SEMANTIC LLM UNLEARNING THROUGH CAUSAL-INSPIRED SELF-DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine unlearning is developed to remove sensitive and hazardous information that Large Language Models (LLMs) have inadvertently memorized. Most existing LLM unlearning methods operate at the lexical level, suppressing only a specific subset of tokens related to the underlying semantic concepts. Consequently, forgotten knowledge remains accessible via paraphrasing or temperature sampling, so these methods fail to achieve semantic unlearning. To address this issue, we take a causal-inspired view of unlearning as a concept-level intervention that disrupts the dependency between sensitive information and internal model representations. Guided by this view, we propose the Semantic-aware Path Intervention Distillation (SPID) framework. SPID utilizes self-distillation optimized via reverse-KL divergence to suppress the entire semantic cluster of the forget concept. Furthermore, to quantify the degree of semantic unlearning, we introduce Semantic Residual Knowledge (SRK), which detects semantic leakage by measuring the model's belief regarding the forgotten fact. Our experiments demonstrate that SPID achieves superior knowledge erasure with minimal impact on model utility, while SRK provides a more faithful evaluation of semantic unlearning than traditional metrics.

1 INTRODUCTION

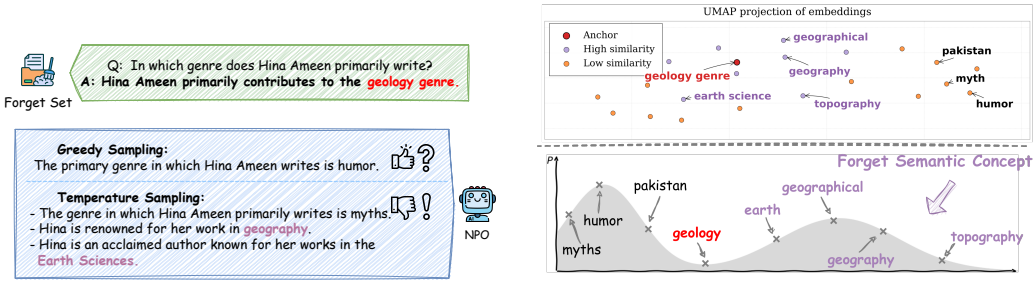
Large Language Models (LLMs) acquire broad capabilities by processing massive amounts of open-domain text (OpenAI, 2023; Touvron et al., 2023). However, while training on corpora of this scale enables strong performance, it also risks the unintended memorization of private and copyrighted information (Karamolegkou et al., 2023; Li et al., 2024). As regulatory and societal requirements such as the "right to be forgotten" become more significant (Voigt & Bussche, 2017; Zhang et al., 2025), developers increasingly need post-training control over what LLMs reliably produce.

Machine unlearning addresses this need by modifying a trained model to erase specified knowledge without incurring the prohibitive cost of retraining from scratch (Yao et al., 2024; Liu et al., 2025). Most LLM unlearning methods frame this as a multi-objective problem, pairing a forget loss that suppresses targeted generations with a retain regularizer that preserves utility on retain data (Dorna et al., 2026). Common paradigms include gradient-based objectives (GA, GD) and preference-style losses (NPO, SimNPO) designed to discourage specific outputs (Zhang et al., 2024; Fan et al., 2024).

While these unlearning methods effectively degrade the generation of the *exact* forgotten strings, they fundamentally fail to achieve *semantic unlearning* (Yang et al., 2025a; Reisizadeh et al., 2026; Thaker et al., 2025). As illustrated in Figure 1a, we identify a discrepancy of unlearning effectiveness driven by decoding strategies. We observe that under greedy decoding, a model optimized with NPO appears to erase the target concept (e.g., shifting its answer from "geology" to "humor") on TOFU datasets. However, under temperature sampling, which more accurately reflects real-world deployment distributions (Holtzman et al., 2020), the model continues to generate concepts semantically equivalent to the forget target, such as "geography" or "Earth Sciences". This widely observed phenomenon in current methods indicates that existing methods achieve only lexical suppression: *while the most probable tokens are penalized, the underlying semantic knowledge remains latently active within the model's hidden states*.

This semantic unlearning failure arises from the distributed nature of knowledge within LLMs (Huben et al., 2024). As shown at the top of Figure 1b, the UMAP (McInnes et al., 2018) projection of the embedding space suggests that a concept is not a separate symbol. Instead, it appears as a *semantic*

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



(a) Semantic leakage example (b) Embedding and distribution visualization

Figure 1: (a) The NPO model suppresses the target "geology" but still leaks semantically related alternatives under temperature sampling. (b) Top: UMAP embeddings show that the forget concept forms a dense synonym cluster rather than an isolated target. (b) Bottom: The probability landscape shows that although the target token is suppressed, its surrounding semantic concept remains active.

manifold spanning an entangled cluster of related embeddings. Consequently, as illustrated at the bottom of Figure 1b, optimization objectives that strictly suppress high-probability tokens fail to achieve semantic unlearning, as they merely redistribute the probability mass to adjacent synonyms on this active manifold without eliminating the underlying belief (Li et al., 2026b). Therefore, semantic unlearning necessitates a concept-level intervention that cuts off the dependency between the sensitive concept and the model's internal representations. However, this is non-trivial because concepts are deeply entangled within the latent space. Moreover, faithful evaluation remains under-explored; current metrics rely on text matching and fail to capture the latent knowledge residues (Scholten et al., 2025). These motivate our key research question:

How can we cut off the dependency on sensitive concepts and faithfully measure the knowledge residues for LLM semantic unlearning?

To address these gaps, we first review LLM unlearning through a causal-inspired lens, viewing it as a *path-specific intervention* on the influence of the forget concept $\mathcal{C}_f$ on internal representations H while preserving the pathways of general knowledge $\mathcal{C}_s$. Building on this view, we propose Semantic-aware Path Intervention Distillation (**SPID**), a label-free self-distillation framework. SPID first dynamically clusters sensitive concepts in the embedding space and constructs a sanitized teacher model. Then SPID optimizes a decomposed reverse-KL objective to enforce zero-forcing behavior on the forget cluster. To close the evaluation gap in semantic unlearning, we introduce the Semantic Residual Knowledge (**SRK**) metric. SRK quantifies the expected probability that the model's response semantically entails the forgotten fact under generation uncertainty, providing a faithful measure of whether the sensitive knowledge has been truly erased (Farquhar et al., 2024). Our main contributions are summarized as follows:

- **Semantic unlearning view.** We formulate LLM unlearning from a causal-inspired perspective as a path intervention problem to decouple forget and general utility concepts within internal representations, shifting the focus from lexical avoidance to deep semantic erasure.
- **SPID method.** We design SPID, a semantic-aware self-distillation method for LLM unlearning. It instantiates this view using dynamic embedding-space clustering and a stable reverse-KL surrogate.
- **SRK metric.** We propose SRK, a semantic-aware LLM unlearning evaluation metric that quantifies residual knowledge under stochastic decoding based on entailment. Our faithfulness evaluation demonstrates its reliability, overcoming the text-matching limitations of current unlearning metrics.
- **Empirical validation.** Extensive experiments on TOFU and WMDP show that SPID achieves superior knowledge erasure compared to state-of-the-art baselines while maintaining model utility. Ablations and analyses further validate each design component.

2 Preliminaries

Notation. Let $\mathcal{V}$ denote a token vocabulary. An LLM, parameterized by θ, processes a prompt $x \in \mathcal{V}^*$ and generates a response $y \in \mathcal{V}^*$ of length $|y|$ in an auto-regressive manner. At each decoding step $i \in \{1, \ldots, |y|\}$, the model produces a conditional distribution $P_\theta(\cdot|x, y_{<i})$, where $y_{<i}$ represents

the prefix tokens generated prior to step i. The joint likelihood of the response y given x is defined as the product of these step-wise probabilities: $P_\theta(y|x) = \prod_{i=1}^{|y|} P_\theta(y_i|x, y_{<i})$.

Unlearning Objective. LLM unlearning aims to post-hoc adjust the parameters of a pre-trained model θ_0 to eliminate specific undesirable knowledge while preserving its general capabilities (Chen et al., 2024). This task is typically defined over two primary datasets:

- **Forget Set ($\mathcal{D}_f$):** A set of data instances representing the specific knowledge, behaviors, or information to be erased. The objective is to ensure that the updated model no longer generates these responses or exhibits the associated proficiency.
- **Retain Set ($\mathcal{D}_r$):** A dataset consisting of general-domain knowledge or specific instances that the model must preserve to maintain its utility and reasoning capacity.

Most existing unlearning methods formulate this task as a multi-objective optimization problem, seeking a parameter set θ that minimizes a composite loss function:

$$\min_\theta \underbrace{\mathbb{E}_{\mathcal{D}_f}\left[\mathcal{L}_f(\theta; D_f)\right]}_{\text{Forget loss } \mathcal{L}_f} + \lambda \underbrace{\mathbb{E}_{\mathcal{D}_r}\left[\mathcal{L}_r(\theta; D_r)\right]}_{\text{Retain loss } \mathcal{L}_r}, \tag{1}$$

where $\mathcal{L}_f$ reduces proficiency on the forget set, $\mathcal{L}_r$ penalizes deviations from the original model on the retain set, and λ controls the trade-off between forgetting and retention (Yuan et al., 2025).

Additional related work and existing methods are provided in Appendices A and C, respectively.

3 LLM Unlearning: A Causal-Inspired Lens

To decouple sensitive concepts from general utility within the LLM's internal representations, we study LLM unlearning through a *causal-inspired lens*. Previous work interprets unlearning as a deconfounding problem at the input-output level (Liu et al., 2024). We extend this view to the *internal semantic space*, focusing on the causal paths for undesirable concept expression in LLM generation.

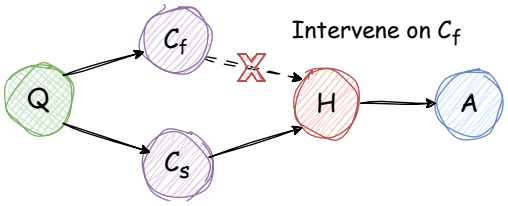


Figure 2: Causal graph $\mathcal{G}$ for LLM unlearning.

Causal Graph of Unlearning. The generation process is represented by a directed acyclic graph $\mathcal{G}$, as illustrated in Figure 2. We define the following variables:

- Q: The input context and query provided to the LLM.
- C: The latent semantic concepts activated by the query, decomposed into disjoint forget concepts $\mathcal{C}_f$ (sensitive knowledge) and safe concepts $\mathcal{C}_s$ (non-sensitive knowledge).
- H: The activated hidden states and parameters where latent knowledge is represented.
- A: The LLM-generated answer.

Given a query Q to the LLM, the standard inference follows the path $Q \to \{\mathcal{C}_f, \mathcal{C}_s\} \to H \to A$. The hidden state H is a function of both the query and the activated concepts, i.e., $H = f_\theta(Q, \mathcal{C}_f, \mathcal{C}_s)$. In a pre-trained model θ_0, $\mathcal{C}_f$ and $\mathcal{C}_s$ are deeply entangled within the representation space H, explaining why lexical-level token suppression fails to eliminate the influence of sensitive information.

Viewing Unlearning as Path Intervention. We view LLM unlearning as a path intervention on the causal graph: updating θ to eliminate the influence of the forget concept $\mathcal{C}_f$ on the hidden state H, while preserving the safe path $\mathcal{C}_s \to H$. Formally, let $H_\theta(Q, \mathcal{C}_f, \mathcal{C}_s)$ denote the model representation, and let $\mathcal{C}_f \nrightarrow H$ denote severing the dependency of H on $\mathcal{C}_f$. The ideal unlearning goal is:

$$P_\theta(A|Q) = P_{\theta_0}(A|\, Q,\ \mathcal{C}_f \nrightarrow H, \mathcal{C}_s). \tag{2}$$

By focusing on this specific causal path, we aim to decouple the undesirable semantic residues from the model's core utility. This goal requires the unlearned model θ to produce answers A based solely on the safe concept $\mathcal{C}_s$, even when Q is semantically related to $\mathcal{C}_f$.

This perspective also explains the nature of the knowledge residues observed in Figure 1b. Since $\mathcal{C}_f$ manifests as a semantic manifold, methods that suppress specific tokens only block a narrow subset

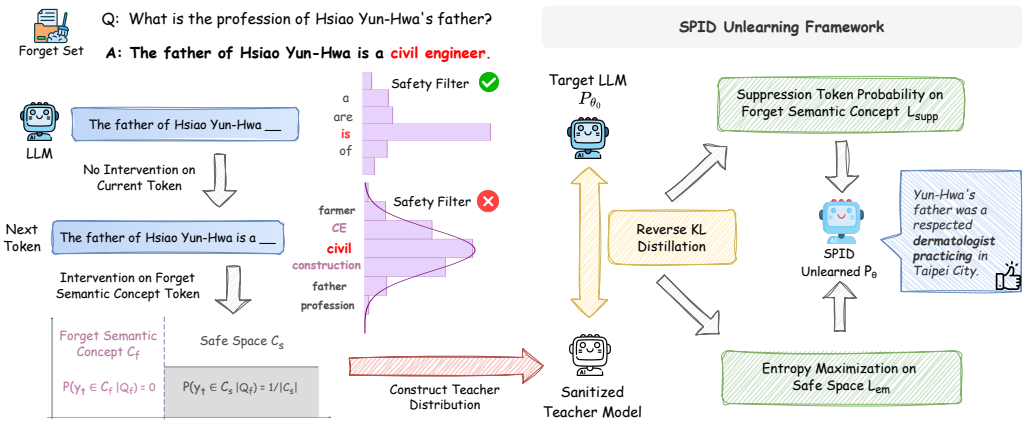


Figure 3: Overview of our SPID framework: constructing a sanitized teacher model (Left) and deriving the SPID objective from reverse-KL distillation (Right).

of paths, leaving the dependency $\mathcal{C}_f \to H$ intact for synonymous expressions. This insight suggests that faithful unlearning requires a semantic-aware intervention that targets the entire concept cluster rather than discrete symbols.

