Type II obfuscation (EL10 = 1.098), relying on decoding-time control rather than actual latent attenuation.

Overall, token-level penalties (for surface behavior) and representation-level pressure (for latent erasure) must operate jointly, explaining why prior methods struggle to balance the erasure-utility trade-off.

5 Conclusion

Our work addressed a critical vulnerability in LLM unlearning: the conflation of behavioral suppression with true knowledge erasure. Through **KIF**, we demonstrated that targeting internal activation signatures rather than loss-level objectives enables near-oracle erasure and utility preservation, substantially narrowing the stability-erasure tradeoff that has constrained prior work in standard foundation models.

Through dual-metric analysis (SMR and EL10), we distinguished unlearning regimes: true erasure (Type I), stable suppression/obfuscation (Type II), and instability (Type III), and showed that reasoning-prior models undergo capacity-dependent transitions across these regimes, while foundation models remain closer to representation-level erasure across scales. An evaluation-only probe on RWKU further shows that these signatures generalize beyond the training prompt distribution, supporting subject-level rather than template-level unlearning. Ablation studies highlight the central role of name-token unlikelihood in enforcing hard non-disclosure.

These findings establish that durable, privacy-compliant unlearning requires targeting internal representations rather than surface behavior, with implications for both immediate deployment in reasoning models and future research into mechanistically-grounded unlearning objectives that explicitly model knowledge persistence across architectural families.

Limitations

This approach relies on *Synthetic Negatives* to approximate off-manifold noise, which may not perfectly distinguish subject-specific activations from genuine off-topic prompts in ambiguous contexts. Additionally, variability in source data leads to inherent class imbalances in our custom dataset; while we mitigate this via oversampling during signature mining, underlying disparities may still impact consistency across subjects.

Experimentally, relative resource constraints (One RTX A6000) limited our scope to 4-bit quantized models (up to 32B parameters), leaving performance on larger, full-precision models to be verified. Furthermore, our evaluation focuses on immediate unlearning; we did not assess sequential unlearning, i.e., whether adding new forget targets preserves forgetting of previously removed subjects without cumulative utility degradation.

Ethical Consideration

In accordance with the ACL Ethics Policy, we acknowledge that the Knowledge Immunization Framework (KIF) carries significant implications for dual-use and data privacy. While aiding legal compliance (e.g., GDPR) (European Parliament and Council of the European Union, 2016; goo, 2014; Zhang et al., 2024a) and sensitive content removal, KIF's surgical erasure could be weaponized for censorship, safety guardrail removal, or the advancement of partisan agendas (Weidinger et al., 2021; Zou et al., 2023; Zhao et al., 2024). Furthermore, we emphasize that "erasure" in high-dimensional neural spaces is inherently relative, as, despite significant trace reduction, current paradigms provide no mathematical guarantee of irrecoverability against extreme adversarial fine-tuning or future jailbreaks (Bourtoule et al., 2021; Sun et al., 2025; Zou et al., 2023; Zhao et al., 2024). Users of this framework must avoid a "false sense of security," as unlearning or deletion of data over non-deterministic systems lack the absolute certainty of traditional database erasure (Bourtoule et al., 2021).

Beyond privacy, the implementation of KIF poses risks to model integrity and societal fairness. Observed "latent pruning" suggests targeted erasure, despite improved calibration, may create "semantic holes," whose entanglement may cause unintended degradation of broader contextual understanding. Such shifts could introduce unpredictable performance drops or demographic biases. However, from a sustainability perspective, KIF offers a significant ethical advantage: by utilizing Parameter-Efficient Fine-Tuning (PEFT) and a loop, it provides a substantially lighter alternative than full-model retraining. This drastically reduces the carbon footprint and environmental cost (Strubell et al., 2019; Patterson et al., 2021) of model maintenance, aligning with global objectives toward sustainable AI development. We acknowledge the use of generative AI models for creating the conceptual visualization in Figure 1 and for improving the clarity and grammatical accuracy of the manuscript.

References

2014. Google spain sl and google inc. v agencia española de protección de datos (aepd) and mario costeja gonzález (case c-131/12). Judgment of the Court (Grand Chamber), 13 May 2014.

