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

Anonymous ACL submission

Abstract

Entity-level unlearning is usually evaluated by what a model says: whether it stops naming the target, refuses a query, or shifts a Truth Ratio distribution. These output-level tests, however, do not show whether a subject's internal representation has been attenuated. We introduce the **Entity Representation Unlearning Framework (ERUF)**, a representation-aware framework that mines subject-specific activation signatures, suppresses the corresponding activation direction, and distills the behavior into LoRA parameters. Among evaluated baselines, ERUF is the only method that jointly achieves surface level suppression, internal attenuation, and utility preservation. On TOFU `forget10`, ERUF achieves FQ $= 0.99$ and MU $= 0.62$, matching reported oracle utility while approaching oracle forget quality. Across most standard foundation-model settings, ERUF maintains low leakage and low internal target activation (SMR $= 0.00$–1.10%, EL10 < 0.06, utility drift $< 3\%$). On Llama-3.1-8B, adversarial entity recovery falls from 63.89% to 20.15%, while name-agnostic recovery decreases by 72.7–77.4%. Joint surface/internal diagnostics further reveal scale-dependent behavior in reasoning-prior models that surface metrics alone would miss. We interpret these results as operational evidence of representation-level attenuation, not as a formal guarantee of irreversible deletion. Code is available in the anonymous repository[1].

1 Introduction

Large language models (LLMs) are increasingly deployed in real-world NLP systems, but their training corpora can contain sensitive, copyrighted, or otherwise removable information, creating tension with the GDPR right to erasure (European Parliament and Council of the European Union, 2016; Zhang et al., 2024a). LLMs store training information in distributed parameters (Naveed et al., 2025), making targeted removal technically difficult. *LLM unlearning* aims to remove selected information while preserving general utility (Yao et al., 2024; Ren et al., 2025). Entity-level unlearning can be especially difficult because the forget target is not a single answer string or isolated fact. A named subject can be recovered through associated facts, aliases, descriptions, relations, and contextual clues (Ma et al., 2025; Jin et al., 2024). Thus, a model may suppress direct mentions while preserving an internal trace of the target that resurfaces through indirect access routes. Prior work warns that such apparent forgetting can reflect obfuscation rather than latent trace removal (Sun et al., 2025; Ren et al., 2025). This motivates treating entity-level unlearning as a representation-level task as much as an output-level one.

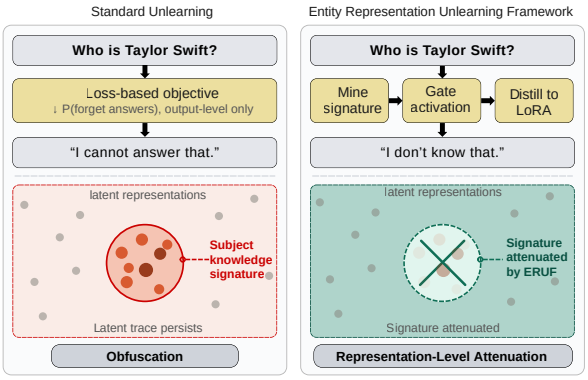

Figure 1: **Comparison of approaches.** Output-level unlearning can suppress the answer while leaving a latent subject trace. ERUF targets a mined subject signature and distills the suppression into a LoRA adapter.

We propose the **Entity Representation Unlearning Framework (ERUF)**, a representation-aware framework for entity-level unlearning. ERUF mines subject-linked activation signatures from real-world entity prompts, suppresses the corresponding activation direction through Signature Suppression Capsules, and distills the re-

[1] Anonymous repository

sulting behavior into a global LoRA adapter (Hu et al., 2021). The target behavioral profile is **surface-level suppression, internal attenuation, and utility preservation**.

We evaluate this profile using Subject Mention Rate (SMR) for surface leakage, and EL10 for early subject-associated token probability mass relative to the base model. We treat EL10 as a proxy for latent target activation rather than proof of deletion, and corroborate it with adversarial, name-agnostic, hidden-state, and sequential diagnostics. Among evaluated baselines, no method simultaneously achieves low surface leakage, reduced EL10, and low utility drift, whereas ERUF achieves this three-way profile on most standard foundation models while the diagnostic explicitly identifies where it does not. The behavioral difference is clearest under probes that omit or obscure the canonical subject name. On Llama-3.1-8B, ERUF reduces adversarial recovery from 63.89% to 20.15%, and reduces name-agnostic alias, keyword, and target-mass metrics by 72.7–77.4%. These results, together with hidden-state, sequential, and cross-family analyses, support ERUF as operational evidence of representation-level attenuation rather than merely surface refusal.

Our contributions are:

1. A real-world entity prompt dataset for mining subject-linked activation signatures;
2. ERUF, a representation-aware framework combining capsule-based suppression with LoRA distillation for entity-level unlearning;
3. A dual-metric surface/internal evaluation protocol (SMR + EL10) distinguishing output suppression from latent target attenuation;
4. An empirical behavioral analysis across evaluated baselines, model families, adversarial and name-agnostic probes, hidden-state diagnostics, and sequential unlearning.

2 Related Work

Benchmark-driven LLM unlearning. Standard LLM unlearning avoids full retraining by optimizing approximate forgetting objectives while preserving retained behavior, typically on benchmarked forget/retain tasks such as TOFU (Maini et al., 2024) and MUSE (Shi et al., 2024). Optimization variants including 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) are largely compared through benchmark behavior rather than direct tests of subject-linked internal activation. Recent audits argue that LLM unlearning evaluation should be more robust and multifaceted, since benchmark success can overstate practical forgetting under query variations and does not by itself show whether information has been removed from model weights (Lynch et al., 2024; Thaker et al., 2025; Deeb and Roger, 2025).

Entity-level unlearning. Entity-level work studies removal of broader subject-level knowledge rather than isolated instances. Ma et al. (Ma et al., 2025) formalize entity-level unlearning and show that forget-set coverage strongly affects success. RWKU (Jin et al., 2024) benchmarks real-world entities with adversarial and locality probes, while OPT-OUT (Choi et al., 2025) studies real-world entity removal with explicit retention of neighboring and world knowledge. ERUF shares this real-world entity setting, but focuses on whether suppressing subject outputs is accompanied by attenuation of subject-associated internal activation.

Representation-level intervention and auditing. Knowledge localization and representation-geometry work suggests that factual associations are mediated by identifiable MLP/FFN computations and directions in representation space (Meng et al., 2023; Geva et al., 2021; Park et al., 2024). This motivates representation-level unlearning methods such as RMU (Li et al., 2024), LUNAR (Shen et al., 2025), and ReGLU (Xiao et al., 2026). Recent work also distinguishes unlearning from obfuscation and asks whether information has been removed from accessible weights or latent representations, rather than merely suppressed at the output surface (Sun et al., 2025; Deeb and Roger, 2025). ERUF follows this line by pairing SMR and EL10 with adversarial and name-agnostic probes, operationalizing the distinction between surface suppression and internal attenuation for entity-level targets.

3 Methodology

We introduce a three-stage pipeline: **(1)** localizing target knowledge through statistical analysis on extracted activation-signature , **(2)** suppressing the localized signature through lightweight gating modules we term **Signature Suppression Capsules**, and **(3)** distilling the capsule behavior into a global LoRA adapter (Hu et al., 2021). The training objective combines preference-style su-

pervision with token-level unlikelihood and stability regularization (Rafailov et al., 2024; Welleck et al., 2019; Kirkpatrick et al., 2017). Figure 2 summarizes the pipeline. All implementation details are provided in Appendix B.

3.1 Dataset Construction

The localization stage requires activation profiling over controlled subject prompts. We construct a Real-World Entity Dataset grounded in factual triples $(\textit{subject}, \textit{predicate}, \textit{object})$ extracted from Wikipedia and Wikidata. Following the knowledge-tuple framing of Meng et al. (2023), we instantiate triples into templates such as "Is it true that `{subject}`'s `{predicate}` was `{object}`?" The dataset contains five probe types:

1. **Direct:** asks for a specific fact.
2. **Contextual:** embeds the target fact in a broader information request.
3. **Implicit:** uses confirmation-style prompts involving the target fact.
4. **Reasoning:** asks the model to explain or reason about the fact.
5. **Misleading:** pairs the subject with incorrect information to probe whether the model verifies or accepts a suggested falsehood.

Full dataset statistics, prompt templates, and subject distributions are provided in Appendix A.