Based on this causal-inspired view, we develop a semantic-aware unlearning method in Section 4. Furthermore, Section 5 introduces a new metric designed to quantify such knowledge residues.

4 Semantic-Aware Unlearning via Path Intervention Distillation

Motivated by the observed knowledge residues and the causal graph in Section 3, we propose **SPID** (Semantic-aware Path Intervention Distillation) for LLM unlearning, as shown in Figure 3. SPID enforces semantic erasure by suppressing the entire semantic cluster of the forget concept. It consists of two components: constructing a sanitized teacher P_s and optimizing the student θ via an objective derived from reverse-KL distillation.

4.1 Sanitized Teacher Construction

Inspired by the path intervention $\mathcal{C}_f \nrightarrow H$, SPID constructs a sanitized teacher P_s as its counterpart at the output level. Specifically, this construction involves two distinct steps: defining the forget semantic cluster $\mathcal{C}_f$ to be zero-forced, and establishing a maximum-entropy distribution over the remaining safe subspace $\mathcal{C}_s$.

SPID operates in a label-free manner. We assume that the model's own highest-confidence predictions on forget prompts serve as a proxy for its latent knowledge. For a given context $c_t = (x, w_{<t})$ where $x \in \mathcal{D}_f$, we identify the semantic anchor $\hat{w}_t$ as follows:

$$\hat{w}_t = \operatorname*{argmax}_{w \in \mathcal{V}} P_\theta(w \mid c_t). \tag{3}$$

To achieve semantic-level erasure and prevent leakage, we perform dynamic semantic clustering in the output embedding space $\mathbf{E} \in \mathbb{R}^{|\mathcal{V}| \times d}$, where semantic relations are encoded by geometric proximity (Ethayarajh, 2019; Khandelwal et al., 2020). Since LLMs represent concepts as high-dimensional clusters rather than discrete tokens, suppressing one token can simply shift probability mass to nearby semantic alternatives.

By calculating the cosine similarity between candidate embeddings and the anchor $\hat{w}_t$, we identify a forget semantic cluster $\mathcal{C}_f$ that covers the conceptual manifold around the anchor:

$$\mathcal{C}_f = \{w \in \mathcal{V} \mid \mathrm{sim}_{\cos}(\mathbf{E}_w, \mathbf{E}_{\hat{w}_t}) > \delta\}, \quad \mathcal{C}_s = \mathcal{V} \setminus \mathcal{C}_f, \tag{4}$$

where δ is the similarity threshold, and $\mathbf{E}_w$ and $\mathbf{E}_{\hat{w}_t}$ are the embeddings of token w and anchor $\hat{w}_t$, respectively. This clustering targets not only the token itself but also its synonyms and paraphrases.

Definition 4.1 (Sanitized Teacher Family)**.** For each context x, define $P_s(\cdot \mid x)$ as any distribution supported on $\mathcal{C}_s$:

$$P_s(w \mid x) = 0,\ \forall w \in \mathcal{C}_f, \qquad \sum_{w \in \mathcal{C}_s} P_s(w \mid x) = 1.$$

Following the Principle of Maximum Entropy (Jaynes, 1957), we define the sanitized teacher P_s as a uniform distribution over the safe subspace $\mathcal{C}_s$:

$$P_s(w \mid c_t) = \begin{cases} 0, & w \in \mathcal{C}_f; \\ \frac{1}{|\mathcal{C}_s|}, & w \in \mathcal{C}_s. \end{cases} \tag{5}$$

To preserve linguistic utility, we incorporate safety gating using a stop-word–based filter on the semantic anchor. Specifically, if the predicted token $\hat{w}_t$ is deemed functional (e.g., 'is', 'the'), we bypass the unlearning update at step t. For computational efficiency, we further approximate the objective by restricting computation to the Top-K token subspace. This teacher serves as a reference to guide the student θ toward an ignorant representational state while preserving linguistic utility.

4.2 Deriving the SPID Objective from Reverse KL Distillation

Given the sanitized teacher P_s, we use reverse-KL self-distillation for unlearning. Unlike forward KL, it provides a "zero-forcing" effect that suppresses sensitive tokens in the forget concept $\mathcal{C}_f$ while preserving the distribution over the safe concept $\mathcal{C}_s$ for utility (Wang et al., 2025a; Wu et al., 2025). We therefore decompose it into a forget-concept suppression loss and a safe-subspace entropy maximization loss.

Formally, the ideal unlearning objective between the teacher model P_s and the student model P_θ is:

$$\mathcal{L}(\theta; x) := D_{\mathrm{KL}}(P_\theta(\cdot \mid x) \parallel P_s(\cdot \mid x)). \tag{6}$$

However, directly optimizing this objective with gradients is intractable due to the hard constraints in the sanitized teacher distribution, which induce numerical singularities:

Theorem 4.2 (Numerical Singularity Induced by Support Constraint)**.** *Let P_s be the maximum-entropy sanitized teacher, defined as uniform over $\mathcal{C}_s$ and zero on $\mathcal{C}_f$. For any student model θ and input $x \in \mathcal{D}_f$, the reverse-KL objective yields:*

$$\mathcal{L}(\theta; x) = \begin{cases} +\infty, & \text{if } P_\theta(\mathcal{C}_f \mid x) > 0; \\ \log |\mathcal{C}_s| - H(P_\theta(\cdot \mid x)), & \text{if } P_\theta(\mathcal{C}_f \mid x) = 0. \end{cases} \tag{7}$$

Theorem 4.2 directly implies the following interpretation of the unlearning task:

Corollary 4.3 (Unlearning as Constrained Entropy Maximization)**.** *Minimizing the reverse KL divergence $\mathcal{L}(\theta; x)$ is equivalent to the constrained optimization:*

$$\max_\theta\ H(P_\theta(\cdot \mid x)), \quad \textit{s.t.} \quad P_\theta(\mathcal{C}_f \mid x) = 0. \tag{8}$$

This result decomposes the unlearning task into two distinct functional requirements: (1) Total Suppression, which demands strictly zero probability mass on the forget cluster $\mathcal{C}_f$, and (2) Entropy Maximization, which encourages a uniform distribution over the safe subspace $\mathcal{C}_s$. The hard constraint $P_\theta(\mathcal{C}_f \mid x) = 0$ creates a barrier for optimization: standard gradient descent cannot traverse regions where the loss is infinite or undefined, necessitating a relaxed objective.

To facilitate stable gradient-based optimization while respecting the dual forces of reverse KL (zero-forcing on $\mathcal{C}_f$ and entropy maximization on $\mathcal{C}_s$), we relax the hard constraints of Corollary 4.3. Specifically, we propose a tractable surrogate objective $\mathcal{L}_{\text{SPID}}$ that achieves this by decomposing the divergence into two differentiable terms:

Suppression Loss ($\mathcal{L}_{supp}$): We approximate the zero-forcing constraint by minimizing the total probability mass assigned to the forget semantic concept $\mathcal{C}_f$. This soft penalty drives the probabilities of semantically similar tokens toward zero without numerical collapse:

$$\mathcal{L}_{supp}(\theta; x) = \sum_{w \in \mathcal{C}_f} P_\theta(w \mid x). \tag{9}$$

Geometrically, it penalizes the projection of H onto the sensitive concept subspace (Lemma E.1).

Entropy Maximization Loss ($\mathcal{L}_{em}$): To match the entropy-maximizing term in Theorem 4.2, we maximize the uncertainty over safe tokens by minimizing the negative entropy of the renormalized safe distribution $\hat{P}_\theta$:

$$\mathcal{L}_{em}(\theta; x) = \sum_{w \in \mathcal{C}_s} \hat{P}_\theta(w \mid x) \log \hat{P}_\theta(w \mid x). \tag{10}$$

Total Objective: Integrating $\mathcal{L}_{supp}$ and $\mathcal{L}_{em}$ with a utility retention loss $\mathcal{L}_r$, we obtain the final objective:

$$\mathcal{L}_{\text{SPID}} = \mathbb{E}_{x \sim \mathcal{D}_f}[\mathcal{L}_{\text{supp}} + \lambda_1 \mathcal{L}_{\text{em}}] + \lambda_2 \mathbb{E}_{x \sim \mathcal{D}_r} \mathcal{L}_r, \tag{11}$$

where $\lambda_1 > 0$ balances suppressing sensitive clusters with maximizing uncertainty over the safe concept, thereby avoiding collapse into anti-facts; $\lambda_2 > 0$ controls the regularization trade-off between unlearning on the forget set $\mathcal{D}_f$ and utility preservation on the retain set $\mathcal{D}_r$.

Optimizing $\mathcal{L}_{\text{SPID}}$ reduces the projection of internal representations onto the sensitive concept subspace, approximating the path-specific intervention $\mathcal{C}_f \nrightarrow H$. This weakens the hidden-state dependency on undesirable concepts (proofs in Appendix E; empirical evidence in Appendix G.6).

5 Semantic Residual Knowledge Metric

To quantify knowledge erasure beyond lexical behavior, we introduce the Semantic Residual Knowledge (**SRK**) framework. Unlike lexical metrics such as ROUGE, SRK evaluates semantic representations, providing a more faithful measure of whether sensitive knowledge is truly removed.

5.1 Definition of Semantic Residual Knowledge

To make the degree of forgotten knowledge measurable, SRK is grounded in established *semantic-aware* evaluation paradigms. Concretely, we adapt entailment-based semantic matching and semantic-equivalence aggregation to the unlearning setting (Liu et al., 2025), and further incorporate sampling-and-clustering strategies for semantic estimation (Farquhar et al., 2024; Dai et al., 2025). Building on these foundations, SRK quantifies fluctuations in residual knowledge and reliably separates genuine conceptual erasure from superficial, lexical-level avoidance.

We employ a parameterized Natural Language Inference (NLI) model Φ as a semantic kernel to quantify residual knowledge in the unlearned LLM θ. Given a generated response y and a "forgotten" reference y^*, we define the soft entailment score $S_\Phi(y, y^*) \coloneqq P_\Phi(\text{entailment} \mid y, y^*)$, which measures the probability that y semantically entails the target forget concept.

SRK: Measuring Semantic Residual. We define SRK as the confidence-weighted expected knowledge residual under the model output distribution $P_\theta(y|x)$. The theoretical SRK is:

$$\text{SRK}_\theta(x, y^*) = \mathbb{E}_{y \sim P_\theta(\cdot|x)}\left[S_\Phi(y, y^*)\right], \tag{12}$$

where $x \in D_f$ is a forget set query. For tractability, we approximate the expectation with importance sampling, following semantic entropy (Farquhar et al., 2024). Given K stochastic samples $\{y_k\}_{k=1}^K$, we compute normalized confidence weights $w_k = \frac{\exp(\ell_k/\tau)}{\sum j{=}1^K \exp(\ell_j/\tau)}$, where τ is the temperature and $\ell_k \coloneqq \log P_\theta(y_k|x)$. The empirical SRK estimator is:

$$\widehat{\text{SRK}}(x, y^*) = \sum_{k=1}^{K} w_k \cdot S_\Phi(y_k, y^*). \tag{13}$$

SRK captures both the frequency and confidence of residual knowledge. A high SRK indicates persistent forget-concept generation, suggesting unresolved functional dependency. Algorithm details are provided in Appendix H.2.

5.2 Faithfulness Analysis of SRK

5.2.1 Evaluation Setup

To validate SRK, we first define criteria for faithful unlearning evaluation, then describe our experimental setup, mainly following OPEN-UNLEARNING (Dorna et al., 2026).

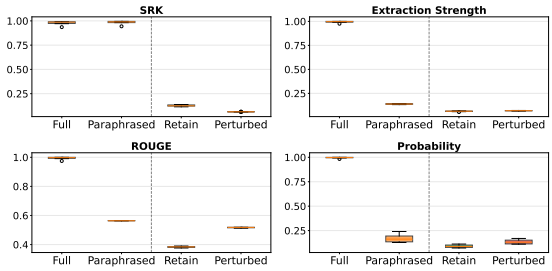


Figure 4: Box plots of multiple metrics across pools under different hyperparameter settings (dashed line separates Positive Pool $\mathcal{P}$ and Negative Pool $\mathcal{N}$).

Table 1: Statistical significance tests across multiple metrics: within Positive Pool $\mathcal{P}$ and between $\mathcal{P}$ and Negative Pool $\mathcal{N}$.