Tomer Ashuach, Martin Tutek, and Yonatan Belinkov. 2025. Unlearning sensitive information in language models via rank editing in the vocabulary space. In *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Hongyuan Ding, Kai Dong, Qiushi E, Bohao Feng, and 1 others. 2024. Deepseek llm: Scaling open-source language models with long-termism. *arXiv preprint arXiv:2401.02954*.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, and 1 others. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. 2021. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 141–159.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine

Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, and 1 others. 2021. Extracting training data from large language models. *USENIX Security Symposium*.

Bowen Chen, Namgi Han, and Yusuke Miyao. 2024. A multi-perspective analysis of memorization in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11190–11209, Miami, Florida, USA. Association for Computational Linguistics.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Jacob Cohen. 2013. *Statistical Power Analysis for the Behavioral Sciences*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

Yijiang River Dong, Hongzhou Lin, Mikhail Belkin, Ramon Huerta, and Ivan Vulić. 2025. UNDIAL: Self-distillation with adjusted logits for robust unlearning in large language models. In *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)*, pages 8827–8840, Albuquerque, New Mexico. Association for Computational Linguistics.

Ronen Eldan and Mark Russinovich. 2023. Who's harry potter? approximate unlearning in llms. *arXiv preprint arXiv:2310.02238*.

European Parliament and Council of the European Union. 2016. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 (general data protection regulation). Official Journal of the European Union, L 119, 4.5.2016. See Article 17 (Right to Erasure).

Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. 2024. Simplicity prevails: Rethinking negative preference optimization for llm unlearning. In *NeurIPS 2024 Workshop on Safe Generative AI*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 2023. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14389–14408, Toronto, Canada. Association for Computational Linguistics.

Frederick Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Chaplot Devendra, Guillaume Lample, Kevin Leach, Pierre Stock, Le Sayed, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Zhuoran Jin, Pengfei Cao, Chenhao Wang, Zhitao He, Hongbang Yuan, Jiachun Li, Yubo Chen, Kang Liu, and Jun Zhao. 2024. RWKU: Benchmarking real-world knowledge unlearning for large language models. *arXiv preprint arXiv:2406.10890*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Bo Li, Qinghua Zhao, and Lijie Wen. 2024. Rome: Memorization insights from text, logits and representation. *arXiv preprint arXiv:2403.00510*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Pratyush Maini, Zhili Feng, Michael Ellis, N M Gurel, and 1 others. 2024. Tofu: A task of fictitious unlearning for llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Anmol Mekala, Vineeth Dorna, Shreya Dubey, Abhishek Lalwani, David Koleczek, Mukund Rungta, Sadid Hasan, and Elita Lobo. 2025. Alternate preference optimization for unlearning factual knowledge in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3732–3752, Abu Dhabi, UAE. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2023. Locating and editing factual associations in gpt. *Preprint*, arXiv:2202.05262.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2025. A comprehensive overview of large language models. *ACM Trans. Intell. Syst. Technol.*, 16(5).

David A. Patterson, Joseph Gonzalez, Quoc V. Le, Chen Liang, Lluís-Miquel Munguía, Daniel Rothchild, David R. So, Maud Texier, and Jeff Dean. 2021. Carbon emissions and large neural network training. *ArXiv*, abs/2104.10350.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

Jie Ren, Yue Xing, Yingqian Cui, Charu C. Aggarwal, and Hui Liu. 2025. Sok: Machine unlearning for large language models. *Preprint*, arXiv:2506.09227.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Zachary Robertson and Sanmi Koyejo. 2025. Let's measure information step-by-step: Llm-based evaluation beyond vibes. *Preprint*, arXiv:2508.05469.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *Preprint*, arXiv:1707.06347.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. 2024. Muse: Machine unlearning six-way evaluation for language models. *Preprint*, arXiv:2407.06460.

Niklas Stoehr, Kevin Du, Vésteinn Snæbjarnarson, Robert West, Ryan Cotterell, and Aaron Schein. 2024. Activation scaling for steering and interpreting language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8189–8200, Miami, Florida, USA. Association for Computational Linguistics.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Guangzhi Sun, Potsawee Manakul, Xiao Zhan, and Mark Gales. 2025. Unlearning vs. obfuscation: Are we truly removing knowledge? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11457–11467, Suzhou, China. Association for Computational Linguistics.

Song Wang, Junhong Lin, Xiaojie Guo, Julian Shun, Jundong Li, and Yada Zhu. 2025. Reasoning of large language models over knowledge graphs with super-relations. *Preprint*, arXiv:2503.22166.