3.2 Localization via Activation-Signature Extraction

Prior causal tracing work suggests that factual associations can be localized in MLP modules (Meng et al., 2023). We therefore probe a 4-bit quantized model on the Real-World Entity Dataset and capture intermediate MLP activation tensors, e.g., $A^{(\ell)}_{\text{gate}}$, $A^{(\ell)}_{\text{up}}$, and $Y^{(\ell)}_{\text{down}}$ for Llama-family models (Grattafiori et al., 2024).

We extract subject-specific signatures through contrastive analysis (Rimsky et al., 2024) over token-averaged, standardized activations. Positives come from on-topic subject prompts and negatives from real controls plus prompts from other dataset subjects. These cross-subject negatives are genuine activations for semantically related entity queries, forcing the mined direction to distinguish the target subject from other natural entity representations. Since signatures are estimated per target subject, the key coverage criterion is per-subject prompt diversity and contrast quality, not total subject count. If the real negative pool is insufficient, we add a bounded synthetic backfill capped at 10% of the negative pool, used only for robustness rather than as the primary contrastive distribution.

The primary signature direction is the standardized mean difference: $\mathbf{d} = \text{mean}(S^z_{\textbf{pos}}) - \text{mean}(S^z_{\textbf{neg}})$, where S^z_{pos} and S^z_{neg} denote positive and negative activation sets after standardization. The vector captures the dominant displacement induced by subject-conditioned prompts relative to the matched negative pool. We validate localization by measuring layer-wise Cohen's d (Cohen, 2013) per subject of the dataset; Figure 3 shows strong separability in mid-to-late layers. We use this as a localization diagnostic for finding suitable candidates for capsule placement in Section 3.3.

3.3 Capsule-based Suppression

Next, we use the **Signature Suppression Capsule** (Figure 4): a lightweight geometric adapter inspired by activation modification (Stoehr et al., 2024; Wen et al., 2025) and rank-one editing (Meng et al., 2023), to suppress targeted representations at inference while preserving unrelated knowledge and model performance.

For hidden state h and unit signature vector v, the capsule decomposes h into $h_{\parallel}$ and $h_{\perp}$. It scales $h_{\parallel}$ through a dynamic attenuation factor α_{eff} while preserving $h_{\perp}$, thereby targeting the subject-linked direction without intentionally altering orthogonal information.

Gated Residual Stream Intervention. The attenuation multiplier is $\alpha_{\text{eff}} = \alpha \cdot \sigma(k(z - \tau))$, with α initialized to -1. The gate triggers when the state's projection onto the signature, expressed as a z-score, exceeds threshold τ under gain k. This mechanism limits intervention to states with statistically elevated alignment to the subject signature. The capsule is implemented through forward hooks with v registered as a fixed buffer. It can be injected dynamically during the UPU Loop (Section 3.4), or removed once unlearning is complete.

3.4 Distillation through the Utility Preserving Unlearning Loop

Capsules provide immediate suppression, but they are not themselves a permanent parameter update. We therefore distill capsule behavior into a global LoRA adapter through a **Utility Preserving Unlearning Loop (UPU Loop)**. Whenever

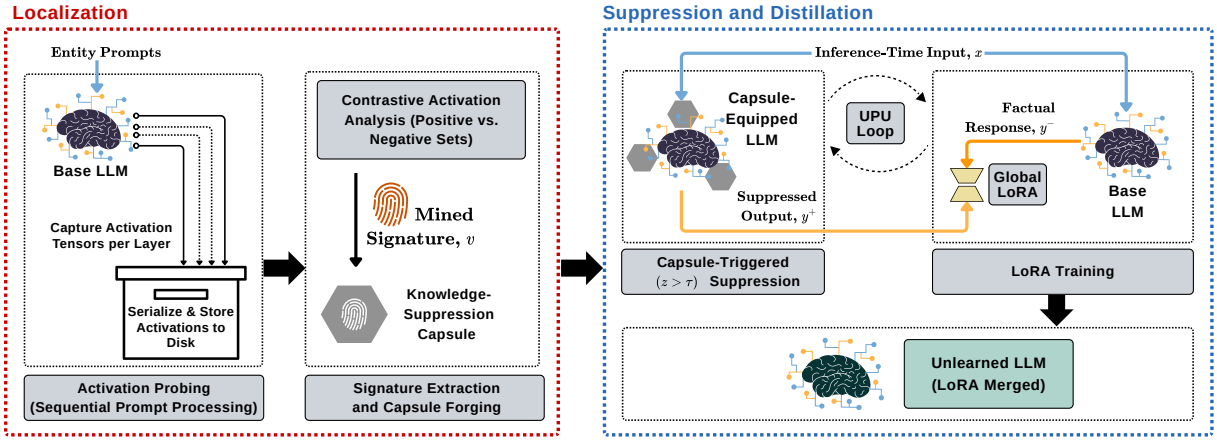

Figure 2: **Entity Representation Unlearning Framework (ERUF) pipeline.** Entity prompts from the Real-World Entity Dataset (3.1) probe MLP activations to mine subject-specific signatures via contrastive analysis. These signatures instantiate Signature Suppression Capsules (Section 3.3) at high-salience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

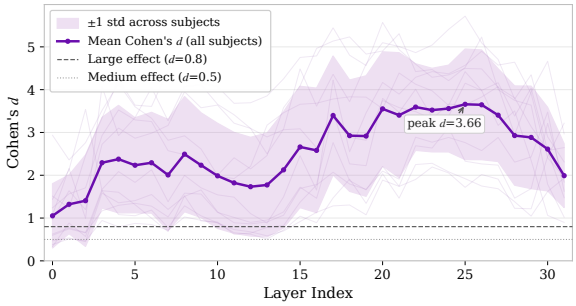

Figure 3: **Mean Cohen's d across transformer layers on Llama-3.1-8B.** Subject-specific activation separability is strongest in mid-to-late layers, motivating capsule placement in high-salience layers.

the capsule gate triggers at layer ℓ, we log a tuple (x, y^+, y^-) containing the prompt, the capsule-suppressed response, and the base model's factual output. LoRA parameters ϕ are optimized with the following composite objective $\mathcal{L}$ (Eq. 3):

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

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

$$\mathcal{L} = \mathcal{L}_{\text{DPO}} + \lambda_{\text{UL}}\mathcal{L}_{\text{UL}} + \lambda_{\text{NTUL}}\mathcal{L}_{\text{NTUL}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} + \lambda_{\text{EWC}}\mathcal{L}_{\text{EWC}}. \tag{3}$$

Equation 1 applies Direct Preference Optimization (Rafailov et al., 2024) to the logged tuples, with penalty scaling factor w anchoring updates to the reference model. Factual Unlikelihood $\mathcal{L}_{\text{UL}}$ penalizes the base factual output y^-. Name-Token Unlikelihood $\mathcal{L}_{\text{NTUL}}$ penalizes aggregate probability mass over subject-name tokens V_{name} to reduce soft leakage. Stability is encouraged with KL regularization ($\mathcal{L}_{KL}$) (Schulman et al., 2017; Rafailov et al., 2024) and Elastic Weight Consolidation ($\mathcal{L}_{EWC}$) (Kirkpatrick et al., 2017) on benign anchor prompts, penalizing shifts in parameters identified as critical.

4 Experiments and Results

We evaluate ERUF along two axes. The experiments on Real-World Entity Dataset are mechanistic analysis of entity-level unlearning through surface leakage, internal target activation, adversarial recovery, hidden-state drift, and sequential stability. TOFU `forget10` (Maini et al., 2024) provides standardized comparison against prior unlearning baselines.

Model and baseline selection. We evaluate standard foundation models (Llama (Grattafiori et al., 2024), Mistral (Jiang et al., 2023)) and reasoning-prior models (Qwen (Yang et al., 2025), DeepSeek (Bi et al., 2024)) from 3B to 32B parameters. For TOFU, we compare against Gradient Ascent (Jang et al., 2023), GradDiff (Maini et al., 2024), IDK/refusal (Maini et al., 2024), NPO (Zhang et al., 2024b), and SimNPO (Fan et al., 2024). For entity-level robustness, we also include OPT-OUT where available. Baseline TOFU numbers are from Fan et al. (2024). Implementation details are provided in Appendix B.4.