Metric	Within $\mathcal{P}$		Between $\mathcal{P}$ & $\mathcal{N}$
	Wilcoxon Test P_{w_p}	TOST Test P_{t_p}	Wilcoxon Test $P_{w_{pn}}$
SRK	**0.1822**	**0.0000**	6.55e-08
Extraction Strength	0.0002	1.0000	6.61e-08
Exact Memorization	0.0002	1.0000	6.75e-08
Truth Ratio	0.0002	1.0000	6.80e-08
ROUGE	0.0002	1.0000	6.33e-08
Probability	0.0002	1.0000	2.36e-06
Para.Prob	0.0376	0.8659	8.60e-06

Criteria for Faithful Evaluation. We argue that a faithful unlearning metric must satisfy two fundamental properties:

1. **Semantic Invariance**: The metric should identify the target concept across linguistic variants (e.g., paraphrases), ensuring measurement on the semantic rather than the lexical level.
2. **Knowledge Discrimination**: The metric must effectively distinguish genuine knowledge retention from incorrect outputs, such as factual perturbations.

Evaluation Setup. To assess these properties, we utilize the TOFU benchmark with the `forget10` task (Maini et al., 2024). Following the OPEN-UNLEARNING protocol, we construct two distinct model pools representing known ground states of knowledge:

- **Positive Pool** ($\mathcal{P}$): Models guaranteed to retain the target knowledge $\mathcal{C}_f$. This includes (i) *Full* models trained on the complete dataset, and (ii) *Paraphrased* models trained on semantically equivalent but linguistically distinct variants of the `forget10` subset.
- **Negative Pool** ($\mathcal{N}$): Models guaranteed to lack the target knowledge $\mathcal{C}_f$. This includes (i) *Retain* models trained exclusively on the `retain90` split, and (ii) *Perturbed* models where the `forget10` subset is replaced with factually incorrect variations.

A faithful metric should yield high and stable scores within different settings in the Positive Pool $\mathcal{P}$, demonstrating semantic invariance. Simultaneously, it must exhibit a significant gap between the scores of $\mathcal{P}$ and $\mathcal{N}$ to prove its discriminative power.

We use GPT-5 to generate paraphrased and perturbed variants. We fine-tune `Llama-3.2-3B` on four configurations (*Full*, *Paraphrased*, *Retain*, *Perturbed*), with 10 instances each, and benchmark SRK against six established metrics. Dataset-generation prompts, full results, extra 1B-model experiments, and experimental settings are in Appendix F; baseline metric definitions are in Appendix D.

5.2.2 RESULT ANALYSIS

To assess metric reliability, Figure 4 shows score distributions, and Table 1 reports statistical tests. We use the Wilcoxon rank-sum test (Wilcoxon, 1945; Mann & Whitney, 1947) for distributional differences and TOST (Schuirmann, 1987) for equivalence. Within $\mathcal{P}$, a *high* Wilcoxon p-value indicates no significant semantic shift, while a *low* TOST p-value confirms equivalence. Between $\mathcal{P}$ and $\mathcal{N}$, a *low* Wilcoxon p-value indicates significant discrimination.

Semantic Invariance. Visually, Figure 4 shows that SRK stays stable from *Full* to *Paraphrased*, while lexical metrics such as *ROUGE* and *Extraction Strength* drop sharply despite unchanged knowledge. Table 1 confirms this pattern: SRK shows no significant Wilcoxon difference ($p_{w_p} = 0.1822$) and passes TOST equivalence ($p_{t_p} \approx 0.0000$), indicating paraphrase invariance. By contrast, baseline metrics exhibit significant within-$\mathcal{P}$ shifts ($p_{w_p} = 0.0002$) and fail equivalence testing ($p_{t_p} = 1.0000$). Thus, prior metrics may mistake linguistic variation for successful unlearning, yielding false positives.

Knowledge Discrimination. Figure 4 shows a clear SRK gap between the positive pool $\mathcal{P}$ and negative pool $\mathcal{N}$. Table 1 confirms this with a highly significant Wilcoxon separation ($p_{w_{pn}} = 6.55 \times 10^{-8}$), showing that SRK distinguishes conceptual retention from factual perturbation.

Collectively, SRK better reflects the underlying semantic belief distribution H, while standard metrics become statistically unstable under syntactic perturbations. Further robustness tests confirm these conclusions across NLI judges of different scales and families, sampling temperatures, and sample counts K (Appendix F.4). We thus adopt SRK in subsequent experiments.

Table 2: Performance of unlearning baselines on TOFU with LLaMA-3.1-8B. ↑/↓ indicates that higher/lower values are better. For each unlearning setting, the best results are in **bold**, the second best are underlined, and the two worst are shaded in red. SRK_f and SRK_r are computed on the forget set and retain set. Additional results on LLaMA-3.2-3B/1B are provided in Table 7, Appendix G.

Method	Forget 1%				Forget 5%				Forget 10%			
	$SRK_f \downarrow$	Mem. ↑	$SRK_r \uparrow$	M.U. ↑	$SRK_f \downarrow$	Mem. ↑	$SRK_r \uparrow$	M.U. ↑	$SRK_f \downarrow$	Mem. ↑	$SRK_r \uparrow$	M.U. ↑
Target Model	0.9736	0.0023	0.9626	0.6276	0.9573	0.0082	0.9633	0.6276	0.9679	0.0061	0.9587	0.6276
Retain Model	0.1564	0.5332	0.9666	0.6177	0.1479	0.5671	0.9603	0.6323	0.1359	0.5731	0.9583	0.6461
GA	0.3351	0.6527	0.3492	0.4773	0.2294	0.7550	0.2005	0.0000	0.2282	0.7861	0.2005	0.0000
GD	0.3645	0.6435	0.3634	0.4879	0.1741	0.7720	0.3940	0.2362	**0.0551**	**0.8794**	0.1241	0.5588
WGA	0.2653	0.4095	0.7497	0.6535	0.1968	0.6774	0.4355	0.5899	0.4538	0.7850	0.3560	0.5661
NPO	0.1907	0.5093	0.5967	0.6355	0.1360	0.5424	0.2985	0.6097	0.1445	0.5158	0.3346	0.5620
SimNPO	0.4934	0.2252	**0.8290**	0.6503	0.2574	0.3712	0.5947	0.5952	0.2166	0.3928	0.5153	0.6040
DPO	**0.1089**	0.2692	0.3449	0.6403	**0.1111**	0.2969	0.2210	0.5036	0.2521	0.1776	0.4129	0.6159
UnDIAL	0.3342	0.5624	0.5047	0.6751	0.2904	0.8103	0.3589	0.6618	0.2996	0.5673	0.6254	0.6324
DUET	0.5769	0.0983	0.7631	0.5938	0.6274	0.0966	0.7335	0.6114	0.6099	0.0975	0.7057	0.5553
BS-S	0.3947	**0.7673**	0.4401	0.6720	0.5073	**0.8661**	0.3404	**0.6757**	0.2021	0.8605	0.3258	0.6279
PMC	0.2409	0.2550	0.4767	0.6484	0.2864	0.3349	0.4954	0.6282	0.2287	0.4022	0.7326	0.5806
SPID (Ours)	0.1592	0.5190	0.7178	**0.6893**	0.2723	0.8047	**0.7748**	0.6354	0.1978	0.7928	**0.8783**	**0.6449**

6 EXPERIMENT

6.1 EXPERIMENT SETUP

Benchmarks and Target Models. We evaluate SPID on two widely recognized unlearning benchmarks: TOFU (Maini et al., 2024) and WMDP (Li et al., 2024). TOFU comprises 200 fictitious author profiles, offering a controlled setting to assess unlearning across three granularity levels: Forget01 (1%), Forget05 (5%), and Forget10 (10%). Conversely, WMDP focuses on unlearning hazardous knowledge, specifically within the biosecurity and cyber domains. Adhering to standard benchmark configurations, we employ the Llama 3 family (Grattafiori et al., 2024) for TOFU and Zephyr-7Bβ (Tunstall et al., 2024) for WMDP.

Baselines. We compare SPID with a diverse set of baselines spanning several unlearning paradigms. These include gradient-based methods (GA, GD, WGA) (Maini et al., 2024; Wang et al., 2025c), and preference optimization approaches (NPO, SimNPO, DPO) (Zhang et al., 2024; Fan et al., 2024). We also evaluate self-distillation methods (UnDIAL (Dong et al., 2025), DUET (Zhong et al., 2026)) and recent leakage-oriented baselines: BS-S (Li et al., 2026b), which targets belief shift through bootstrapped sequences, and PMC (Scholten et al., 2026), which targets robustness to stochastic decoding. For WMDP, we additionally report RMU (Li et al., 2024).

Evaluation Metrics. We evaluate unlearning efficacy and utility preservation with benchmark-specific metrics. For TOFU, we report SRK on forget and retain sets, plus *Memorization (Mem.)*, the harmonic mean of ES, EM, Truth Ratio, and Paraphrased Probability (Dorna et al., 2026), and *Model Utility (M.U.)*, the harmonic mean of nine retain-, real-author-, and world-knowledge metrics (Maini et al., 2024). For WMDP, where SRK is inapplicable due to the multiple-choice format, we use hazardous-knowledge QA accuracy for unlearning and MMLU for general capability.

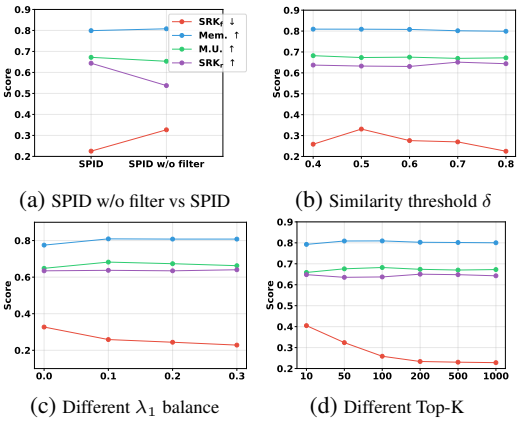
We adopt the standard configurations for each benchmark. Further details on experiment setup, additional results, and analysis are deferred to Appendix G.

6.2 EXPERIMENT RESULTS

TOFU Results. Table 2 demonstrates that SPID achieves a superior trade-off between deep semantic erasure and utility preservation, effectively addressing the limitations of lexical-level suppression. First, aligning with our observations on semantic leakage, methods like GD appear effective on lexical-level Memorization (Mem.) benchmarks (e.g., on *Forget01*) yet fail to reduce SRK_f, indicating that the underlying semantic knowledge remains latent rather than erased. In contrast, SPID consistently minimizes both metrics across all unlearning ratios, confirming semantic forgetting. This holds for complex multi-token facts as well, with no measurable gap between short- and long-span knowledge (Appendix G.5). We also observe a critical discrepancy in baselines: while they maintain Model Utility (M.U.) at levels comparable to the Target Model, they suffer a sharp decline in Semantic Retention (SRK_r). Since M.U. primarily relies on greedy sampling and lexical matching, this gap

Table 3: Comparison with baselines on WMDP with retain regularization. ↑ / ↓ indicates larger / smaller values are preferable.

Method	Unlearn		Retain
	Bio ↓	Cyber ↓	MMLU ↑
Original	0.6465	0.4438	0.5852
GD	0.2739	0.2657	0.4265
WGA	0.2590	0.2989	0.4806
NPO	0.2647	0.3067	0.4434
SimNPO	0.2617	0.3163	0.4453
RMU	0.2537	0.2617	0.4967
BS-S	**0.248**2	**0.2431**	0.3574
SPID (ours)	0.2628	0.2748	**0.5305**



(a) SPID w/o filter vs SPID (b) Similarity threshold δ

(c) Different λ_1 balance (d) Different Top-K

Figure 5: Results for SPID sensitivity analysis.

suggests that baselines merely preserve lexical patterns of the retain set, while the internal semantic knowledge has already been disrupted. SPID, on the contrary, sustains high scores on both metrics, proving that it preserves the structural integrity of knowledge beyond mere lexical retention.

WMDP Results. Table 3 reports the results on the WMDP benchmarks. As SPID operates directly on the semantic concept, the embedding clustering effectively distinguishes and cuts off hazardous information. Consequently, our method realizes an improved trade-off between unlearning and utility (MMLU), despite the shift from TOFU's QA pairs to raw corpora. This validates the robustness of SPID across input formats and its capability for semantic unlearning.

6.3 Further Analysis

Figure 5 reports sensitivity and component analyses of SPID on TOFU Forget10 on 3B model.

Effectiveness of the Filtering Mechanism. Figure 5a validates the importance of our dynamic filtering. Compared with "SPID w/o filter", we observe a significant performance gap. Removing the filter results in a sharp increase in SRK_f (from 0.22 to 0.33), indicating incomplete erasure of the target knowledge. More importantly, SRK_r drops from 0.64 to 0.53, showing more harm to retained knowledge. This confirms that the filtering is essential for disentangling the forget set from general knowledge, ensuring that the unlearning process is both effective and precise.