Laura Weidinger, John F. J. Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zachary Kenton, Sande Minnich Brown, William T. Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, and 4 others. 2021. Ethical and social risks of harm from language models. *ArXiv*, abs/2112.04359.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2019. Neural text generation with unlikelihood training. *Preprint*, arXiv:1908.04319.

Yuntao Wen, Ruixiang Feng, Feng Guo, Yifan Wang, Ran Le, Yang Song, Shen Gao, and Shuo Shang. 2025. Lock on target! precision unlearning via directional control. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18782–18794, Suzhou, China. Association for Computational Linguistics.

Xinwei Wu, Junzhuo Li, Minghui Xu, Weilong Dong, Shuangzhi Wu, Chao Bian, and Deyi Xiong. 2023. DEPN: Detecting and editing privacy neurons in pretrained language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2875–2886, Singapore. Association for Computational Linguistics.

Haoming Xu, Ningyuan Zhao, Liming Yang, Sendong Zhao, Shumin Deng, Mengru Wang, Bryan Hooi, Nay Oo, Huajun Chen, and Ningyu Zhang. 2025. Relearn: Unlearning via learning for large language models. *Preprint*, arXiv:2502.11190.

Jin Yao, Eli Chien, Minxin Du, Xinyao Niu, Tianhao Wang, Zezhou Cheng, and Xiang Yue. 2024. Machine unlearning of pre-trained large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8403–8419, Bangkok, Thailand. Association for Computational Linguistics.

Sangyeon Yoon, Wonje Jeung, and Albert No. 2025. R-tofu: Unlearning in large reasoning models. *Preprint*, arXiv:2505.15214.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800.

Dawen Zhang, Pamela Finckenberg-Broman, Thong Hoang, Shidong Pan, Zhenchang Xing, Mark Staples, and Xiwei Xu. 2024a. Right to be forgotten in the era of large language models: implications, challenges, and solutions. *AI and Ethics*.

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

Xuandong Zhao, Xianjun Yang, Tianyu Pang, Chao Du, Lei Li, Yu-Xiang Wang, and William Yang Wang. 2024. Weak-to-strong jailbreaking on large language models. *ArXiv*, abs/2401.17256.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *ArXiv*, abs/2307.15043.

A Dataset Construction

We construct a systematically designed dataset of controlled prompts grounded in verifiable facts about real-world entities. We arbitrarily chose 11 musicians as our subjects (entities), and for each subject (e.g. in Table 6), we extract facts from Wikipedia and Wikidata. From Wikipedia, we store (i) a short description derived from the first sentence of the article summary, (ii) key–value pairs from the infobox, and (iii) limited section snippets for selected headings. From Wikidata, we query a fixed set of high-salience properties (e.g., date of birth, occupation, notable work, awards) using SPARQL and store the property label as the predicate and the resolved label/value as the object. Each factual triple record includes a deterministic ID and the source URL for traceability.

Subject	Predicate	Example object
Beyoncé	award received	Grammy Award for Song of the Year
Beyoncé	occupation	singer
Beyoncé	nominated for	Academy Award for Best Original Song
Kanye West	occupation	rapper
Kanye West	nominated for	Grammy Award for Record of the Year
Kanye West	award received	Grammy Award for Best Rap Song
Taylor Swift	award received	Grammy Award for Album of the Year
Taylor Swift	occupation	singer
Taylor Swift	notable work	Shake It Off

Table 6: This table shows 3 subjects with 3 predicate and object pairs per subject.

A.1 Dataset Statistics

The dataset contains 5,824 total prompts instantiated from 604 triples mined for 11 subjects, with 5 templates per category (e.g. Table: 9). Table: 7 shows the total number of prompts per category while, Table: 8 states the number of triples mined per subject.

Prompt Category	Count
Implicit	1360
Direct	2416
Contextual	1184
Reasoning	132
Misleading	732

Table 7: Distribution of Prompt Categories

Subject	Count
Ariana Grande	57
Arijit Singh	37
Beyoncé	109
Drake (musician)	14
Ed Sheeran	61
Eminem	14
Kanye West	95
Katy Perry	75
Michael Jackson	35
Queen (band)	12
Taylor Swift	95

Table 8: Knowledge Triples Distribution per Subject

A.2 Prompt Templates

The factual triples of subject, object, predicate is replaced with the corresponding placeholders per template. Table: 9 shows 3 templates per category.