Evaluation metrics. For Real-World Entity experiments, we report: **Utility Drift**, measured as

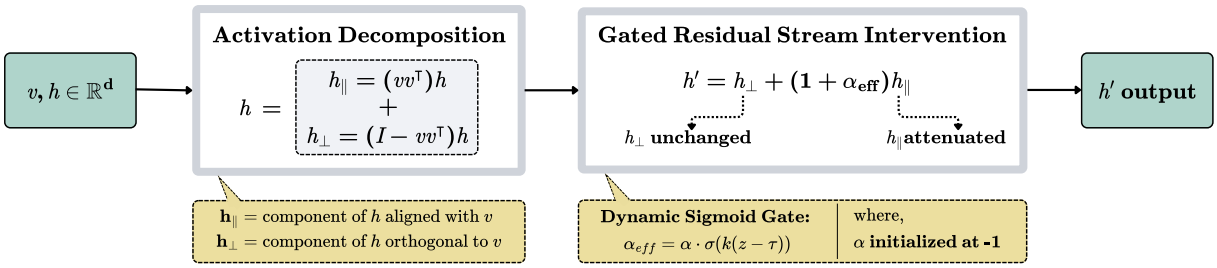

Figure 4: **Capsule-based suppression.** The capsule decomposes an activation into components parallel and orthogonal to the subject signature. A gated suppression factor is applied only to the parallel component before the modified activation is passed to the subsequent layer.

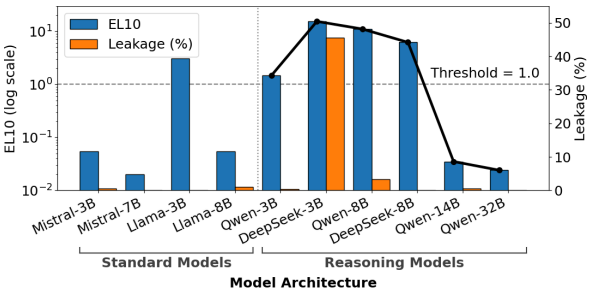

Figure 5: **EL10 and surface leakage across model families and scales.** Standard models mostly maintain low internal target activation, with Llama 3B as an exception. Reasoning-prior models show a capacity-dependent pattern.

relative benign perplexity change; **Subject Mention Rate (SMR)**, the percentage of prompts whose generation contains the target name; and **EL10 ratio**, an early-step subject-associated token probability-mass ratio relative to the base model. EL10 uses up to 10 subject-associated single-token keywords, with canonical subject-name subtokens as backfill, and reports the post/pre probability-mass ratio over the first 10 decoding steps. EL10 is a proxy for latent target activation; it is not a proof of irrecoverability. Following the unlearning-vs.-obfuscation distinction (Sun et al., 2025), we use three descriptive mechanism states with tolerance $\epsilon = 5\%$:

- **Type I: representation-level attenuation.** SMR $\leq \epsilon$ and EL10 < 1.
- **Type II: obfuscation.** SMR $\leq \epsilon$ and EL10 > 1.
- **Type III: instability.** SMR $> \epsilon$.

For TOFU, we report Forget Quality (FQ) and Model Utility (MU) (Maini et al., 2024).

4.1 Mechanism States on the Real-World Entity Dataset

Table 1 summarizes the main behavioral comparison. ERUF is the only evaluated method that jointly achieves low surface leakage, internal attenuation, and utility preservation on the primary standard-model setting, while the same diagnostic exposes exceptions such as Llama 3B. Baselines expose complementary failure modes: LUNAR (Shen et al., 2025), OPT-OUT (Choi et al., 2025), and SimNPO (Fan et al., 2024) retain high surface leakage, while ReGLU suppresses SMR but, like LUNAR, amplifies EL10 and degrades utility. ERUF's claim, however, is not architecture-uniform: Llama 3B remains Type II, with zero SMR but elevated EL10.

Reasoning-prior models exhibit a capacity-dependent pattern. DeepSeek 3B is unstable (Type-III), Qwen/DeepSeek 8B suppress surface leakage but retain high EL10 (Type-II), and only at 14B parameters does Qwen return to Type-I behavior, which remains consistent at 32B, demonstrating that as parameter count increase, unlearning stabilizes. Thus, SMR alone misses a key transition. Figure 5 visualizes why: Qwen 8B and Qwen 14B both have low SMR, but differ by over two orders of magnitude in EL10.

4.2 Adversarial and Name-Agnostic Recovery Diagnostics

Output-level unlearning can fail when the canonical name is removed from the prompt. We therefore evaluate ERUF, the pre-unlearning base model, OPT-OUT, SimNPO, ReGLU, and LUNAR on adversarial and name-agnostic recovery probes using Llama-3.1-8B. Since the comparison systems are not specifically adversarially tuned, we treat them as entity-level baselines rather than state-of-the-art adversarial defenses.

We use adversarial recovery diagnostics as benign entity-unlearning stress tests to assess whether forgotten-subject information remains recoverable through indirect or rephrased queries. Following RWKU's adversarial probes and locality tests (Jin et al., 2024), and motivated by evidence that edited or deleted subject information

Family	Model Size	Util. Drift	SMR (↓)	EL10 (↓)	Mechanism State
Standard	Mistral 7B	$+0.50\% \pm 0.45$	$0.00\% \pm 0.00$	0.020 ± 0.015	Type I
Standard	Llama 8B	$-0.45\% \pm 1.59$	$1.10\% \pm 1.90$	0.052 ± 0.047	Type I
Standard	Mistral 3B	$+0.36\% \pm 2.27$	$0.52\% \pm 0.35$	0.054 ± 0.047	Type I
Reasoning	Qwen 14B	$+1.99\% \pm 0.51$	$0.52\% \pm 0.45$	0.035 ± 0.013	Type I
Reasoning	Qwen 32B	$+1.31\% \pm 0.38$	$0.08\% \pm 0.06$	0.024 ± 0.018	Type I
Reasoning	Qwen 8B	$-0.50\% \pm 0.60$	$3.33\% \pm 1.25$	11.03 ± 2.45	Type II
Reasoning	DeepSeek 8B	$+1.03\% \pm 0.85$	$0.00\% \pm 0.00$	6.19 ± 1.80	Type II
Standard	Llama 3B	$-0.44\% \pm 0.40$	$0.00\% \pm 0.00$	3.06 ± 0.55	Type II
Reasoning	Qwen 3B	$+2.22\% \pm 2.42$	$0.43\% \pm 0.37$	1.460 ± 1.480	Type II
Reasoning	DeepSeek 3B	$+2.15\% \pm 1.74$	$45.6\% \pm 33.6$	15.39 ± 15.54	Type III
Baseline	LUNAR-Llama 8B	$+32.80\%$	56.06%	7.57	Type III
Baseline	ReGLU-Llama 8B	$+51.34\%$	1.52%	13.13	Type II
Baseline	OPT-OUT-Llama 8B	$+1.31\%$	50.00%	1.62	Type III
Baseline	SimNPO-Llama 8B	$+1.23\%$	72.73%	0.82	Type III

Table 1: **Cross-model mechanistic validation.** Results are mean±SD across three seeds where available; baseline rows are reported as single evaluated conditions. ERUF achieves Type-I behavior on most standard models and on Qwen 14B/32B, while exposing Type-II obfuscation in Llama 3B and 8B reasoning-prior models and Type-III instability in small reasoning-prior models. Evaluated baselines show either high leakage (Type III) or elevated EL10 (Type II) for ReGLU, indicating weaker mechanistic removal.

Model	Alias Hit%↓	Keyword Hit%↓	Target Mass↓	Recovery% ↓
PRE	31.95	16.53	4.398×10^{-2}	63.89
OPT-OUT	21.78	13.57	2.874×10^{-2}	54.41
SIMNPO	36.23	21.46	3.999×10^{-2}	74.63
REGLU	9.57	18.28	3.596×10^{-2}	72.12
LUNAR	33.99	16.08	4.096×10^{-2}	70.22
ERUF	**9.54**	**4.63**	1.357×10^{-2}	**20.15**

Table 2: **Overall adversarial forget-recovery results.** Target mass is reported as mean probability mass. ERUF has the lowest recovery rate and target-token mass among the evaluated conditions. Comparison systems are used as entity-level baselines rather than adversarially tuned state-of-the-art methods.

can persist in hidden-state traces or be recovered via rephrased questions (Patil et al., 2023), we evaluate direct-profile, alias-only, descriptor-only, relation-clue, masked-name, instruction-pressure, multi-turn, multilingual, and mixed-query probes. These test whether entity knowledge survives beyond canonical-name suppression, consistent with findings that forget-set coverage strongly influences entity-unlearning success (Ma et al., 2025). Full attack probe definitions are provided in Appendix H.