Hyperparameter Sensitivity Analysis:

- **Similarity Threshold δ (Figure 5b):** SPID is stable for $\delta \in [0.4, 0.8]$, with utility (Mem., M.U., SRK_r) largely unchanged. A stricter threshold (e.g., $\delta = 0.8$) yields the lowest SRK_f, indicating more precise forgetting.
- **Effect of Max Entropy λ_1 (Figure 5c):** Increasing λ_1 not only enhances the unlearning efficacy (SRK_f, and Mem.) but also leads to a corresponding improvement on utility (SRK_r, and M.U.). This highlights the critical role of the maximum entropy objective.
- **Top-K Retrieval (Figure 5d):** As K increases from 10 to 200, SRK_f drops markedly and then saturates beyond $K = 200$, suggesting that adequate context is beneficial while excessively large neighborhoods are not required.

7 Conclusion

In this work, we identified semantic leakage as a critical limitation in current LLM unlearning. To address this, we propose SPID, a causal-inspired framework that unlearns at the concept level as a structural path intervention. By leveraging a stable reverse-KL objective with embedding-space semantic clustering, SPID effectively suppresses the entire semantic cluster of each forget concept. Additionally, to overcome the limitations of lexical-level evaluations, we introduce SRK, which faithfully quantifies knowledge leakage. Extensive experiments on benchmarks demonstrate that SPID outperforms existing baselines, achieving a better trade-off between thorough knowledge erasure and the preservation of general model utility.

AI USE STATEMENT

We used LLM tools to assist parts of this work. GPT was used for writing assistance and language polishing. Codex was used to assist coding, debugging, and repository-level editing. In addition, we used the GPT-5 API to generate the paraphrased and perturbed variants used in the faithfulness evaluation dataset in Section 5, and as an LLM-based judge to extract minimal factual spans for the knowledge-complexity analysis in Appendix G.5. All LLM-assisted outputs were reviewed and edited by the authors, and we take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work aims to mitigate privacy breaches and the unauthorized disclosure of protected or hazardous information by removing such knowledge from LLMs at the semantic level (SPID) and by auditing whether it remains accessible under paraphrasing or stochastic decoding (SRK). Our experiments are conducted on publicly available unlearning benchmarks (TOFU and WMDP) and involve neither crowdsourcing nor human subjects. Unlearning methods also carry dual-use risks, such as the removal of safety-relevant knowledge; practical deployment should therefore pair unlearning with access control, audit logs, retained safety evaluations, misuse monitoring, and human oversight. We view SPID as a risk-reduction tool rather than a standalone guarantee of legal or safety compliance.

REPRODUCIBILITY STATEMENT

The causal-inspired view and the theoretical results underlying SPID are presented in Sections 3 and 4, with complete proofs and assumptions in Appendix E. Details of the evaluation metrics and their faithfulness evaluation are given in Section 5 and Appendix D and F. The experimental setup, including benchmarks, baselines, hyperparameters, and compute resources, is described in Section 6 and Appendix G, and pseudocode for both SPID and SRK is provided in Appendix H. Our code is available in an anonymized repository at `https://anonymous.4open.science/r/SPID-0DC9`.

REFERENCES

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

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pp. 463–480. IEEE, 2015.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Kongyang Chen, Zixin Wang, Bing Mi, Waixi Liu, Shaowei Wang, Xiaojun Ren, and Jiaxing Shen. Machine unlearning in large language models. *arXiv preprint arXiv:2404.16841*, 2024.

Lu Dai, Yijie Xu, Jinhui Ye, Hao Liu, and Hui Xiong. Seper: Measure retrieval utility through the lens of semantic perplexity reduction. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=ixMBnOhFGd`.

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

Vineeth Dorna, Anmol Reddy Mekala, Wenlong Zhao, Andrew McCallum, J Zico Kolter, Zachary Chase Lipton, and Pratyush Maini. Openunlearning: Accelerating LLM unlearning via unified benchmarking of methods and metrics. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026. URL `https://openreview.net/forum?id=Gy67Zh5X1i`.

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

Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 55–65, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1006. URL https://aclanthology.org/D19-1006/.

Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Simplicity prevails: Rethinking negative preference optimization for llm unlearning. *arXiv preprint arXiv:2410.07163*, 2024.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, June 2024. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-024-07421-0. URL https://www.nature.com/articles/s41586-024-07421-0. Publisher: Nature Publishing Group.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=XPZIaotutsD.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106:620–630, May 1957. doi: 10.1103/PhysRev.106.620. URL https://link.aps.org/doi/10.1103/PhysRev.106.620.

Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. Copyright violations and large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7403–7412, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.458. URL https://aclanthology.org/2023.emnlp-main.458/.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HklBjCEKvH.

Junyi Li, Yongqiang Chen, and Ningning Ding. CiPO: Counterfactual unlearning for large reasoning models through iterative preference optimization. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3152–3170, San Diego, California, United States, July 2026a. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.143. URL https://aclanthology.org/2026.acl-long.143/.

Kemou Li, Qizhou Wang, Yue Wang, Fengpeng Li, Jun Liu, Bo Han, and Jiantao Zhou. LLM unlearning with LLM beliefs. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=qCfYOLAzti.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew Bo Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Ariel Herbert-Voss, Cort B Breuer, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam Alfred Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Ian Steneker, David Campbell, Brad Jokubaitis, Steven Basart, Stephen Fitz, Ponnurangam Kumaraguru, Kallol Krishna Karmakar, Uday Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine*

Learning, volume 235 of *Proceedings of Machine Learning Research*, pp. 28525–28550. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/li24bc.html.

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

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, Kush R. Varshney, Mohit Bansal, Sanmi Koyejo, and Yang Liu. Rethinking machine unlearning for large language models. *Nat. Mac. Intell.*, 7(2):181–194, 2025. doi: 10.1038/S42256-025-00985-0. URL https://doi.org/10.1038/s42256-025-00985-0.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Yujian Liu, Yang Zhang, Tommi Jaakkola, and Shiyu Chang. Revisiting who's harry potter: Towards targeted unlearning from a causal intervention perspective. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8708–8731, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.495. URL https://aclanthology.org/2024.emnlp-main.495/.

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

H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947. ISSN 00034851. URL http://www.jstor.org/stable/2236101.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

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

Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *ACM Trans. Intell. Syst. Technol.*, 16(5), September 2025. ISSN 2157-6904. doi: 10.1145/3749987. URL https://doi.org/10.1145/3749987.

OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. doi: 10.48550/ARXIV.2303.08774. URL https://doi.org/10.48550/arXiv.2303.08774.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=HPuSIXJaa9.

Hadi Reisizadeh, Jiajun Ruan, Yiwei Chen, Soumyadeep Pal, Sijia Liu, and Mingyi Hong. Leak@k: Unlearning does not make LLMs forget under probabilistic decoding. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=sREOiWynXf.

Yan Scholten, Stephan Günnemann, and Leo Schwinn. A probabilistic perspective on unlearning and alignment for large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=51WraMid8K.

Yan Scholten, Sophie Xhonneux, Leo Schwinn, and Stephan Günnemann. Model collapse is not a bug but a feature in machine unlearning for LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=1MCQzboBrR.

Donald J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680, December 1987. doi: 10.1007/BF01068419.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. MUSE: Machine unlearning six-way evaluation for language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=TArmA033BU.

Pratiksha Thaker, Shengyuan Hu, Neil Kale, Yash Maurya, Zhiwei Steven Wu, and Virginia Smith. Position: Llm unlearning benchmarks are weak measures of progress. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 520–533. IEEE, 2025.

Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. Unrolling sgd: Understanding factors influencing machine unlearning. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pp. 303–319. IEEE, 2022.

Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. Memorization without overfitting: Analyzing the training dynamics of large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=u3vEuRr08MT.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Lewis Tunstall, Edward Emanuel Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro Von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M Rush, and Thomas Wolf. Zephyr: Direct distillation of LM alignment. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=aKkAwZB6JV.

Stefan Vasilev, Christian Herold, Baohao Liao, Seyyed Hadi Hashemi, Shahram Khadivi, and Christof Monz. Unilogit: Robust machine unlearning for LLMs using uniform-target self-distillation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 22453–22472, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1154. URL https://aclanthology.org/2025.findings-acl.1154/.

Paul Voigt and Axel von dem Bussche. *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Springer Publishing Company, Incorporated, 1st edition, 2017. ISBN 3319579584.

Bichen Wang, Yuzhe Zi, Yixin Sun, Yanyan Zhao, and Bing Qin. Balancing forget quality and model utility: A reverse KL-divergence knowledge distillation approach for better unlearning in LLMs. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1306–1321, Albuquerque, New Mexico, April 2025a. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.60. URL https://aclanthology.org/2025.naacl-long.60/.

Qizhou Wang, Bo Han, Puning Yang, Jianing Zhu, Tongliang Liu, and Masashi Sugiyama. Towards effective evaluations and comparisons for LLM unlearning methods. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=wUtCieKuQU.

Qizhou Wang, Jin Peng Zhou, Zhanke Zhou, Saebyeol Shin, Bo Han, and Kilian Q Weinberger. Rethinking LLM unlearning objectives: A gradient perspective and go beyond. In *The Thirteenth International Conference on Learning Representations*, 2025c. URL https://openreview.net/forum?id=huo8MqVH6t.

Yue Wang, Qizhou Wang, Feng Liu, Wei Huang, Yali Du, Xiaojiang Du, and Bo Han. GRU: Mitigating the trade-off between unlearning and retention for LLMs. In *Forty-second International Conference on Machine Learning*, 2025d. URL https://openreview.net/forum?id=EAjhGr1Oeo.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, December 1945. URL http://www.jstor.org/stable/3001968.

Taiqiang Wu, Chaofan Tao, Jiahao Wang, Runming Yang, Zhe Zhao, and Ngai Wong. Rethinking Kullback-Leibler divergence in knowledge distillation for large language models. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 5737–5755, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.383/.

Nakyeong Yang, Minsung Kim, Seunghyun Yoon, Joongbo Shin, and Kyomin Jung. FaithUn: Toward faithful forgetting in language models by investigating the interconnectedness of knowledge. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12999–13014, Suzhou, China, November 2025a. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.657. URL https://aclanthology.org/2025.emnlp-main.657/.

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

Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=8Dy42ThoNe.

Xiaojian Yuan, Tianyu Pang, Chao Du, Kejiang Chen, Weiming Zhang, and Min Lin. A closer look at machine unlearning for large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Q1MHvGmhyT.

Dawen Zhang, Pamela Finckenberg-Broman, Thong Hoang, Shidong Pan, Zhenchang Xing, Mark Staples, and Xiwei Xu. Right to be forgotten in the era of large language models: Implications, challenges, and solutions. *AI and Ethics*, 5(3):2445–2454, 2025.

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

Yisheng Zhong, Zhengbang Yang, and Zhuangdi Zhu. DUET: Distilled LLM unlearning from an efficiently contextualized teacher. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Xa6QRrXrKX.

A Related Work

We provide a further discussion on the related work of our paper in this section.

LLM Unlearning Machine unlearning aims to remove the influence of designated training data from a trained model while avoiding full retraining (Cao & Yang, 2015; Bourtoule et al., 2021). Compared with classical settings where forgetting targets are well-defined samples, unlearning for large language models (LLMs) must address broader safety and compliance needs (e.g., privacy/copyright and hazardous knowledge) and the fact that LLM knowledge is distributed over parameters and representations as abstract *semantic concepts* rather than isolated records (Eldan & Russinovich, 2023; Shi et al., 2025; Li et al., 2024). Most practical LLM unlearning methods therefore adopt a post-hoc recipe, optimizing a forget loss on a forget set D_f while regularizing with a retain set D_r to preserve general utility (Yao et al., 2024; Maini et al., 2024; Liu et al., 2025; Nguyen et al., 2025).

A dominant line of work formulates unlearning as targeted optimization by directly editing model weights. Early and widely used strategies perform gradient ascent on the negative log-likelihood (or cross-entropy) of forget examples, effectively suppressing the *lexical probability* of specific sequences, and combine it with gradient descent (or KL regularization) on retain data to mitigate utility drop (Yao et al., 2024; Maini et al., 2024; Dorna et al., 2026). Subsequent refinements modulate update strength via token-wise or example-wise weighting, scheduling, or multi-objective shaping to stabilize training and reduce over-forgetting (Yang et al., 2025b; Wang et al., 2025d).

Another family reframes forgetting as preference realignment: the model is encouraged to prefer neutral/refusal responses over generating undesirable content. Methods based on DPO-style preference learning and its negative/inverse variants (e.g., NPO, SimPO, AltPO) provide bounded objectives anchored to a reference policy, which can improve stability relative to unbounded gradient ascent and offer a convenient mechanism to steer generations away from forbidden targets (Rafailov et al., 2023; Zhang et al., 2024; Fan et al., 2024; Mekala et al., 2025; Li et al., 2026a).

Beyond loss design, several methods intervene on internal representations or leverage distillation to achieve more localized edits. Representation-level methods (e.g., steering or misdirection of hidden states) attempt to erase a target concept by directly manipulating the feature space that encodes it (Li et al., 2024). Distillation-based approaches instead construct a modified teacher distribution that removes undesired information, and train a student model to match this "sanitized" behavior, providing a flexible alternative to direct weight editing (Wang et al., 2025a; Vasilev et al., 2025). For instance, UnDIAL (Dong et al., 2025) applies a penalty to the highest-probability tokens, whereas DUET (Zhong et al., 2026) employs a refusal prefix to generate the target distribution.