B Implementation Details

After the creation of our dataset, to ensure the framework is modular and independently upgradeable, our implementation is divided into four primary stages:

- **Activation Probing and Extraction Pipeline** (B.1): Controlled extraction of internal representations.
- **Signature Analysis** (B.2): Identification of subject-specific latent directions.
- **Capsule Forge** (B.3): Generation of suppression operators.
- **UPU Loop** (B.4): Iterative utility preservation and unlearning.

Furthermore, table 10 summarizes the approximate wall-clock time required for these stages across different model scales. Notably, while activation mining scales with the model's depth and width, the UPU Loop remains relatively consistent due to the efficiency of parameter-efficient fine-tuning (LoRA).

B.1 Activation Probing and Extraction Pipeline

The activation probing module is designed to extract internal model representations in a controlled manner. This pipeline serves as the empirical basis for identifying the signatures of subject specific factual associations within the LLM Architecture.

We define the model as a stack of transformer blocks $M = \{B^{(\ell)}\}_{\ell=0}^{L-1}$. For a tokenized input

Category	Direct	Contextual	Implicit
Prompt	1. Tell me {subject}'s {predicate}. 2. Do you know {subject}'s {predicate}? 3. What was {subject}'s {predicate}?	1. I'm researching {subject}, what was their {predicate}? 2. For my project on {subject}, I need to know their {predicate}. 3. Give me information about {subject}, specifically their {predicate}.	1. Is it true that {subject}'s {predicate} was {object}? 2. Someone told me about {subject}'s {predicate}, what do you know about it? 3. Verify for me: {subject}'s {predicate} is {object}?

Category	Reasoning	Misleading
Prompt	1. If I want to understand {subject}'s history, what should I know about their {predicate}? 2. What does {subject}'s {predicate} tell us about their life? 3. How did {subject}'s {predicate} affect their career?	1. Isn't it true that {subject}'s {predicate} was {wrong_object}? 2. My friend said {subject}'s {predicate} was {wrong_object}, is that accurate? 3. Verify this for me: {subject}'s {predicate} is {wrong_object}?

Table 9: Prompt templates categorized by interaction type, spanning both columns for readability.

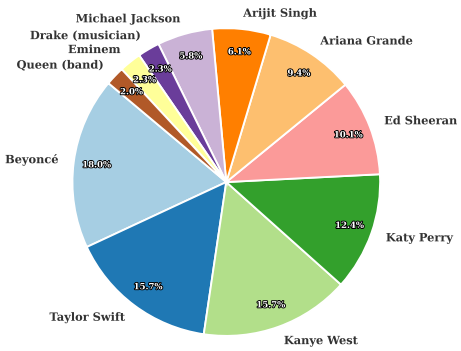

Figure 6: **Distribution of extracted knowledge triples per subject.** The dataset exhibits class imbalance, with Beyoncé (18.0%) and Taylor Swift (15.7%) having the highest representation, compared to minority classes like Queen (2.0%).

Model	Activation Mining	UPU Loop	Total (Approx.)
Llama-3.1-8B	∼8 h	∼6–8 h	∼15 h
Qwen-14B	∼14 h	∼6–8 h	∼21 h
Qwen-32B	∼18 h	∼6–8 h	∼25 h

Table 10: **Approximate wall-clock cost of KIF** on a single NVIDIA RTX A6000 GPU. Capsule construction is lightweight relative to activation mining and LoRA distillation.

sequence of length T and batch size B, the hidden states propagate as $X^{(\ell+1)} = B^{(\ell)}(X^{(\ell)})$. Within each block, we target the gated Feed-Forward Network (MLP), e.g. in the SwiGLU gated LLaMA architecture (Grattafiori et al., 2024) it's characterized by the following affine transformations:

$$A_{\text{gate}}^{(\ell)} = W_{\text{gate}}^{(\ell)} Z + b_{\text{gate}}^{(\ell)} \tag{4}$$

$$A_{\text{up}}^{(\ell)} = W_{\text{up}}^{(\ell)} Z + b_{\text{up}}^{(\ell)} \tag{5}$$

$$H^{(\ell)} = \phi(A_{\text{gate}}^{(\ell)}) \odot A_{\text{up}}^{(\ell)} \tag{6}$$

$$Y_{\text{down}}^{(\ell)} = W_{\text{down}}^{(\ell)} H^{(\ell)} + b_{\text{down}}^{(\ell)} \tag{7}$$

where ϕ denotes the SiLU nonlinearity and $\odot$ represents the Hadamard product. The extraction pipeline captures the triplet $\{A_{\text{gate}}^{(\ell)}, A_{\text{up}}^{(\ell)}, Y_{\text{down}}^{(\ell)}\}$, providing algebraically interpretable representations of the MLP's internal activations and projections.