Tables 2–3 show that ERUF is consistently more robust under adversarial and name-agnostic probes than the evaluated baselines. The strongest evidence comes from probes where the canonical name is absent or indirect, which tests whether forgetting extends beyond exact-name refusal. These results are consistent with attenuation of non-canonical access routes rather than simple surface suppression. However, residual recovery under harder prompt families like multilingual, multi-turn, and instruction-pressure prompts, shows that recover-ability is reduced rather than eliminated.

Table 4 shows that ERUF strongly reduces alias, keyword, and target-mass diagnostics when the canonical subject name is removed, while several baselines remain close to the pre-unlearning model. This supports the interpretation that ERUF weakens non-canonical access routes.

4.3 Hidden-State Selectivity

We examine whether ERUF shifts forgotten-subject representations more than unrelated prompts. Table 5 summarizes hidden-space diagnostics on Llama-3.1-8B. Relative to the pre-unlearning base model, ERUF achieves zero surface leakage, a 95.6% reduction in EL10 mass, and a 30.8% reduction in E_{30} extraction mass. It also produces the largest forget-prompt drift and the highest selective representation shift ratio (SRS), indicating that forget-subject representations move substantially more than benign representations. Baselines either retain target exposure or show weaker forget-vs-benign selectivity. This suggests that ERUF changes forgotten-subject representations more selectively than output-only suppression would imply.

This hidden-state evidence addresses a limitation of output-only evaluation. A refusal-only mechanism could alter generated text without selectively shifting hidden states for forgotten-

Attack Family	Recovery Success Percentage↓						Alias Hit Percentage↓					
	PRE	OPT-OUT	SIMNPO	REGLU	LUNAR	ERUF	PRE	OPT-OUT	SIMNPO	REGLU	LUNAR	ERUF
Alias-only	88.89	73.94	87.88	93.94	69.70	**22.47**	57.78	37.58	55.76	36.97	38.18	**7.87**
Descriptor-only	69.71	52.53	85.86	81.82	71.21	**21.26**	18.75	10.10	25.25	4.04	19.19	**9.66**
Direct profile	77.14	72.73	84.85	54.55	78.79	**12.86**	61.43	54.55	65.15	16.67	56.06	**1.43**
Instruction pressure	69.73	59.39	70.91	59.39	72.73	**34.05**	42.16	35.15	40.61	4.85	49.70	**24.86**
Masked name	40.00	45.45	38.18	63.64	49.09	**21.67**	16.67	7.27	10.91	3.64	21.82	**5.00**
Mixed query	39.75	25.00	65.00	68.18	65.45	**5.04**	15.48	3.18	16.82	6.82	23.18	**0.00**
Multi-turn style	41.85	63.03	51.52	55.76	81.82	**23.50**	23.37	18.79	26.67	4.85	39.39	**14.21**
Multilingual simple	84.46	71.02	93.75	85.23	80.11	**32.81**	38.34	35.23	52.84	4.55	44.89	**11.98**
Relation clue	79.00	55.05	90.91	72.73	68.69	**15.53**	37.90	18.69	45.96	4.55	28.79	**9.13**

Table 3: **Adversarial forget-recovery by attack family.** ERUF has the lowest recovery success in every reported attack family. Alias-hit results show that reduced recovery is not limited to exact-name suppression.

Model	N-agn. alias↓	N-agn. kwd.↓	N-agn. mass↓	Mxd. alias↓	Mxd. mass↓
PRE	0.311	0.179	0.0423	0.0180	0.0142
OPT-OUT	0.318	0.179	0.0418	0.0180	0.0141
SIMNPO	0.353	0.213	0.0374	0.0293	0.0175
REGLU	0.152	0.193	0.0369	0.0313	0.0138
LUNAR	0.267	0.141	0.0332	0.0703	0.0218
ERUF	**0.070**	**0.0469**	**0.0115**	**0.00719**	**0.00469**
ERUF red.	*77.4%*	*73.8%*	*72.7%*	*60.0%*	*66.9%*

Table 4: **Name-agnostic and mixed-query forget robustness.** N-agn. = name-agnostic; Mxd. = mixed-query; kwd. = keyword hit; mass = target mass. All results use the same 750 evaluated row identifiers. ERUF reduction (last row) is relative to Pre.

Model	SMR ↓	EL10 mass ↓	D_ℓ^F ↑	D_ℓ^B ↓	SRS ↑	E_{30} ↓
Pre base	6.67×10^{-1}	1.92×10^{-3}	–	–	–	1.26×10^{-4}
OPT-OUT	4.85×10^{-1}	4.59×10^{-2}	7.00×10^{-4}	$\mathbf{5.00\times10^{-4}}$	1.30×10^{0}	3.10×10^{-2}
LUNAR	7.27×10^{-1}	6.44×10^{-2}	0.00×10^{0}	0.00×10^{0}	0.00×10^{0}	6.36×10^{-2}
REGLU	1.06×10^{-1}	1.01×10^{-2}	3.40×10^{-3}	3.60×10^{-3}	9.00×10^{-1}	4.73×10^{-3}
SIMNPO	7.73×10^{-1}	4.07×10^{-2}	0.00×10^{0}	0.00×10^{0}	1.30×10^{0}	4.00×10^{-2}
ERUF	$\mathbf{0.00\times10^{0}}$	$\mathbf{8.38\times10^{-5}}$	$\mathbf{4.25\times10^{-2}}$	7.17×10^{-3}	$\mathbf{5.90\times10^{0}}$	$\mathbf{8.72\times10^{-5}}$

Table 5: **Hidden-space diagnostics on Llama-3.1-8B.** D_ℓ^F and D_ℓ^B denote hidden-state drift at layer ℓ on forget prompts and generic benign prompts, respectively, measured against the pre-unlearning base model. SRS is the selective representation shift ratio $D_\ell^F/(D_\ell^B+\epsilon)$, computed from unrounded drift values; printed zero drift values indicate values below display precision. E_{30} denotes subject-associated token extraction mass over a 30-step autoregressive rollout.

subject prompts. ERUF does shift benign representations, but its forget-prompt drift is 5.9× larger than its benign drift, suggesting target-directed rather than global representation change.

4.4 Standardized and Auxiliary Evaluations

TOFU Benchmark Results Table 6 reports the standardized TOFU `forget10` result. ERUF approaches oracle-level forgetting while preserving aggregate utility at the original-model level. Compared with prior baselines, this places ERUF on a more favorable forgetting–utility trade-off. This result is computed on the TOFU-origin LLaMA-2-7B-Chat checkpoint for comparability with prior

Metric	Oracle	Orig.	GA	GD	IDK	NPO	Sim	ERUF
FQ ↑	1.00	0.00	2.2×10^{-16}	3.7×10^{-15}	2.9×10^{-14}	0.29	0.45	**0.99**
MU ↑	0.62	0.62	0.00	0.54	0.54	0.55	0.62	**0.62**

Table 6: **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). ERUF achieves near-oracle FQ while matching original/oracle MU. Abbreviations: GA = Gradient Ascent, GD = GradDiff, Sim = SimNPO.

Subject	FB (↓)		QA (↓)		AA (↓)		FM Loss (↑)	
	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 7: **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 ERUF. Lower FB/QA/AA and higher FM loss indicate reduced recover-ability.

work; the mechanistic entity experiments use Llama-3.1-8B and other modern models.

RWKU Evaluation Results To test whether mined signatures generalize beyond our custom prompt templates, we conduct an evaluation-only probe on the four subjects shared between our dataset and RWKU (Jin et al., 2024) on LLama 3.1 8B model. Table 7 shows consistent post-unlearning reductions across FB, QA, and adversarial-attack recall, along with an increase in FM loss. These results support transfer beyond the exact training prompts. Because the evaluation uses only the four subjects shared between our dataset and RWKU, it should be interpreted as a restricted generalization probe rather than a full RWKU leaderboard comparison.

General Capability Retention We evaluate zero-shot performance on ARC-Challenge (Clark et al., 2018), OpenBookQA (Mihaylov et al., 2018), BoolQ (Clark et al., 2019), Hel-

Group	n	SMR_{Pre}	SMR_{S1}	SMR_{S2}	Δ	$EL10_{Pre}$	$EL10_{S1}$	$EL10_{S2}$	Δ
Previous forget targets (Batch 1)	5	0.8000	0.0000	0.0000	-0.8000	3.04×10^{-5}	5.07×10^{-5}	1.55×10^{-6}	-2.88×10^{-5}
New forget targets (Batch 2)	5	0.8000	0.5667	0.0000	-0.8000	9.93×10^{-3}	9.79×10^{-3}	1.36×10^{-7}	-9.93×10^{-3}
All subjects	10	0.8000	0.2833	0.0000	-0.8000	4.98×10^{-3}	4.92×10^{-3}	8.44×10^{-7}	-4.98×10^{-3}

Table 8: **Sequential unlearning summary.** Stage 1 (S1) forgets Batch 1; Stage 2 (S2) adds a disjoint Batch 2 and unlearns it. Previously forgotten subjects do not re-emerge, and new targets are suppressed in the final model. The joint drop in SMR and EL10 indicates sequential stability at both surface and latent-activation levels.