However, despite the empirical success of these approaches, a critical limitation persists: most current methods predominantly focus on suppressing specific lexical targets (e.g., exact token sequences) while often overlooking the underlying *semantic* persistence. Consequently, models may still exhibit semantic leakage, where the "forgotten" knowledge remains accessible via paraphrasing, synonymous expressions, or probabilistic sampling (Scholten et al., 2025; Reisizadeh et al., 2026; Scholten et al., 2026). PMC (Scholten et al., 2026) addresses a related robustness-to-decoding setting by exploiting partial model collapse to reduce leakage under stochastic generation, but it primarily reshapes the output distribution rather than explicitly erasing the underlying semantic concept. While recent attempts like BS-S (Li et al., 2026b) aim to update model beliefs by bootstrapping diverse output sequences, they primarily operate through data augmentation at the lexical level. Addressing this issue requires moving beyond lexical-level suppression to semantic-level erasure. In this work, we revisit LLM unlearning from a causal-inspired perspective, framing the task as cutting off the dependency between sensitive concepts and model representations. Guided by this view, we propose SPID (Semantic-aware Path Intervention Distillation), which suppresses the entire semantic cluster of the forget concept to achieve faithful semantic unlearning.

LLM Unlearning Evaluation Evaluating the efficacy of unlearning remains a critical challenge, particularly in distinguishing between superficial rephrasing and real semantic erasure. Standard protocols typically assess performance through a dual-objective lens: measuring the suppression of forget set samples (e.g., via Exact Match, ROUGE) while monitoring general capabilities on retain sets (Maini et al., 2024; Shi et al., 2025; Dorna et al., 2026). However, recent studies highlight a profound *semantic gap* in these metrics (Thaker et al., 2025; Reisizadeh et al., 2026). Reliance on

lexical matching often yields false positives, where a model is deemed to have "forgotten" a concept simply because it suppresses specific token sequences, despite retaining the underlying *semantic belief* accessible through paraphrasing, synonymous expressions, or varying decoding strategies (Yang et al., 2025a; Li et al., 2026b). This failure to capture *semantic equivalence* necessitates a shift towards evaluation paradigms that measure conceptual rather than just textual absence. In response to this gap, we introduce Semantic Residual Knowledge (SRK), an entailment-based metric designed to quantify *semantic leakage* under generation uncertainty, providing a faithful assessment of LLM Unlearning.

B NOTATION

Table 4: Summary of frequently used notation.

Symbol	Meaning
θ_0, θ	Parameters of the original model and the unlearned model.
$P_\theta(\cdot \mid x)$	Output distribution of the model parameterized by θ given input or context x.
$\mathcal{D}_f, \mathcal{D}_r$	Forget set and retain set used for unlearning and utility preservation.
Q, H, A	Query, hidden representation, and generated answer in the causal-inspired view of LLM unlearning.
$\mathcal{C}_f, \mathcal{C}_s$	Forget and safe semantic concepts in the causal graph. In SPID, they are operationalized as the context-dependent sensitive token cluster and its safe complement.
c_t, $\hat{w}_t$	Decoding context at step t and the semantic anchor token predicted with highest probability.
$\mathbf{E}_w$, δ	Output embedding of token w and the cosine-similarity threshold for semantic clustering.
Top-K	Candidate token subset used to make semantic clustering computationally efficient.
P_s	Sanitized teacher distribution that assigns zero probability to the forget cluster and maximum entropy over the safe subspace.
$\mathcal{L}_{supp}$, $\mathcal{L}_{em}$, $\mathcal{L}_r$	Suppression loss, entropy-maximization loss, and retain-set regularization loss.
λ_1, λ_2	Weights for entropy maximization and retain-set regularization in SPID.
SRK, Φ, y^*	Semantic Residual Knowledge metric, NLI evaluator, and target semantic string.

C UNLEARNING BASELINE METHODS

We introduce the baselines we compared with in this section.

Gradient Ascent (GA) serves as the foundational method in LLM Unlearning (Thudi et al., 2022). It operates by directly reversing the standard training objective: instead of minimizing the prediction error, GA maximizes the cross-entropy loss on the forget set $\mathcal{D}_f$.

$$\mathcal{L}_{\text{GA}}(\theta; \mathcal{D}_f) = \mathbb{E}_{(x,y)\sim\mathcal{D}_f}[\log P_\theta(y \mid x)]. \tag{14}$$

KL Divergence acts as a standard regularization term to maintain model utility. It constrains the unlearned model P_θ to remain statistically close to the reference model P_{θ_0} on the retain set $\mathcal{D}_r$, thereby ensuring that the model's behavior on non-targeted tasks does not deviate excessively:

$$\mathcal{L}_{\text{KL}}(\theta, D_r) = \mathbb{E}_{x\sim\mathcal{D}_r}\left[\text{KL}(P_\theta(\cdot \mid x) \,\|\, P_{\theta_0}(\cdot \mid x))\right]. \tag{15}$$

In this paper, we use this KL divergence for the retain set regularization.

Gradient Difference (GD) combines gradient ascent on the forget set with a KL-based retention objective to mitigate catastrophic forgetting (Yao et al., 2024):

$$\mathcal{L}_{\text{GD}}(\theta) = \mathcal{L}_{\text{GA}}(\theta; \mathcal{D}_f) + \lambda\mathcal{L}_{\text{KL}}(\theta, D_r) \tag{16}$$

where λ controls the trade-off between effective unlearning and utility preservation.

Weighted Gradient Ascent (WGA) identifies a critical limitation in standard GA: the tendency to disproportionately penalize samples that are already effectively forgotten. To mitigate this,

WGA (Wang et al., 2025c) introduces a dynamic, token-wise weighting scheme based on the model's current confidence. This mechanism ensures that the optimization process remains focused on tokens that still retain strong traces of the targeted information:

$$\mathcal{L}_{\text{WGA}}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}_f}\left[\sum_{i=1}^{|y|} w_i \log P_\theta(y_i \mid x, y_{<i})\right], \tag{17}$$

where $w_i = P_\theta(y_i \mid x, y_{<i})^\alpha$, with α controlling the strength of the counteraction.

Direct Preference Optimization (DPO) represents the alignment framework for preference-based unlearning (Maini et al., 2024; Rafailov et al., 2023). It strategically redefines preference pairs by treating a generic refusal (e.g., "I don't know") as the preferred response (y_{idk}) and the original knowledge as the rejected response (y). This guides the model to explicitly favor ignorance over the forgotten facts:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\frac{1}{\beta}\mathbb{E}_{(x,y)\sim\mathcal{D}_f}\left[\log\sigma\left(\beta\log\frac{P_\theta(y_{\text{idk}} \mid x)}{P_{\theta_0}(y_{\text{idk}} \mid x)} - \beta\log\frac{P_\theta(y \mid x)}{P_{\theta_0}(y \mid x)}\right)\right]. \tag{18}$$

Negative Preference Optimization (NPO) streamlines the preference learning framework by focusing exclusively on the negative signal. Unlike methods requiring explicit pairs, NPO simply suppresses the likelihood of the forget set samples relative to the reference model. This approach has been shown to enhance training stability during unlearning (Zhang et al., 2024):

$$\mathcal{L}_{\text{NPO}}(\theta) = -\frac{1}{\beta}\mathbb{E}_{(x,y)\sim\mathcal{D}_f}\left[\log\sigma\left(-\beta\log\frac{P_\theta(y \mid x)}{P_{\theta_0}(y \mid x)}\right)\right]. \tag{19}$$

SimNPO proposes a memory-efficient evolution of NPO. It removes the computational overhead of maintaining a reference model P_{θ_0} during training. Instead, it employs a constant margin δ within the preference loss, achieving comparable forgetting performance while significantly reducing resource requirements (Fan et al., 2024):

$$\mathcal{L}_{\text{SimNPO}}(\theta) = -\frac{1}{\beta}\mathbb{E}_{(x,y)\sim\mathcal{D}_f}\left[\log\sigma\left(-\frac{\beta}{|y|}\log P_\theta(y \mid x) - \delta\right)\right], \tag{20}$$

where $|y|$ denotes the sequence length of the response, and δ serves as the margin parameter replacing the reference log-probabilities.

UnDIAL tackles the instability often observed in gradient-based methods via a self-distillation mechanism (Dong et al., 2025). It systematically reduces the model's confidence in specific target tokens by adjusting the logits directly, and then minimizes the divergence between the model's output and these synthetic, suppressed logits:

$$\begin{aligned} &z_{\text{adj}}(x) = z_{\theta_0}(x) - \beta\cdot\mathbf{1}_y, \\ &\mathcal{L}_{\text{UnDIAL}}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}_f}\left[\text{KL}\left(\text{softmax}(z_{\text{adj}}(x)) \,\|\, P_\theta(\cdot \mid x)\right)\right], \end{aligned} \tag{21}$$

where $z_{\theta_0}(x)$ represents the reference logits, $\mathbf{1}_y$ is the one-hot vector for the target label, and β controls the intensity of the penalty applied to the forget token.

DUET constructs a "teacher" distribution using in-context learning to guide the unlearning process (Zhong et al., 2026). By prepending a safety prompt x_{ic} (e.g., "You must refuse to answer..."), it transforms the original model's logits into a target distribution that exhibits refusal behavior. The objective is to distill this refusal pattern into the student model by aligning their top-K logits:

$$\begin{aligned} &z_{\text{teach}}(x) = z_{\theta_0}(x_{\text{ic}} \oplus x), \\ &\mathcal{L}_{\text{DUET}}(\theta) = \mathbb{E}_{x\sim\mathcal{D}_f}\left[\sum_{j\in\text{TopK}} \ell(z_\theta(x)_j, z_{\text{teach}}(x)_j)\right], \end{aligned} \tag{22}$$

where $x_{\text{ic}} \oplus x$ denotes the concatenation of the safety prompt and the input, and the loss ℓ (e.g., Huber loss) minimizes the distance between the student and teacher on the most dominant tokens.

Bootstrapping-Sequence (BS-S) targets the model's *global beliefs* to prevent harmful continuations from re-emerging (Li et al., 2026b). Unlike token-level methods, BS-S extends unlearning to the

sequence level by constructing an augmented forget set $\hat{\mathcal{D}}_f = \{(x, \hat{y}^{(j)})\}_{j=1}^{N}$, where $\hat{y}^{(j)} \sim P_\theta(\cdot \mid x)$ are high-confidence generations sampled from the model itself. By including these trajectories, BS-S exposes and suppresses deeper memorization. The final objective balances the original and bootstrapped targets:

$$\mathcal{L}_{\text{BSS}}(\theta) = (1 - \lambda_{\text{BSS}})\mathcal{L}(\theta; \mathcal{D}_f) + \lambda_{\text{BSS}}\mathcal{L}(\theta; \hat{\mathcal{D}}_f), \tag{23}$$

where λ_{BSS} is a hyperparameter and $\mathcal{L}$ can be any base unlearning loss.

Partial Model Collapse (PMC) leverages the model collapse phenomenon for targeted unlearning (Scholten et al., 2026). Instead of optimizing directly on the ground-truth forget answers, PMC samples multiple responses from the current model for each forget question and fine-tunes on a selected self-generated response, thereby inducing a localized collapse of the output distribution on sensitive queries. Given a forget question $q \in \mathcal{D}_f$, PMC first samples candidates $\{\tilde{y}^{(j)}\}_{j=1}^{n} \sim P_\theta(\cdot \mid q)$ and selects one response according to a reward-guided preference rule:

$$\hat{y} \sim \text{BT}_\tau \left(\tilde{y}^{(1)}, \ldots, \tilde{y}^{(n)}; r \right), \tag{24}$$

where $r(\cdot)$ scores the desirability of a sampled response for unlearning and BT_τ denotes a Bradley-Terry-style selection distribution with temperature τ. The resulting objective combines self-generated forget responses with retain-set likelihood preservation:

$$\mathcal{L}_{\text{PMC}}(\theta) = -\mathbb{E}_{q \sim \mathcal{D}_f} \left[\log P_\theta(\hat{y} \mid q) \right] - \lambda \mathbb{E}_{(x,y) \sim \mathcal{D}_r} \left[\log P_\theta(y \mid x) \right], \tag{25}$$

where λ controls the trade-off between targeted collapse on forget queries and utility preservation on retain data.

Representation Misdirection Unlearning (RMU) forces forgetting by steering the model's internal hidden states toward random noise rather than specific tokens (Li et al., 2024). It maps the representations of the forget set $\mathcal{D}_f$ to a random vector u, while regularizing the retain set $\mathcal{D}_r$ to match the original model's latent states:

$$\mathcal{L}_{\text{RMU}}(\theta) = \mathbb{E}_{\mathcal{D}f} \left[|M_\theta(x, y) - \omega \cdot u|_2^2 \right] + \lambda \mathbb{E}_{\mathcal{D}r} \left[|M_\theta(x, y) - M_{\theta_0}(x, y)|_2^2 \right], \tag{26}$$

where M_θ represents intermediate-layer activations, u is a random control vector, and ω is a scaling factor.

D Evaluation Metrics Details

D.1 Meta Metrics

In this section, we introduce the metrics we used in our experiments in Section 5.