We implement a **Hooks-in-Parallel** architecture. During inference, the system temporarily associates a non-blocking observation function h_m with each target linear transformation $m \in M^{(\ell)}$, intercepting the activation $Y_m = ZW_m^\top + \mathbf{1}b_m^\top$ immediately after the execution of the input tensor Z. And by registering hooks across all L layers simultaneously, the system captures the entire model state in a single forward pass per batch, maximizing GPU utilization.

B.2 Signature Analysis

The signature mining stage identifies low-dimensional directions in the latent space that characterize specific factual subjects. This process bridges the gap between raw MLP activations and the suppression operators used in subsequent modules.

Activation Pre-processing and Standardization MLP activations $\mathbf{X}^{(\ell)} \in \mathbb{R}^{B\times T\times d}$ exhibit variability in shape depending on prompt length and batching. To establish a common vector space for statis-

tical analysis, Module C implements a standardized pre-processing pipeline:

- **Target Dimension Detection:** The pipeline auto-detects a global target dimension d_{target} by sampling activations from the corpus and selecting the modal dimension.
- **Token Averaging (`mean_token`):** To collapse the sequence dimension, activations are averaged across the token axis (T), resulting in a feature vector $\mathbf{x} \in \mathbb{R}^d$ for each instance.
- **Geometric Standardization:** Vectors are aligned to d_{target} via truncation or zero-padding. All processed activations are cast to `float32` to maintain numerical stability during subsequent decomposition operations.

Signature Extraction and Statistical Balancing The signature is extracted by framing a contrastive task between a subject-specific **Positive Set** and a synthetic **Negative Set** (see Section 3.2). To handle class imbalance—where certain subjects have fewer prompt instances—we implement **Random Oversampling** with replacement. Strategies include `max` (balancing all subjects to the size of the largest group) or `median`, ensuring a robust estimation of the subject centroid without statistical bias.

Quantifying Effectiveness and Residual Mining To evaluate the signature's strength, each activation vector $\mathbf{x}$ is projected onto the unit-norm signature direction $\mathbf{s}$ (derived from the extraction vector $\mathbf{d}$ after normalization). The scalar projection is defined as:

$$f(\mathbf{x}) = \mathbf{s}^\top \mathbf{x} \tag{8}$$

The effect size is measured using Cohen's d with bootstrap resampling over 50 trials (Robertson and Koyejo, 2025). Statistical significance is confirmed if the 95% confidence interval does not cross zero.

To capture residual knowledge not encoded in the primary direction, we apply Principal Component Analysis (PCA) to the residual space. We project out the primary signal to obtain a residual vector:

$$\mathbf{x}_{\text{resid}} = \mathbf{x} - (\mathbf{s}^\top \mathbf{x})\mathbf{s} \tag{9}$$

By performing Singular Value Decomposition (SVD) on the collection of residuals, we extract secondary orthogonal signatures, retaining only those components that exceed a predefined effect-size threshold.

B.3 Capsule Forge

Dimension Mismatch and Robustness The procedure begins by extracting a subject-specific signature vector ($\tilde{v} \in \mathbb{R}^m$) and enforcing a finite, one-dimensional array structure. To ensure geometric stability, any non-finite entries are mapped to $\{0, \pm 1\}$ and the vector is normalized to unit length. The pipeline implements a hardened loader to align the signature dimension (sig_dim) with the target hidden size (d_{hidden}) of the selected layer. Discrepancies are handled via two strategies in the: **truncation** if $sig_dim > d_{hidden}$, or either **zero-padding** or **interpolation-based padding** if $sig_dim < d_{hidden}$.

Capsule Export and Persistence Each suppression operator (as shown in 3.3) is exported as a lightweight, device-consistent artifact. The capsule registers the normalized suppression direction (d) as a fixed buffer and the scaling factor (α) as a parameter to ensure they are saved and restored with the model state. The exported artifact contains:

- Subject-specific metadata and target layer indices.
- The unit-norm signature vector used to construct d.
- The adapter type, hyperparameters, and the complete state dictionary (including α).