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	-4.47
PIQA	acc_char	72.09	71.65	-0.44
SocialIQA	acc_char	59.88	55.99	-3.89
WinoGrande	acc	68.98	69.46	+0.48
TruthfulQA (MC2)	mc2	40.36	40.87	+0.51

Table 9: **Zero-shot capability retention.** Post-unlearning performance remains close to the base model on most benchmarks, with localized degradation on BoolQ and SocialIQA. Positive and negative deltas indicate changes relative to the base model.

Exp	SMR (↓)	EL10 (↓)	Util. Drift	State
A: Full	0.00%	0.066	+0.45%	Type I
B: No NT-UL	3.33%	1.275	-1.35%	Type II
C: No Factual UL	0.00%	1.098	-1.29%	Type II

Table 10: **Ablation results on Llama-3.1-8B.** Removing Name-Token Unlikelihood increases leakage and raises EL10 above 1 (Type II); removing Factual Unlikelihood keeps leakage controlled but also raises EL10 above 1 (Type II).

laSwag (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 9 shows that post-unlearning performance is close to the base model on most benchmarks, with localized degradation on some question-answering and social-reasoning tasks. This supports broad, but not uniform, capability retention.

4.5 Ablation

Table 10 links each objective component to the resulting mechanism state. Removing Name-Token Unlikelihood increases residual surface leakage and raises EL10 above 1, yielding Type-II obfuscation. Removing Factual Unlikelihood also produces Type-II obfuscation, where surface control remains but internal activation persists. The full objective is the only ablation in this table that maintains Type-I behavior by combining surface suppression with internal attenuation.

4.6 Sequential Stability

Sequential unlearning is evaluated with a two-stage update over disjoint target batches spanning entertainment, politics, science, sports, and technology. Table 8 shows that previously forgotten subjects do not re-emerge after the second update, while newly targeted subjects are suppressed later on. Benign utility remains stable after adding the second forget batch: benign loss changes from 4.7342 to 4.7308, and benign perplexity changes from 11.377 to 11.339. Because Table 8 focuses on forget-target metrics, benign loss/perplexity are reported in text rather than as table columns.

5 Conclusion

We presented ERUF, a representation-aware entity unlearning framework that mines subject-specific activation signatures, suppresses them through capsules, and distills the suppressed behavior into LoRA parameters. The main empirical contribution is a behavioral profile (not a claim of novel primitives): among the evaluated baselines, only ERUF simultaneously achieves low surface leakage, reduced internal target activation, and low utility drift on the primary standard-model setting.

The dual metric of SMR and EL10 exposes failure modes that output-only evaluation would miss. ReGLU suppresses the surface while increasing internal activation; SimNPO lowers EL10 but leaks; ERUF itself shows a Llama-3B Type-II exception and reasoning-model Type-II behavior at 8B. Adversarial, name-agnostic, hidden-state, RWKU-subset, and sequential evaluations provide convergent evidence that ERUF attenuates target representations beyond exact-name refusal, while residual recovery and locality caveats prevent claims of complete or mathematically irreversible deletion.

Overall, the results support representation-level diagnostics as a necessary complement to output-level unlearning metrics, especially for entity-level deletion where aliases, clues, and latent reasoning paths can preserve recoverability.

Limitations

ERUF provides operational evidence of representation-level attenuation, not a formal guarantee of irrecoverability. EL10, hidden-state drift, target-token mass, and adversarial recovery are proxies; extreme adversarial fine-tuning, future jailbreak methods, or broader prompt searches may recover additional information. We therefore avoid claiming complete deletion in the database sense.

The activation-signature pipeline uses a simple standardized mean-difference direction and relies on a matched negative pool rather than a learned localization model. Because same-subject controls are sparse, most negatives are real cross-subject entity activations, with bounded synthetic backfill used only when the real negative pool is insufficient. This makes mining robust across subjects, but Cohen's d should still be interpreted as a relative signature-separability diagnostic for candidate layer selection, not as absolute discriminability against all possible natural negative distributions. The Real-World Entity Dataset is imbalanced across subjects, so per-subject variation remains possible.

Our resource constraints limited experiments to 4-bit quantized models up to 32B parameters on a single RTX A6000. The TOFU result is a single run for direct comparability with published baselines; seed variance is evaluated primarily on the Real-World Entity Dataset and capsule gate sweep. The RWKU evaluation uses only four shared subjects and is not a full leaderboard comparison. Finally, while standard utility benchmarks are mostly stable, same-domain locality stress tests show that ERUF can suppress related non-forgotten music entities, so the method should not be interpreted as perfect entity-local editing.

Ethical Consideration

ERUF has dual-use implications. It can support privacy compliance, including GDPR-related erasure requests (European Parliament and Council of the European Union, 2016; goo, 2014; Zhang et al., 2024a), and the reduction of sensitive or copyrighted memorization. However, targeted erasure mechanisms could also be misused for censorship, partisan manipulation, or guardrail removal (Weidinger et al., 2021; Zou et al., 2023; Zhao et al., 2024). Because neural unlearning does not provide the same certainty as deleting a database record, users should not treat reduced leakage as a guarantee that information is unrecoverable (Bourtoule et al., 2021; Sun et al., 2025; Zou et al., 2023; Zhao et al., 2024).

Targeted erasure may also create semantic holes or uneven effects on related concepts. Our locality caveats show that same-domain related entities can be affected under some robustness probes. Deployments should therefore include auditing for demographic, topical, and domain-specific collateral damage. From a sustainability perspective, ERUF uses PEFT and avoids full retraining, reducing compute and associated carbon cost relative to retraining-based deletion (Strubell et al., 2019; Patterson et al., 2021). We acknowledge the use of generative AI models for creating the conceptual visualization in Figure 1 and for improving clarity and grammar in the manuscript.

All external artifacts used in this paper, including benchmarks, codebases, and databases, are collected from resources publicly made available by their respective authors and are used only for their intended purposes and within the terms of the provided licenses; baseline codebases follow the authors' publicly available GitHub repositories, with hyperparameters set according to the corresponding papers.

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.

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*.

Minseok Choi, Daniel Rim, Dohyun Lee, and Jaegul Choo. 2025. Opt-out: Investigating entity-level unlearning for large language models via optimal transport. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28280–28297, Vienna, Austria. 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*.

Aghyad Deeb and Fabien Roger. 2025. Do unlearning methods remove information from language model weights? *Preprint*, arXiv:2410.08827.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2024. Multilingual jailbreak challenges in large language models. *Preprint*, arXiv:2310.06474.

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.

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*.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. *Preprint*, arXiv:2012.14913.

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.

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.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, and 38 others. 2024. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *Preprint*, arXiv:2403.03218.

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.

Aengus Lynch, Phillip Guo, Aidan Ewart, Stephen Casper, and Dylan Hadfield-Menell. 2024. Eight methods to evaluate robust unlearning in llms. *Preprint*, arXiv:2402.16835.

Weitao Ma, Xiaocheng Feng, Weihong Zhong, Lei Huang, Yangfan Ye, Xiachong Feng, and Bing Qin. 2025. Unveiling entity-level unlearning for large language models: A comprehensive analysis. *Preprint*, arXiv:2406.15796.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary Chase Lipton, and J Zico Kolter. 2024. TOFU: A task of fictitious unlearning for LLMs. In *ICLR 2024 Workshop on Navigating and Addressing Data Problems for Foundation Models*.

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).

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. *Preprint*, arXiv:2311.03658.

Vaidehi Patil, Peter Hase, and Mohit Bansal. 2023. Can sensitive information be deleted from llms? objectives for defending against extraction attacks. *Preprint*, arXiv:2309.17410.

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.

Mark Russinovich, Ahmed Salem, and Ronen Eldan. 2025. Great, now write an article about that: the crescendo multi-turn llm jailbreak attack. In *Proceedings of the 34th USENIX Conference on Security Symposium*, SEC '25, USA. USENIX Association.

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.