1. **Exact Memorization (EM):** This metric evaluates the precision of the model's recall by calculating the proportion of tokens in the ground truth sequence y that are correctly predicted via greedy decoding (Tirumala et al., 2022; Wang et al., 2025b). Formally:

$$\text{EM} = \frac{1}{|y|} \sum_{k=1}^{|y|} \mathbb{1} \left\{ \arg\max_{w \in \mathcal{V}} p_\theta(w \mid x, y_{<k}) = y_k \right\}. \tag{27}$$

2. **Extraction Strength (ES):** ES measures the intensity of memorization by identifying the shortest prefix of y required for the model to reproduce the remaining suffix verbatim (Wang et al., 2025b; Carlini et al., 2021).

$$\text{ES} = 1 - \frac{1}{|y|} \min \{k \mid \text{generate}(x, y_{<k}) = y_{>k}\}. \tag{28}$$

3. **Probability Metrics:** We measure the model's confidence in the ground truth using:
   - **Base Probability:** The likelihood assigned to the canonical answer: $p_\theta(y \mid x)$.
   - **Paraphrased Probability:** The likelihood assigned to a paraphrased answer y_{para} to mitigate template-matching bias: $p_\theta(y_{\text{para}} \mid x)$.

4. **ROUGE:** We utilize ROUGE-L (Lin, 2004) to assess the textual overlap between the generated output $y^*$and the ground truth y.
5. **Truth Ratio:** This metric quantifies the model's discernment by comparing the probability of a correct paraphrased answer (y_{para}) against a plausible but incorrect perturbation (y_{pert}). Following OPEN-UNLEARNING (Dorna et al., 2026), we utilize a normalized formulation to specifically measure knowledge retention:

$$\text{Truth Ratio} = \frac{p_\theta(y_{\text{para}} \mid x)}{p_\theta(y_{\text{para}} \mid x) + p_\theta(y_{\text{pert}} \mid x)}. \tag{29}$$

E THEORETICAL PROOFS AND DETAILS

In this section, we provide the detailed proofs for the numerical singularity of the Reverse-KL objective and its equivalence to constrained entropy maximization, as discussed in Section 4, and formalize the geometric interpretation of the suppression loss $\mathcal{L}_{supp}$.

E.1 PROOF OF THEOREM 4.2

Theorem 4.2 (Numerical Singularity Induced by Support Constraint). Let P_s be the maximum-entropy sanitized teacher, defined as uniform over $\mathcal{C}_s$ and zero on $\mathcal{C}_f$. For any student model θ and input $x \in \mathcal{D}_f$, the reverse-KL objective yields:

$$\mathcal{L}(\theta; x) = \begin{cases} +\infty, & \text{if } P_\theta(\mathcal{C}_f \mid x) > 0; \\ \log|\mathcal{C}_s| - H(P_\theta(\cdot \mid x)), & \text{if } P_\theta(\mathcal{C}_f \mid x) = 0. \end{cases} \tag{30}$$

Proof. Recall the definition of the Kullback-Leibler divergence:

$$D_{KL}(P_\theta(\cdot|x)||P_s(\cdot|x)) = \sum_{w\in\mathcal{V}} P_\theta(w|x) \log \frac{P_\theta(w|x)}{P_s(w|x)}.$$

According to the definition of the sanitized teacher P_s in Section 4:

For $w \in \mathcal{C}_f$, $P_s(w|x) = 0$.For $w \in \mathcal{C}_s$, $P_s(w|x) = \frac{1}{|\mathcal{C}_s|}$ (Uniform distribution).We decompose the summation over the vocabulary $\mathcal{V}$ into the disjoint sets $\mathcal{C}_f$ and $\mathcal{C}_s$:

$$\mathcal{L}(\theta; x) = \underbrace{\sum_{w\in\mathcal{C}_f} P_\theta(w|x) \log \frac{P_\theta(w|x)}{P_s(w|x)}}_{\text{Term 1}} + \underbrace{\sum_{w\in\mathcal{C}_s} P_\theta(w|x) \log \frac{P_\theta(w|x)}{P_s(w|x)}}_{\text{Term 2}}.$$

Case 1: $P_\theta(\mathcal{C}_f|x) > 0$:

If the model assigns any probability mass to the forget concept, there exists at least one token $w' \in \mathcal{C}_f$ such that $P_\theta(w'|x) > 0$. However, $P_s(w'|x) = 0$.Consequently, the term $\log \frac{P_\theta(w'|x)}{0}$ diverges to $+\infty$. Thus, the total divergence is infinite.

Case 2: $P_\theta(\mathcal{C}_f|x) = 0$:

If the model successfully suppresses the forget concept, then $P_\theta(w|x) = 0$ for all $w \in \mathcal{C}_f$. Following the convention $\lim_{p\to 0} p \log p = 0$, Term 1 becomes 0.The total probability mass must sum to 1. Since $P_\theta(\mathcal{C}_f|x) = 0$, it follows that $\sum_{w\in\mathcal{C}_s} P_\theta(w|x) = 1$.We expand Term 2:

$$\begin{aligned} \text{Term 2} &= \sum_{w\in\mathcal{C}_s} P_\theta(w|x) \log \frac{P_\theta(w|x)}{1/|\mathcal{C}_s|} \\ &= \sum_{w\in\mathcal{C}_s} P_\theta(w|x) \left(\log P_\theta(w|x) - \log \frac{1}{|\mathcal{C}_s|} \right) \\ &= \sum_{w\in\mathcal{C}_s} P_\theta(w|x) \log P_\theta(w|x) + \log|\mathcal{C}_s| \sum_{w\in\mathcal{C}_s} P_\theta(w|x). \end{aligned}$$

Substituting the definitions of entropy $H(P) = -\sum P \log P$ and the probability sum of 1:

$$\text{Term 2} = -H(P_\theta(\cdot|x)) + \log |\mathcal{C}_s|.$$

Therefore, $\mathcal{L}(\theta; x) = \log |\mathcal{C}_s| - H(P_\theta(\cdot|x))$. □

E.2 Proof of Corollary 4.3

Corollary 4.3 (Unlearning as Constrained Entropy Maximization). Minimizing the reverse KL divergence $\mathcal{L}(\theta; x)$ is equivalent to the constrained optimization:

$$\max_\theta H(P_\theta(\cdot|x)), \quad \text{s.t.} \quad P_\theta(\mathcal{C}_f|x) = 0.$$

Proof. We seek to minimize the objective derived in Theorem 4.2:

$$\min_\theta \mathcal{L}(\theta; x).$$

To avoid the infinite loss regime identified in Case 1 of Theorem 4.2, the optimization must satisfy the hard constraint:

$$P_\theta(\mathcal{C}_f|x) = 0.$$

This justifies the zero-forcing requirement in our unlearning definition.

Under this constraint, the loss function becomes:

$$\mathcal{L}(\theta; x) = \log |\mathcal{C}_s| - H(P_\theta(\cdot|x)).$$

Since the safe subspace $\mathcal{C}_s$ is determined by the teacher construction and is independent of the student's parameters θ during the update step, $\log |\mathcal{C}_s|$ is a constant. Minimizing a function is equivalent to maximizing its negative. Therefore:

$$\min_\theta \{-H(P_\theta(\cdot|x))\} \iff \max_\theta H(P_\theta(\cdot|x)).$$

Combining the constraint and the objective, we arrive at the constrained entropy maximization problem. This theoretical result directly motivates the design of our SPID loss components: $\mathcal{L}_{supp}$ relaxes the hard constraint $P_\theta(\mathcal{C}_f|x) = 0$, and $\mathcal{L}_{em}$ implements the entropy maximization. □

E.3 Geometric Interpretation of the Suppression Loss

We formalize the statement in Section 4 that $\mathcal{L}_{supp}$ penalizes the projection of the hidden state onto the sensitive concept subspace. Fix a context x and let $h \in \mathbb{R}^d$ denote the final hidden state fed to the unembedding layer, so that the logits are $z_w = \mathbf{E}_w^\top h$ and $P(w) = [\text{softmax}(z)]_w$ for $w \in \mathcal{V}$. As in the proof of Corollary 4.3, the clusters $\mathcal{C}_f$ and $\mathcal{C}_s = \mathcal{V} \setminus \mathcal{C}_f$ are held fixed during an update step (they are piecewise constant in θ), and we assume both are nonempty. Let $p_f = \sum_{w \in \mathcal{C}_f} P(w)$, so that $\mathcal{L}_{supp} = p_f$; since softmax probabilities are strictly positive, $0 < p_f < 1$. We define the probability-weighted conditional centroids

$$\mu_f = \frac{1}{p_f} \sum_{w \in \mathcal{C}_f} P(w)\, \mathbf{E}_w, \qquad \mu_s = \frac{1}{1 - p_f} \sum_{w \in \mathcal{C}_s} P(w)\, \mathbf{E}_w, \qquad \bar{\mu} = p_f\, \mu_f + (1 - p_f)\, \mu_s, \quad (31)$$

and the centered span of the forget-cluster embeddings $\mathcal{S}_f = \text{span}\{\mathbf{E}_w - \bar{\mu} : w \in \mathcal{C}_f\}$.

Lemma E.1 (Geometry of the Suppression Gradient)**.** *Under the setting above, the following statements hold.*

(i) The gradient of the suppression loss with respect to the hidden state is

$$\nabla_h \mathcal{L}_{supp} = p_f(1 - p_f)\,(\mu_f - \mu_s) = \sum_{w \in \mathcal{C}_f} P(w)\,(\mathbf{E}_w - \bar{\mu}), \quad (32)$$

and hence $\nabla_h \mathcal{L}_{supp} \in \mathcal{S}_f$.

(ii) Suppose $\mu_f \neq \mu_s$ and let $u = (\mu_f - \mu_s)/\|\mu_f - \mu_s\|$. A gradient step $h^+ = h - \eta \nabla_h \mathcal{L}_{supp}$ with step size $\eta > 0$ leaves the component of h orthogonal to $\mathcal{S}_f$ unchanged and satisfies

$$\langle h^+, u\rangle - \langle h, u\rangle = -\eta\, p_f(1-p_f)\, \|\mu_f - \mu_s\| < 0. \tag{33}$$

(iii) Let $h = h_\theta(x)$ be twice continuously differentiable in θ with Jacobian $J = \partial h/\partial\theta$. A parameter update $\theta^+ = \theta - \eta \nabla_\theta \mathcal{L}_{supp}$ changes the component of h along u by

$$\langle h_{\theta^+}(x), u\rangle - \langle h_\theta(x), u\rangle = -\eta\, p_f(1-p_f)\, \|\mu_f - \mu_s\|\; u^\top J J^\top u + O(\eta^2), \tag{34}$$

whose first-order term is non-positive because $JJ^\top$ is positive semidefinite.

Proof. (i) Since $\mathcal{L}_{supp} = p_f = \sum_{w\in\mathcal{C}_f} P(w)$, the softmax Jacobian $\partial P(w)/\partial z_v = P(w)\,(\mathbb{1}[w=v] - P(v))$ gives

$$\frac{\partial p_f}{\partial z_v} = \sum_{w\in\mathcal{C}_f} P(w)\,(\mathbb{1}[w=v] - P(v)) = P(v)\,(\mathbb{1}[v\in\mathcal{C}_f] - p_f)\,. \tag{35}$$

Combining this with $\partial z_v/\partial h = \mathbf{E}_v$ via the chain rule,

$$\begin{aligned}\nabla_h \mathcal{L}_{supp} &= \sum_{v\in\mathcal{V}} P(v)\,(\mathbb{1}[v\in\mathcal{C}_f] - p_f)\,\mathbf{E}_v \\ &= \sum_{v\in\mathcal{C}_f} P(v)\,\mathbf{E}_v - p_f \sum_{v\in\mathcal{V}} P(v)\,\mathbf{E}_v = p_f\,\mu_f - p_f\,\bar{\mu},\end{aligned} \tag{36}$$

where we used $\sum_{v\in\mathcal{V}} P(v)\,\mathbf{E}_v = p_f\,\mu_f + (1-p_f)\,\mu_s = \bar{\mu}$. Substituting $\bar{\mu} = p_f\,\mu_f + (1-p_f)\,\mu_s$ yields $p_f\,\mu_f - p_f\,\bar{\mu} = p_f(1-p_f)(\mu_f - \mu_s)$. Alternatively, writing $p_f\,\mu_f = \sum_{w\in\mathcal{C}_f} P(w)\,\mathbf{E}_w$ and $p_f\,\bar{\mu} = \sum_{w\in\mathcal{C}_f} P(w)\,\bar{\mu}$ yields $\nabla_h\mathcal{L}_{supp} = \sum_{w\in\mathcal{C}_f} P(w)\,(\mathbf{E}_w - \bar{\mu})$, which is a linear combination of the vectors spanning $\mathcal{S}_f$.

(ii) By (i), $h^+ - h = -\eta\nabla_h\mathcal{L}_{supp} \in \mathcal{S}_f$, so the component of h orthogonal to $\mathcal{S}_f$ is unchanged. Moreover, $\nabla_h\mathcal{L}_{supp} = p_f(1-p_f)\,\|\mu_f - \mu_s\|\,u$, hence

$$\langle h^+ - h, u\rangle = -\eta\, p_f(1-p_f)\,\|\mu_f - \mu_s\|\,\langle u, u\rangle = -\eta\, p_f(1-p_f)\,\|\mu_f - \mu_s\|, \tag{37}$$

which is strictly negative because $0 < p_f < 1$ and $\mu_f \neq \mu_s$.