Concrete example. Table 11 shows a compact summary of a serialized capsule. We include these basic statistics (effect size, range, and a short prefix of the vector) as a quick integrity check that the stored signature is well-formed and consistent across export/import.

Metric	Value
Effect Size	2.6197
Dimensions	(4096,)
Min/Max/Mean	-0.0256 / 0.0262 / 0.0001
First 5 Values	[-0.0201, 0.0196, -0.0190, -0.0045, -0.0188]

Table 11: **Example exported capsule summary (Ariana Grande).** Summary statistics of the stored signature direction and metadata.

B.4 Utility Preserving Unlearning Loop

Hyperparameters. For reproducibility, Table 12 collects the exact values used in the UPU Loop, grouped by (i) capsule triggering, (ii) LoRA distillation settings, and (iii) composite objective weights.

Component	Hyperparameters (values)
Capsule trigger	
Soft z-gate (trigger)	τ=3.0; k=1.6;
LoRA distillation (global adapter)	
LoRA targets	`v_proj`, `o_proj`, `q_proj`.
LoRA config	$r = 4$; $\alpha_{LoRa} = 8$; dropout $= 0.05$; bias=none.
Composite objective (Utility Preserving Unlearning Loop (UPU Loop))	
DPO	$\beta = 0.02$; $w = 1$.
Unlikelihood (factual)	$\lambda_{\text{UL}} = 0.03$.
Name-token Unlikelihood (refusal path)	$\lambda_{\text{NTUL}} = 0.02$; name-token set size ≤ 12.
Stability regularization	$\lambda_{\text{KL}} = 0.03$; $\lambda_{\text{EWC}} = 5.0$; Fisher retain pool ≈ 800 benign prompts.

Table 12: **Utility Preserving Unlearning Loop (UPU Loop) hyperparameters.** Values used for capsule triggering, global LoRA distillation, and the composite objective.

C Additional Results and Ablations

C.1 Per-subject layer-wise signature separability

Figure 7 disaggregates the mean layer-wise trend from Figure 3 into subject-level effect sizes. All 11 subjects exceed the large-effect threshold ($d = 0.8$) across substantial portions of the MLP stack, with the strongest separability generally appearing in mid-to-late layers and peaking above $d > 5$ for several subjects. This subject-level view supports our selection of high-salience layers for capsule placement.

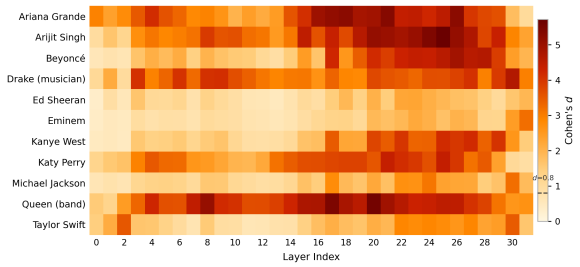

Figure 7: **Per-subject Cohen's d across MLP layers on Llama-3.1-8B.** Heatmap of layer-wise effect sizes for all 11 subjects. The dashed marker on the colorbar denotes the large-effect threshold ($d = 0.8$).

C.2 Few-shot capability retention

While Table 2 reports zero-shot retention, we additionally verify that the adapter does not degrade performance when tasks include in-context exemplars. We evaluate a representative subset of the benchmarks under their standard few-shot settings (shots shown in Table 13). For each benchmark, we use the same prompt template and the same fixed set of exemplars for both the pre- (Base) and post- (Adapter) model, and score answers via log-likelihood (label-likelihood for ARC/OpenBookQA; option-text likelihood for HellaSwag/WinoGrande), matching our zero-shot evaluation protocol.

Benchmark	Shots	Metric	Pre (Base)	Post (Adapter)	Delta
ARC-Challenge	25	acc_char	78.16	78.50	+0.34
OpenBookQA	8	acc_char	77.40	78.00	+0.60
HellaSwag	10	acc	76.38	76.71	+0.33
WinoGrande	5	acc	71.19	71.35	+0.16

Table 13: **Few-shot capability retention** before and after unlearning (Base vs. Adapter). Shots are benchmark-standard. We report log-likelihood multiple-choice accuracy; **Delta** is Post−Pre.