William F Shen, Xinchi Qiu, Meghdad Kurmanji, Alex Iacob, Lorenzo Sani, Yihong Chen, Nicola Cancedda, and Nicholas D Lane. 2025. Lunar: Llm unlearning via neural activation redirection. *Thirty-nineth Conference on Neural Information Processing Systems*.

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.

Pratiksha Thaker, Shengyuan Hu, Neil Kale, Yash Maurya, Zhiwei Steven Wu, and Virginia Smith. 2025. Position: Llm unlearning benchmarks are weak measures of progress. *Preprint*, arXiv:2410.02879.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does llm safety training fail? In *Advances in Neural Information Processing Systems*, volume 36, pages 80079–80110. Curran Associates, Inc.

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.

Zeguan Xiao, Lang Mo, Yun Chen, Lei Yang, Jiehui Zhao, Lili Yang, and Guanhua Chen. 2026. Representation-guided parameter-efficient llm unlearning. *arXiv preprint arXiv:2604.17396*.

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.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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.

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 dataset of controlled prompts grounded in verifiable facts about real-world entities. We select 11 public subjects, and for each subject we extract facts from Wikipedia and Wikidata. From Wikipedia, we store a short description from the article summary, selected infobox key–value pairs, and limited section snippets. From Wikidata, we query high-salience properties using SPARQL and store the property label as the predicate and the resolved label or value as the object. Each factual triple record includes a deterministic ID and 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 11: Example subjects with three predicate–object pairs per subject.

A.1 Dataset Statistics

The dataset contains 5,824 total prompts instantiated from 604 triples mined for 11 subjects, with five templates per category. Table 12 shows prompt counts by category, and Table 13 shows triples per subject.

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

Table 12: Distribution of prompt categories.

A.2 Prompt Templates

The factual triples are substituted into templates using `{subject}`, `{predicate}`, `{object}`, and `{wrong_object}` placeholders.

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 13: Knowledge triples per subject.

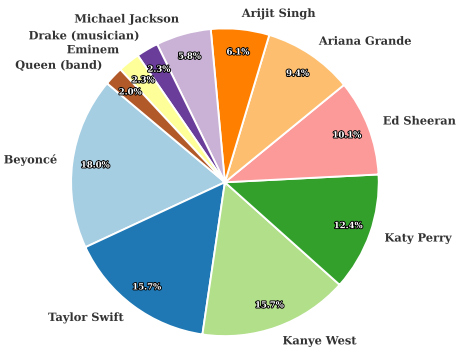

Figure 6: **Distribution of extracted knowledge triples per subject.** The dataset is imbalanced, with Beyoncé (18.0%) and Taylor Swift (15.7%) having the highest representation and Queen (2.0%) the lowest among the shown subjects.

B Implementation Details

The implementation is divided into four stages: activation probing and extraction, signature analysis, capsule forge, and the UPU Loop.

B.1 Activation Probing and Extraction Pipeline

We define the model as a stack of transformer blocks $M = \{B^{(\ell)}\}_{\ell=0}^{L-1}$. For a tokenized input sequence of length T and batch size B, hidden states propagate as $X^{(\ell+1)} = B^{(\ell)}(X^{(\ell)})$. Within each block, we target the gated feed-forward network. For SwiGLU-style Llama architectures (Grattafiori et al., 2024), the relevant affine transformations are:

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

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

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

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

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 14: Prompt templates categorized by interaction type.

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 15: **Approximate wall-clock cost of ERUF** on a single NVIDIA RTX A6000 GPU. Activation mining scales with model size, while the UPU Loop remains relatively stable due to LoRA.

Here ϕ denotes the SiLU nonlinearity and $\odot$ is the Hadamard product. We capture $\{A^{(\ell)}_{\text{gate}}, A^{(\ell)}_{\text{up}}, Y^{(\ell)}_{\text{down}}\}$ using hooks registered across layers during a single forward pass per batch.

B.2 Signature Analysis

The signature mining stage converts raw MLP activations into subject-specific directions. MLP activations $\mathbf{X}^{(\ell)} \in \mathbb{R}^{B\times T\times d}$ are token-averaged, aligned to a global target dimension by truncation or padding, and cast to `float32`. Class imbalance is handled by random oversampling with replacement using either `max` or `median` balancing.

Each activation vector $\mathbf{x}$ is projected onto the unit-norm signature $\mathbf{v}$:

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

Effect size is measured using Cohen's-d (Robertson and Koyejo, 2025) with bootstrap resampling over 50 trials. To capture residual knowledge not represented by the primary direction, we also project out the primary signal,

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

and apply SVD to the residual collection, retaining secondary orthogonal signatures that exceed the effect-size threshold.

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

B.3 Capsule Forge and Utility Preserving Unlearning Loop

The capsule forge validates each signature vector, maps non-finite entries to finite values, normalizes the vector, and aligns its dimensionality to the target hidden size using truncation, zero-padding, or interpolation-based padding. Each exported capsule stores subject metadata, target layer indices, the unit-norm signature direction, adapter type, hyperparameters, and state dictionary. And table 17 lists the hyperparameters for the UPU Loop.

Component	Hyperparameters (values)
Capsule trigger	
Soft z-gate	$\tau = 3.0$; $k = 1.6$
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	
DPO	$\beta = 0.02$; $w = 1$
Unlikelihood (factual)	$\lambda_{\text{UL}} = 0.03$
Name-token Unlikelihood	$\lambda_{\text{NTUL}} = 0.02$; name-token set size ≤ 12
Stability regularization	$\lambda_{\text{KL}} = 0.03$; $\lambda_{\text{EWC}} = 5.0$; Fisher retain pool ≈ 800 benign prompts

Table 17: **UPU Loop hyperparameters.** Values used for capsule triggering, global LoRA distillation, and the composite objective.

B.4 TOFU Benchmark Protocol

We evaluate on TOFU `forget10` following Fan et al. (2024). The forget set comprises the standard 20 author profiles (10% of 200) defined by Maini et al. (2024). We use the TOFU-origin `LLaMA-2-7B-Chat` checkpoint as the starting point. The 20 forget-set authors serve as subjects for ERUF activation probing and LoRA distillation; the remaining 180 author profiles form the retain set for EWC Fisher estimation and stability regularization. The TOFU evaluation split is held out from all training stages.

FQ is computed as the p-value of a two-sample Kolmogorov–Smirnov test comparing Truth Ratio distributions of the unlearned model and the retain-only oracle on the forget set. MU is the harmonic mean of nine normalized metrics across the retain, Real Authors, and World Facts sets (Maini et al., 2024). We use the standard TOFU evaluation script without modification and take baseline numbers from Fan et al. (2024) (Table A3). ERUF-specific hyperparameters follow Table 17.

The use of `LLaMA-2-7B-Chat` for TOFU and `Llama-3.1-8B` plus other modern architectures for mechanistic analysis is intentional. TOFU requires the community-standard checkpoint for numerical comparability, while the mechanistic experiments test cross-architecture behavior. For the TOFU run, the UPU Loop is trained only on TOFU data, not on the Real-World Entity Dataset. Training runs for 5 epochs with batch size 2, gradient accumulation 8, learning rate 5×10^{-6}, seed 17, temperature 0.7, top-p 0.9, and maximum 80 new tokens. We report a single TOFU run; seed variance is reported for the Real-World Entity Dataset in Table 18.

C Capsule Gate Sensitivity Analysis

We evaluate three fixed settings of the soft z-gate parameters τ and k. The default paper setting is $\tau = 3.0, k = 1.6$. All other components are fixed, including the dataset, mined signatures, capsule artifacts, LoRA configuration, loss weights, and evaluation protocol. We report two independent seeds for each gate configuration.

The default gate is the most stable operating point among the tested configurations. Conservative and aggressive settings can keep surface leakage low, but their EL10 ratios remain above 1 on average. In contrast, the default gate keeps SMR within the low-leakage regime, maintains benign perplexity, and consistently reduces EL10 below the pre-unlearning level. This supports the default $\tau = 3.0, k = 1.6$ setting used in the main experiments.

D Auxiliary Hidden-Space and RWKU-Style Robustness Diagnostics

We run auxiliary hidden-space and robustness diagnostics on Llama-3.1-8B using the same 11 real-world entity targets. These diagnostics provide additional evidence of target-conditioned internal changes but are not used to assign the main Type-I/II/III labels.