(iii) By the chain rule, $\nabla_\theta\mathcal{L}_{supp} = J^\top\nabla_h\mathcal{L}_{supp}$. A first-order Taylor expansion of $h_\theta(x)$ around θ gives

$$h_{\theta^+}(x) - h_\theta(x) = J(\theta^+ - \theta) + O(\eta^2) = -\eta\, JJ^\top\nabla_h\mathcal{L}_{supp} + O(\eta^2). \tag{38}$$

Taking the inner product with u and substituting $\nabla_h\mathcal{L}_{supp} = p_f(1-p_f)\,\|\mu_f - \mu_s\|\,u$ yields the stated expression. Finally, $u^\top JJ^\top u = \|J^\top u\|^2 \geq 0$. □

F Metrics Faithful Evaluation Details

F.1 Evaluation Experiment Settings

For each size of model, we adopt the hyperparameter grid of OPEN-UNLEARNING (Dorna et al., 2026): five learning rates (1×10^{-5} – 5×10^{-5}) and two checkpoint choices (epochs 5 and 10). Crossing these with two pools and two dataset variants produces 40 models overall. For SRK, we sample $k = 10$ times with temperature $\tau = 1.0$, and use `microsoft/deberta-large-mnli` NLI model (He et al., 2021).

We independently generate the training-set variants for each model pool using the following prompts in Appendix F.5 with paraphrased and perturbed examples. In practice, for robustness, we define the final entailment score as the maximum of the bidirectional entailment scores (i.e. $S_\Phi(y, y^*) \coloneqq \max[P_\Phi(\text{entailment} \mid y, y^*)\,,\; P_\Phi(\text{entailment} \mid y^*, y)]$).

F.2 Full Results Visualization

We further present the full results of these metrics in Figure 6.

F.3 Additional result on 1B model

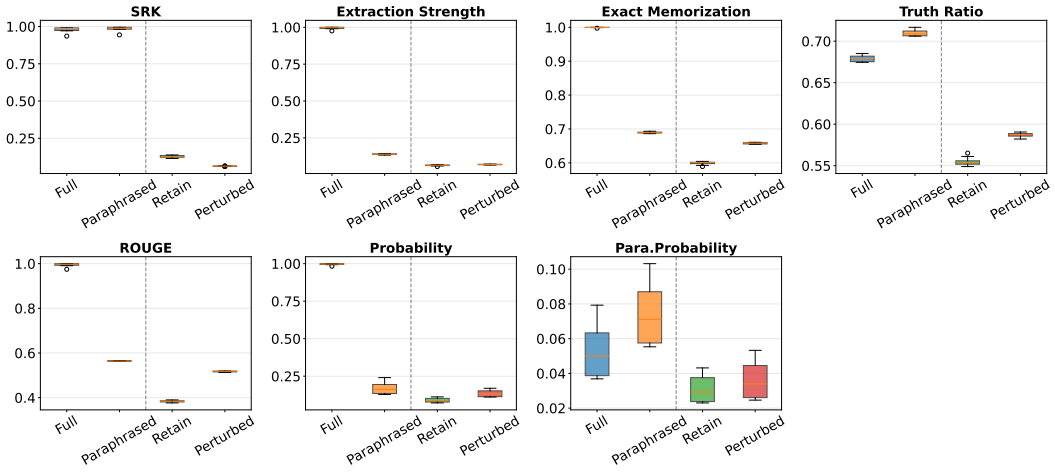


Figure 6: Box plots of all metrics across pools under different hyperparameter settings on 3B Model (dashed line separates Positive Pool $\mathcal{P}$ and Negative Pool $\mathcal{N}$).

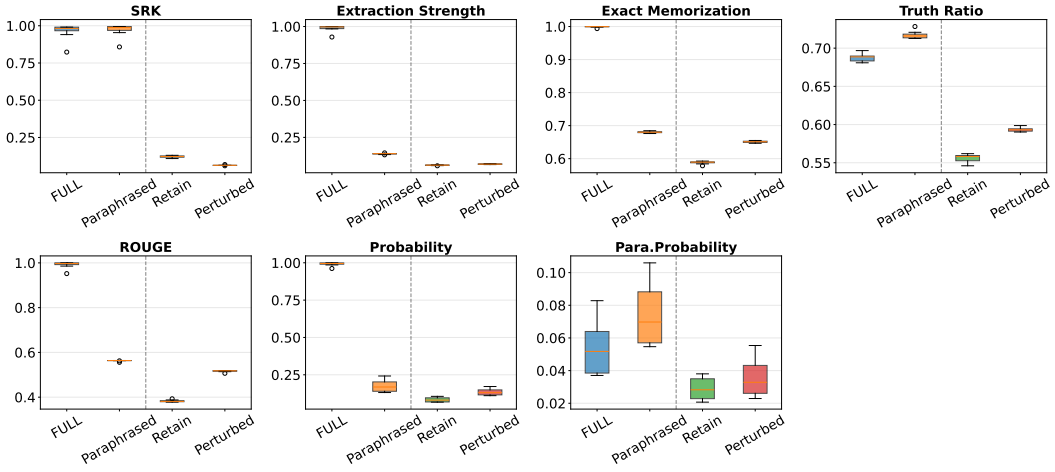


Figure 7: Box plots of all metrics across pools under different hyperparameter settings on 1B Model (dashed line separates Positive Pool $\mathcal{P}$ and Negative Pool $\mathcal{N}$).

To verify the scalability and consistency of our findings, we conduct the same statistical significance tests on the 1B model as reported for the 3B model in the main text, shown as Figure 7. Table 5 confirms that the results align perfectly with our previous observations.

Table 5: Statistical significance tests across multiple metrics on 1B Model: within Positive Pool $\mathcal{P}$ and between $\mathcal{P}$ and Negative Pool $\mathcal{N}$.

	Within $\mathcal{P}$		Between $\mathcal{P}$ & $\mathcal{N}$
Metric	Wilcoxon Test P_{w_p}	TOST Test P_{t_p}	Wilcoxon Test $P_{w_{pn}}$
SRK	**0.1055**	**0.0000**	6.68e-08
Extraction Strength	0.0020	1.0000	6.61e-08
Exact Memorization	0.0020	1.0000	6.61e-08
Truth Ratio	0.0020	1.0000	6.80e-08
ROUGE	0.0020	1.0000	6.24e-08
Probability	0.0020	1.0000	1.80e-06
Para.Probability	0.0020	0.5129	2.69e-06

SRK remains the only metric to satisfy semantic invariance within the positive pool $\mathcal{P}$. It shows no statistically significant difference under the Wilcoxon test ($p_{w_p} = 0.1055 > 0.05$) and successfully establishes statistical equivalence via the TOST test ($p_{t_p} \approx 0.0000$). In contrast, standard metrics like *Extraction Strength* and *ROUGE* fail to capture this invariance ($p_{t_p} = 1.0000$), behaving similarly to noise under syntactic perturbations. Furthermore, SRK maintains robust discriminative power between $\mathcal{P}$ and $\mathcal{N}$ ($p_{w_{pn}} \approx 6.68 \times 10^{-8}$), distinguishing conceptual retention from factual errors. These results validate that SRK reliably reflects the underlying semantic belief distribution regardless of model size.

F.4 Robustness of SRK to NLI Judges and Sampling Configurations

SRK relies on an NLI judge Φ and on a sampling configuration (temperature τ and number of samples K). To verify that our conclusions do not hinge on these choices, we re-run the full faithfulness protocol of Section 5 (Table 1) with four additional NLI judges of different scales and families, namely `deberta-base-mnli`, `deberta-xlarge-mnli`, `deberta-v2-xlarge-mnli` (He et al., 2021), and `roberta-large-mnli` (Liu et al., 2019); with sampling temperatures $\tau \in \{0.5, 1.0, 1.5\}$; and with numbers of samples $K \in \{5, 10, 20\}$. We vary one factor at a time and keep the others at their default values (`deberta-large-mnli`, $\tau = 1.0$, $K = 10$). For each variant, we report the Pearson and Spearman correlations between its SRK scores and those under the default configuration, together with the same statistical tests as in Table 1.

Table 6: Robustness of SRK to the NLI judge, sampling temperature τ, and number of samples K. Correlations are computed against the default configuration (shaded). Within $\mathcal{P}$, a high Wilcoxon p-value and a low TOST p-value indicate semantic invariance; between $\mathcal{P}$ and $\mathcal{N}$, a low Wilcoxon p-value indicates knowledge discrimination.

	Correlation w/ default		Within $\mathcal{P}$		Between $\mathcal{P}$ & $\mathcal{N}$
Setting	Pearson	Spearman	Wilcoxon Test P_{w_p}	TOST Test P_{t_p}	Wilcoxon Test $P_{w_{pn}}$
NLI judge Φ					
`deberta-large-mnli` (default)	1.0000	1.0000	0.182	0.000	6.6e-08
`deberta-base-mnli`	0.9999	0.9702	0.565	0.000	6.4e-08
`deberta-xlarge-mnli`	0.9990	0.9966	0.205	0.000	6.4e-08
`deberta-v2-xlarge-mnli`	0.9997	0.9972	0.205	0.000	6.4e-08
`roberta-large-mnli`	1.0000	0.9960	0.205	0.000	6.4e-08
Sampling temperature τ					
0.5	0.9999	0.9857	0.099	0.000	6.2e-08
1.0 (default)	1.0000	1.0000	0.182	0.000	6.6e-08
1.5	0.9975	0.9762	0.186	0.039	6.8e-08
Number of samples K					
5	0.9966	0.9783	0.156	0.000	6.4e-08
10 (default)	1.0000	1.0000	0.182	0.000	6.6e-08
20	1.0000	0.9945	0.156	0.000	6.7e-08

As shown in Table 6, SRK scores under every setting stay highly correlated with those under the default configuration (Pearson ≥ 0.996, Spearman ≥ 0.970). All variants preserve semantic invariance within $\mathcal{P}$: the Wilcoxon test finds no significant difference ($p_{w_p} \geq 0.099$), and the TOST test establishes equivalence at the 0.05 level ($p_{t_p} \leq 0.039$; the largest value occurs at $\tau = 1.5$). They also separate retained from erased knowledge ($p_{w_{pn}} \leq 6.8 \times 10^{-8}$). SRK is therefore robust to the choice of NLI judge and to the sampling configuration.

F.5 Prompts for Paraphrased and Perturbed Forget10 Dataset

Prompt of Paraphrased

You are a careful editor.
Task: Paraphrase ONLY the given ANSWER to the given QUESTION.

Hard requirements:
– Keep the QUESTION unchanged (do not rewrite it).
– Keep ALL facts in the answer identical (names, dates, numbers, places, relations, events, causality).
– Do NOT add new facts or remove any factual claims.
– Do NOT change the meaning, certainty, or level of detail.
– If you are unsure how to paraphrase without changing facts, copy the original answer verbatim.

Output requirements:
– Output ONLY the paraphrased answer text.
– No prefacing, no bullet labels, no quotes around the whole answer, no extra commentary.

QUESTION: {question}

ANSWER: {answer}

Paraphrased ANSWER:

Prompt of Perturbed

You are generating perturbed answers for training data.

Given a QUESTION and its ORIGINAL ANSWER, write a Perturbed answer that is likely incorrect.

Requirements:
– Do NOT change or rewrite the QUESTION.
– Do NOT copy the ORIGINAL ANSWER verbatim.
– The Perturbed answer must contain at least one concrete factual error (e.g., a wrong name/date/number/place/relationship), while remaining fluent.
– Keep the writing style, tone, and level of detail similar to the ORIGINAL ANSWER.
– Match the ORIGINAL ANSWER length as closely as possible (target within 10% characters).

Output:
– Output ONLY the Perturbed answer text.
– No bullet labels, no explanations, no metadata.

QUESTION: {question}

ANSWER: {answer}

Perturbed ANSWER:

F.6 EXAMPLES

Examples of Paraphrased cases

Question: What is the profession of Hsiao Yun–Hwa's father?

Answer: The father of Hsiao Yun–Hwa is a civil engineer.

Paraphrased Answer: Hsiao Yun–Hwa's father works as a civil engineer.

Examples of Perturbed cases

Question: What is the profession of Hsiao Yun–Hwa's father?

Answer: The father of Hsiao Yun–Hwa is a civil engineer.

Perturbed Answer: The father of Hsiao Yun–Hwa is a commercial airline pilot.

G UNLEARNING EXPERIMENT DETAILS

G.1 EXPERIMENT DETAILS

We mainly follow the default experimental settings of TOFU (Maini et al., 2024) and WMDP-Bio (Li et al., 2024). Unless otherwise noted, we adopt the same data splits and preprocessing protocols as the original benchmarks. For TOFU, we use a learning rate of 1×10^{-5} and train for 10 epochs. For WMDP, following Li et al. (2024) and Wang et al. (2025b), we report the best results selected from epochs 80 and 125, with learning rates of 1×10^{-5} and 4×10^{-6}.