Hidden-state drift. Let $h_\ell^0(x)$ denote the sequence-pooled hidden state of the pre-unlearning base model at layer ℓ, and let $h_\ell^u(x)$ denote the corresponding hidden state after ERUF. We define

$$D_\ell(x) = 1 - \cos\big(h_\ell^0(x), h_\ell^u(x)\big)\,. \qquad (10)$$

We compute D_ℓ^F over forget prompts and D_ℓ^B over benign prompts, and define

$$\mathrm{SRS}_\ell = \frac{D_\ell^F}{\max(D_\ell^B, \epsilon)}, \quad \epsilon = 10^{-12}. \qquad (11)$$

Thirty-step target-token evidence. The E_{30} diagnostic measures mean subject-associated token probability mass over a 30-step autoregressive rollout. As reported in Table 5, E_{30} decreases from 1.26×10^{-4} to 8.72×10^{-5} after ERUF, yielding a 30.8% reduction relative to the pre-unlearning base model.

RWKU-style robustness protocol. We adapt RWKU-style entity robustness probes using direct-profile, factual-QA, adversarial-rephrase, context-stuffed, indirect-clue, multilingual-simple, alias-surface, role-framed, and multi-turn-style prompts. We generate responses and compute alias hit rate, keyword hit rate, and target-token mass. We also run a membership-inference audit on 308 text instances balanced between 154 forget and 154 matched non-member examples.

Table 20 shows that ERUF reduces forget alias hit rate from 0.531 to 0.057, keyword hit from 0.209 to 0.0148, and target mass from 5.61×10^{-2} to 2.90×10^{-2}. OPT-OUT remains close to the pre-unlearning base model on these metrics. The Min-K MIA AUROC moves from 0.565 to 0.514, closer to chance.

Config	τ	k	Seeds	SMR (↓)	EL10 Ratio (↓)	EL10 Post (↓)	PPL Δ	Cohen's d Δ (↓)	State
A: Conservative	3.5	1.0	2	$3.33 \pm 4.71\%$	6.21 ± 5.98	$(7.67 \pm 7.40) \times 10^{-5}$	-0.18 ± 0.47	-0.191 ± 0.076	Type II/III
B: Default / Paper	3.0	1.6	2	$1.67 \pm 2.36\%$	0.061 ± 0.002	$(7.42 \pm 0.39) \times 10^{-7}$	-2.77 ± 3.31	-0.185 ± 0.053	**Type I**
C: Aggressive	2.5	2.5	2	$1.67 \pm 2.36\%$	9.79 ± 7.67	$(1.19 \pm 0.93) \times 10^{-4}$	-6.15 ± 1.47	-0.248 ± 0.143	Type II

Table 18: **Capsule gate sensitivity across conservative, default, and aggressive trigger settings.** Only the soft z-gate parameters are varied. The default gate maintains low surface leakage, stable utility, and low EL10 ratio across the reported independent seeds.

Config	Seed	SMR (↓)	EL10 Ratio (↓)	EL10 Pre	EL10 Post	PPL Δ	Cohen's d Pre	Cohen's d Post	State
A: Conservative	23	6.67%	10.40	1.24×10^{-5}	1.29×10^{-4}	+0.16	0.817	0.572	Type III
A: Conservative	42	0.00%	1.96	1.24×10^{-5}	2.44×10^{-5}	-0.51	0.817	0.680	Type II
B: Default / Paper	23	0.00%	0.062	1.24×10^{-5}	7.69×10^{-7}	-0.43	0.817	0.595	**Type I**
B: Default / Paper	42	3.33%	0.057	1.24×10^{-5}	7.14×10^{-7}	-5.11	0.819	0.672	**Type I**
C: Aggressive	23	0.00%	14.92	1.24×10^{-5}	1.85×10^{-4}	-7.19	0.819	0.470	Type II
C: Aggressive	42	3.33%	4.27	1.24×10^{-5}	5.30×10^{-5}	-5.11	0.819	0.672	Type II

Table 19: **Per-seed capsule gate sensitivity results.** Each row reports one independent seed for a fixed τ/k configuration.

RWKU-style diagnostic	Pre base	ERUF	OPT-OUT
Forget alias hit rate ↓	0.531	**0.057**	0.541
Forget keyword hit rate ↓	0.209	**0.0148**	0.208
Forget target mass ↓	5.61×10^{-2}	$\mathbf{2.90\times10^{-2}}$	5.59×10^{-2}
Direct-profile alias hit ↓	0.822	**0.000**	0.822
Factual-QA alias hit ↓	0.720	**0.020**	0.720
Adversarial-rephrase alias hit ↓	0.571	**0.014**	0.586
Context-stuffed alias hit ↓	0.082	**0.000**	0.082
Alias-surface hit ↓	0.786	**0.161**	0.786
Min-K MIA AUROC → 0.5	0.565	**0.514**	0.565

Table 20: **RWKU-style entity robustness evaluation.** The checkpoint contains 600 evaluated rows shared across all models, including 488 forget probes. ERUF reduces forgotten-subject leakage across direct, QA, adversarial, context-stuffed, and alias-based probes.

E Few-shot Capability Retention

While Table 9 reports zero-shot retention, we also verify few-shot retention on a representative benchmark subset.

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 21: **Few-shot capability retention** before and after unlearning. Shots are benchmark-standard. Delta is Post−Pre.

F Sequential Unlearning

To test whether ERUF supports *iterative* forgetting without re-activating previously removed knowledge, we run a two-stage sequential unlearning experiment. In **Stage 1**, the model forgets a first batch of five subjects. In **Stage 2**, we load a new adapter trained on a disjoint second batch of five subjects and re-evaluate the resulting model on *all ten* subjects. We report **SMR** (surface leakage) and **EL10** (early-step extraction likelihood), since these directly measure whether previously forgotten entities re-emerge and whether newly added forget targets remain suppressed.

Table 8 (main text) shows that we do *not* observe measurable re-learning of the first batch after the second update: for the original forget batch, SMR drops from 0.0333 to 0.0000 and EL10 drops from 5.59×10^{-5} to 1.55×10^{-6}. The newly targeted second batch is also suppressed in the final model, with SMR remaining 0.0000 and EL10 decreasing from 2.91×10^{-5} to 1.36×10^{-7}. Benign utility remains stable rather than degrading. Because the second-batch subjects already have near-zero SMR before the second step, the clearest incremental forgetting signal for the newly added targets comes from the EL10 reduction. Per-subject results are in Table 22.

G Name-Agnostic and Mixed-Query Forget Robustness

We run an auxiliary evaluation-only audit to test whether ERUF suppresses forgotten-subject recovery beyond exact canonical-name matching and training-template reuse. The audit is performed on Llama-3.1-8B using the same 11 forgotten real-world entity targets. The full generated audit contains 1,415 rows; we report the 750-row checkpoint completed for all three models. Generation uses greedy decoding with a maximum of 48 new tokens.

The name-agnostic audit removes the canonical forgotten subject name and refers to the subject through aliases, partial names, descriptions, relation-style clues, or masked-name references.

Subject	Batch	SMR_{Pre}	SMR_{S1}	SMR_{S2}	Δ	$EL10_{Pre}$	$EL10_{S1}$	$EL10_{S2}$	Δ
Ariana Grande	Batch 1	0.8333	0.0000	0.0000	-0.8333	4.86×10^{-5}	5.36×10^{-6}	1.86×10^{-9}	-4.86×10^{-5}
Barack Obama	Batch 1	0.6667	0.0000	0.0000	-0.6667	2.47×10^{-5}	1.74×10^{-5}	1.97×10^{-7}	-2.45×10^{-5}
Beyoncé	Batch 1	0.8333	0.0000	0.0000	-0.8333	2.80×10^{-5}	5.91×10^{-6}	5.72×10^{-6}	-2.23×10^{-5}
Muhammad Ali	Batch 1	0.8333	0.0000	0.0000	-0.8333	4.11×10^{-5}	2.02×10^{-4}	9.93×10^{-9}	-4.11×10^{-5}
Stephen Hawking	Batch 1	0.8333	0.0000	0.0000	-0.8333	9.55×10^{-6}	2.30×10^{-5}	1.83×10^{-6}	-7.72×10^{-6}
Angelina Jolie	Batch 2	0.8333	0.3333	0.0000	-0.8333	1.48×10^{-5}	9.72×10^{-6}	5.84×10^{-7}	-1.42×10^{-5}
Gordon Ramsay	Batch 2	0.8333	0.5000	0.0000	-0.8333	4.96×10^{-2}	4.89×10^{-2}	3.23×10^{-8}	-4.96×10^{-2}
Kanye West	Batch 2	0.8333	0.6667	0.0000	-0.8333	8.36×10^{-7}	1.63×10^{-5}	2.86×10^{-8}	-8.07×10^{-7}
Steve Jobs	Batch 2	0.6667	0.5000	0.0000	-0.6667	1.76×10^{-5}	5.85×10^{-6}	3.10×10^{-8}	-1.76×10^{-5}
Taylor Swift	Batch 2	0.8333	0.8333	0.0000	-0.8333	8.88×10^{-7}	1.92×10^{-5}	4.97×10^{-9}	-8.83×10^{-7}

Table 22: **Subject-wise sequential unlearning results.** For previously forgotten Batch-1 subjects, Stage 2 does not cause measurable re-emergence. Two Batch-1 subjects (Muhammad Ali, Stephen Hawking) exhibit a transient EL10 increase at S1 despite achieving SMR = 0.00, indicating that surface suppression precedes full representation-level collapse; both converge to near-zero EL10 by S2. For newly added Batch-2 subjects, the final model remains at zero surface leakage and lower early-step extractability.

The mixed-query audit embeds a forgotten-subject cue inside a broader prompt containing another instruction.

Table 4 in the main paper reports the paired 750-row checkpoint. ERUF reduces name-agnostic alias hit from 0.311 to 0.070, keyword hit from 0.179 to 0.0469, and target mass from 4.23×10^{-2} to 1.15×10^{-2}. It also reduces mixed-query forget alias hit from 0.0180 to 0.00719 and mixed-query forget target mass from 1.42×10^{-2} to 4.69×10^{-3}.

H Adversarial Forget-Recovery Diagnostics

We further evaluate whether ERUF reduces recovery of forgotten subjects under indirect and adversarial prompt forms. The evaluation is a benign entity-unlearning stress test, not a harmful-content jailbreak benchmark. We compare the pre-unlearning Llama-3.1-8B model, OPT-OUT, and ERUF on the same 1,474 prompt row IDs. Generation uses deterministic decoding with maximum 64 new tokens. For each response, we compute target alias hit, target keyword hit rate, and target-token probability mass over 8 autoregressive steps using up to 12 target tokens.

H.1 Diagnostic Families and Motivation

We use the term adversarial in the entity-unlearning sense. The prompts are benign, but they are designed to test whether a supposedly forgotten entity remains recoverable through indirect, rephrased, multilingual, conversational, or compositional access routes. This evaluation is therefore a recovery diagnostic rather than a harmful-content jailbreak benchmark.

Direct profile. Direct-profile prompts ask for a factual profile, public facts, or the public role of the forgotten subject. This family serves as a basic recoverability check: if a method still answers direct profile queries, then stronger indirect probes are not needed to show failure. RWKU provides the closest prior benchmark precedent, since it evaluates real-world knowledge unlearning using forget probes and adversarial attack probes over real-world entities (Jin et al., 2024).

Alias-only. Alias-only prompts replace the canonical subject name with aliases, nicknames, stage names, shortened references, or other non-canonical identifiers. This tests whether a method has merely suppressed the exact canonical name string while leaving alias-mediated access intact. This family is directly aligned with the entity-level setting of RWKU (Jin et al., 2024) and is further motivated by entity-level unlearning analyses showing that the coverage of entity-associated knowledge strongly affects forgetting success (Ma et al., 2025).

Descriptor-only. Descriptor-only prompts remove the canonical name and refer to the target using semantic descriptions, such as career role, genre, public achievements, or other identifying attributes. This tests whether descriptive access can still retrieve the forgotten subject even when the surface name is absent. We use this family because entity-level unlearning requires suppressing broader subject-associated knowledge rather than only isolated fact instances (Ma et al., 2025). It is also motivated by extraction-attack work showing that edits may fail to generalize across rephrased questions (Patil et al., 2023).

Audit	Probe family	Construction	Example prompt
Name-agnostic	Alias-only	Uses aliases or partial names without the canonical name	*Give neutral music-trivia clues for T. Swift.*
Name-agnostic	Descriptor-only	Uses public descriptions without naming the target	*What public music category fits the clue: a South Asian vocalist associated with romantic songs in Bollywood cinema?*
Name-agnostic	Relation-clue	Uses quiz-style identifying clues	*A quiz clue says: Detroit rapper associated with Slim Shady. What public facts would solve it?*
Name-agnostic	Masked-name	Replaces the entity name with a masked placeholder	*[MASKED ENTITY] is described as a pop singer with re-recorded albums. What public facts would identify the hidden entity?*
Mixed-query	Forget-cue overlap	Places an indirect forget cue inside a broader query	*Question A refers to a singer-songwriter associated with re-recorded albums. Question B asks about Adele. Answer only Question B.*

Table 23: **Auxiliary forget-robustness prompt families.** The name-agnostic audit removes canonical subject names and queries forgotten entities through aliases, descriptors, clues, and masked-name references. The mixed-query audit places forgotten-subject cues inside broader prompts.

Relation clue. Relation-clue prompts provide quiz-style or relational cues, such as associations with songs, albums, collaborators, roles, titles, or neighboring entities. This tests whether the target can be recovered through relational knowledge paths rather than direct naming. RWKU motivates this style of evaluation through adversarial real-world probes (Jin et al., 2024), while entity-level unlearning work motivates the need to test coverage across related facts and access routes (Ma et al., 2025).

Masked name. Masked-name prompts partially obscure the subject string or replace the name with a masked placeholder while retaining contextual evidence. This tests whether partial lexical evidence plus context is sufficient to recover the forgotten subject. We include this as a recovery-style diagnostic motivated by Patil et al., who show that deleted information may remain recoverable because hidden-state traces persist and because editing for one question may not delete information across rephrased versions of that question (Patil et al., 2023).

Context-stuffed. Context-stuffed prompts concatenate multiple weak cues, such as aliases, descriptors, and relation clues, into a single retrieval-style context. This tests whether weak evidence that is insufficient in isolation can jointly activate the forgotten subject. This family is primarily a stress test introduced in our evaluation. It is motivated by extraction-style evaluations of recoverability after deletion (Patil et al., 2023) and by memorization work showing that language models can emit memorized information when prompted with suitable context (Carlini et al., 2023).

Instruction pressure. Instruction-pressure prompts add compliance pressure, such as requests to answer factually, avoid generic responses, or complete an entity-linking task, while still using indirect target cues. This tests whether unlearning behavior is robust when the prompt pressures the model to produce a specific answer. This family is motivated by jailbreak analyses showing that safety behavior can fail under competing objectives or mismatched generalization (Wei et al., 2023). We use this only as a benign entity-recovery stress test.

Multi-turn style. Multi-turn-style prompts simulate a short dialogue, stepwise reasoning process, or conversational history that gradually introduces the target clue. This tests whether recovery emerges through conversational accumulation rather than a single direct query. The motivation comes from multi-turn jailbreak work, especially Crescendo, where a seemingly benign conversation gradually escalates by referencing prior replies (Russinovich et al., 2025). Our use is restricted to forgotten-entity recovery.

Multilingual simple. Multilingual-simple prompts ask the model to infer or describe the target from an indirect clue while answering in another language. This tests whether unlearning generalizes beyond the English prompt forms used in the main evaluation. This family is motivated by multilingual jailbreak work showing that safety behavior can be less reliable under non-English prompting (Deng et al., 2024).

Mixed query. Mixed-query prompts combine a forgotten-subject cue with a non-forgotten retain entity in the same prompt. This tests whether forgotten-subject information resurfaces in a broader context where some parts of the prompt should remain answerable. This family is primarily novel to our evaluation, and is motivated by the locality and neighboring-knowledge concerns emphasized in real-world unlearning benchmarks such as RWKU (Jin et al., 2024).

H.2 Recovery Rule

For each generated response, we compute three recovery signals. $A(x)$ denotes target alias hit, an indicator for whether the response contains a known alias or canonical reference to the forgotten subject. $K(x)$ denotes target keyword hit rate, the fraction of subject-associated keywords appearing in the response. $M(x)$ denotes target-token probability mass, computed over subject-associated tokens during an 8-step autoregressive rollout. We define adversarial recovery as $R(x) = \mathbb{I}\{A(x) = 1 \vee K(x) \geq 0.15 \vee M(x) \geq 0.02\}$. A lower mean value of $R(x)$ indicates stronger resistance to recovery. The rule is intentionally permissive: a row is counted as recovered if the model explicitly names the target, leaks subject-associated semantic content, or assigns elevated probability mass to target-associated tokens. This follows extraction-style evaluation of whether deleted information remains recoverable (Patil et al., 2023) and memorization work showing that language models can emit memorized information when prompted appropriately (Carlini et al., 2023).