Following default settings, for all methods except Gradient Ascent (GA), we employ the KL divergence as the retain loss ($\mathcal{L}_r$) to preserve general model utility, and we fix the regularization coefficient at $\lambda_2 = 1$ for the TOFU dataset. For the WMDP benchmark, we adopt $\lambda_2 = \{1, 10, 35\}$.

Hyperparameter Settings:

- **GA & GD:** No additional hyperparameter tuning.
- **WGA:** We tune the inverse temperature $\alpha \in \{0.05, 0.1, 0.5, 1, 5, 7\}$.
- **NPO:** We tune the reference weight $\beta \in \{0.05, 0.1, 0.5, 1\}$.
- **SimNPO:** We tune the reference weight $\beta \in \{2.0, 2.5, 3.0, 3.5, 4.5\}$ and sweep the stabilization offset $\delta \in \{0.125, 0.15, 0.20, 0.25\}$.
- **DPO:** We tune $\beta \in \{0.1\}$.
- **UnDIAL:** We tune the penalty intensity $\beta \in \{1, 100\}$.
- **DUET:** Following the original instructions, we sample using top-$K \in \{100, 200\}$ and set the prefix prompt to "As a responsible AI, I must refuse questions that could pose safety risks."
- **BS-S:** We tune the bootstrapping coefficient $\lambda_{\text{BSS}} \in \{0.2, 0.3, 0.6\}$ and the sampling count $N \in \{1, 3, 5\}$.
- **PMC:** Following the original PMC setting, we use vanilla PMC with the ROUGE-L dissimilarity reward $r(x) = 1 - \text{ROUGE-L}(x, y_0)$, where y_0 is the original greedy response. We set top-$p = 0.95$, tune $\lambda \in \{0.5, 0.75, 1, 2, 3\}$, the number of sampled responses $n \in \{10, 15\}$, and the sampling temperature in $\{0.8, 0.9, 1.0, 1.25, 1.5\}$. For response selection, we use the highest-reward sample, corresponding to the Bradley–Terry approximation used in the original implementation.
- **SPID:** We sweep the similarity threshold $\delta \in [0.4, 0.8]$, set top-$K = 100$, and tune $\lambda_1 \in \{0.1, 0.5, 1, 2\}$. For stop words, we utilize the stopwords set from the NLTK (Natural Language Toolkit) package.

Table 7: Performance of unlearning baselines on TOFU with LLaMA-3.2-3B/1B models. ↑/↓ indicates whether higher/lower values are better. For each unlearning setting, the best and runner-up results are highlighted in **bold** and underlined, respectively, and the two worst are shaded in red; SRK_f and SRK_r are computed on the forget and retain sets.

Method	Forget 1%				Forget 5%				Forget 10%			
	$SRK_f \downarrow$	Mem. ↑	$SRK_r \uparrow$	M.U. ↑	$SRK_f \downarrow$	Mem. ↑	$SRK_r \uparrow$	M.U. ↑	$SRK_f \downarrow$	Mem. ↑	$SRK_r \uparrow$	M.U. ↑
LLaMA-3.2-3B												
Target Model	0.8476	0.0122	0.7419	0.6661	0.8099	0.0305	0.7406	0.6661	0.7872	0.0325	0.7461	0.6661
Retain Model	0.2322	0.5421	0.7365	0.6623	0.1422	0.5688	0.7421	0.6594	0.1407	0.5739	0.7236	0.6498
GA	0.3348	0.3874	**0.6427**	0.6653	0.1717	0.7138	0.2812	0.0000	0.3385	0.8210	0.3326	0.0000
GD	0.3937	0.3472	0.6313	0.6499	0.4020	0.8108	0.4153	0.0000	0.3268	0.7916	0.3264	0.4452
WGA	0.2776	**0.8449**	0.2812	0.5921	0.2771	0.3501	0.3045	0.6192	0.2518	0.5230	0.6204	0.6410
NPO	0.3266	0.3806	0.6219	0.6660	0.3905	0.2343	0.3952	0.6474	0.0778	0.5577	0.3957	0.6367
SimNPO	0.3764	0.3163	0.6213	0.6524	0.2341	0.4043	0.4551	0.6281	0.1936	0.4145	0.4158	0.6207
DPO	0.2413	0.2945	0.4644	0.6539	**0.0478**	0.5153	0.0933	0.1805	**0.0412**	0.6162	0.5578	0.3813
UnDIAL	0.2453	0.5687	0.4664	**0.6831**	0.5374	0.8067	0.4383	0.6742	0.1273	0.8001	0.6322	0.6632
DUET	0.3249	0.2685	0.3711	0.6126	0.3112	0.2543	0.3107	0.5620	0.4886	0.1267	0.6281	0.6129
BS-S	0.2805	0.8308	0.3348	0.6287	0.1963	**0.8425**	0.2991	0.5680	0.1963	**0.8453**	0.3383	0.4703
PMC	**0.0650**	0.5672	0.3522	0.6651	0.0920	0.4574	0.4733	0.6625	0.1666	0.4148	0.4293	0.6612
SPID (Ours)	0.2518	0.5427	0.6127	0.6773	0.2641	0.8159	**0.5574**	**0.6824**	0.2251	0.7991	**0.6438**	**0.6719**
LLaMA-3.2-1B												
Target Model	0.5741	0.0676	0.4920	0.5992	0.5605	0.0848	0.5039	0.5992	0.5428	0.0868	0.4842	0.5992
Retain Model	0.1905	0.5447	0.5072	0.5986	0.1385	0.5768	0.4988	0.5991	0.1280	0.5848	0.5167	0.5911
GA	0.2685	0.4179	0.4136	0.5880	0.1630	0.7705	0.1350	0.0000	0.5501	0.8101	0.0000	0.0000
GD	0.3113	0.4355	0.2792	0.5162	0.1852	0.5108	0.1853	0.4552	0.1616	0.5516	0.1658	0.4427
WGA	0.2453	0.5069	0.3101	0.5896	0.1892	0.6219	0.1696	0.3656	0.1332	0.7420	0.2672	0.5434
NPO	0.2757	0.4227	0.4130	0.5885	0.5173	0.5268	**0.4841**	0.4493	0.1305	0.4941	0.2901	0.5467
SimNPO	0.3252	0.3472	**0.4259**	0.5925	0.2039	0.4140	0.3261	0.5657	0.1664	0.4265	0.3088	0.5700
DPO	**0.0703**	0.3900	0.2656	0.5580	0.1006	0.3873	0.3589	0.5645	**0.0279**	0.6395	0.3853	0.5713
UnDIAL	0.2392	0.5616	0.3332	**0.6044**	0.1436	0.8141	0.2914	0.5661	0.1685	0.8116	**0.4676**	0.5714
DUET	0.2806	0.2797	0.2771	0.5467	0.2516	0.3166	0.2999	0.5299	0.2377	0.3358	0.4386	0.5508
BS-S	0.2671	**0.8337**	0.2688	0.5266	0.2862	**0.8373**	0.2022	0.5366	0.2397	**0.8323**	0.2572	0.5260
PMC	0.2614	0.3345	0.3872	0.5631	0.1740	0.3688	0.3081	0.5684	0.1707	0.3815	0.3371	0.5603
SPID (Ours)	0.2258	0.4815	0.3705	0.6023	0.1288	0.7820	0.4117	**0.5836**	0.2303	0.7772	0.4403	**0.5828**

G.2 Hardware Resources

All experiments were conducted on a server equipped with two Intel Xeon Platinum 8358P CPUs (2.6 GHz), two NVIDIA A800 GPUs (80 GB each), and 1 TB of RAM. For the LLM API, we used GPT-5 via the Poe platform [1].

G.3 Additional Result of TOFU dataset

Table 7 reports full TOFU results under three unlearning budgets (Forget 1%/5%/10%) for both LLaMA-3.2-3B and LLaMA-3.2-1B, with the Target and Retain Models as references.

Overall, SPID achieves the best utility–retention trade-off across unlearning ratios. On LLaMA-3.2-3B, SPID attains the highest M.U. at Forget 5% (0.6824) and Forget 10% (0.6719), while simultaneously achieving the **best retain-set semantic score** SRK_r (0.5574 and 0.6438). Although DPO yields the lowest SRK_f (strongest forgetting), it significantly degrades retain utility at larger forget ratios, whereas SPID maintains competitive forgetting (SRK_f) without collapsing SRK_r and M.U. On LLaMA-3.2-1B, SPID again delivers the best M.U. at Forget 5% (0.5836) and Forget 10% (0.5828), with consistently strong SRK_r (0.4117/0.4403), indicating that the proposed SPID scales to smaller models.

G.4 Computational Overhead

We measure the training time of each method on TOFU Forget10 with LLaMA-3.2-3B on a single NVIDIA A800 GPU (10 epochs, effective batch size 16). SPID takes 274.0 s (Table 8), which is faster than half of the baselines, including the two leakage-oriented methods BS-S (600.8 s) and PMC (1455.9 s). SPID is slower than gradient-based methods such as GA (112.0 s) and GD (184.5 s), but the extra cost comes with a substantially better forgetting–utility trade-off (Tables 2 and 7).

[1] https://poe.com/

Table 8: Wall-clock time (in seconds) of a full unlearning run on TOFU Forget10 with LLaMA-3.2-3B on a single NVIDIA A800 GPU (10 epochs, effective batch size 16). Methods are sorted by running time.

Method	GA	GD	DUET	WGA	UnDIAL	SPID (Ours)	SimNPO	NPO	BS-S	DPO	PMC
Full run (s)	112.0	184.5	185.2	185.2	218.5	274.0	311.6	427.6	600.8	720.7	1455.9

G.5 Performance on Complex Multi-Token Knowledge

We further examine whether SPID remains effective when the target knowledge spans many tokens. To measure knowledge complexity, we use GPT-5 as an LLM-based judge on the TOFU Forget10 dataset with the LLaMA-3.1-8B model of Table 2. For each query and its ground-truth and perturbed answers, we prompt GPT-5 to extract the minimal factual span that directly answers the query. We then tokenize the span with the evaluated model's own tokenizer, so that span lengths are counted in that model's tokens. An example is shown below.

Example of factual span extraction

Query: What is the full name of the author born in Taipei on 05/11/1991?

Answer: The author's full name is Hsiao Yun–Hwa.

Extracted factual span: Hsiao Yun–Hwa

We treat factual spans with fewer than 12 tokens as short-span knowledge and spans with 12 or more tokens as complex multi-token knowledge. This splits the 400 QA pairs in TOFU Forget10 into 280 short-span and 120 complex multi-token instances. We then evaluate SPID and the leakage-oriented baseline BS-S separately on the two subsets using Mem. and SRK_f.

Table 9: Results on short-span (< 12 tokens) and complex multi-token (≥ 12 tokens) knowledge in TOFU Forget10 with LLaMA-3.1-8B. The last column reports the difference between the two subsets, with its confidence interval in brackets.

Method	Metric	Short (< 12)	Complex (≥ 12)	Complex $-$ Short
SPID (Ours)	Mem. $\uparrow$	0.791	0.794	$+0.003$ $[-0.008, +0.014]$
	$SRK_f \downarrow$	0.198	0.197	-0.002 $[-0.038, +0.036]$
BS-S	Mem. $\uparrow$	0.857	0.868	$+0.011$ $[-0.017, +0.036]$
	$SRK_f \downarrow$	0.192	0.226	$+0.034$ $[-0.001, +0.069]$

As shown in Table 9, SPID performs consistently across the two groups: for both Mem. and SRK_f, the complex-minus-short difference is close to zero and its confidence interval covers zero, so we see no measurable degradation on complex multi-token knowledge. BS-S, by contrast, shows a larger increase in SRK_f on complex spans ($+0.034$, with a confidence interval that only marginally includes zero), which suggests more residual semantic leakage as factual spans grow longer.

G.6 Representation-Level Probing Analysis

We probe the hidden representations to examine whether the output-level objective of SPID changes how the forget knowledge is encoded. We conduct the analysis on TOFU Forget10 with the 3B model and compare four models: the Base model before fine-tuning on TOFU, the Retain model fine-tuned only on the retain set, the Target model fine-tuned on the full TOFU data, and the Target model after unlearning with SPID. All measures use the hidden states from the last layer.

Linear decodability (AUC). For each QA pair, we label the correct answer as 1 and a structurally similar perturbed answer as 0, and train a logistic-regression probe on the last-layer hidden states of

each model. We report the AUC with 95% confidence intervals. A higher AUC indicates that the forget facts are more linearly decodable from the hidden states.

Concept-direction alignment (Cos.). For each model, we compute the true–perturbed direction, i.e., the direction that separates correct from perturbed answers in its hidden states, and report its cosine similarity with the concept direction given by the Target model, on the forget set (Cos. Forget) and the retain set (Cos. Retain). The Target model therefore scores 1 by construction.

Table 10: Representation-level probing on TOFU Forget10 with the 3B model. AUC is the linear decodability of forget facts, with 95% confidence intervals in brackets. Cos. is the cosine similarity between the true–perturbed direction of each model and the concept direction given by the Target model.

Metric	Base	Retain	Target	SPID (Ours)
AUC	0.858 [.833, .884]	0.855 [.829, .881]	0.945 [.924, .962]	0.828 [.801, .858]
Cos. Forget	0.430	0.758	1.000	0.601
Cos. Retain	0.433	0.915	1.000	0.918

As shown in Table 10, the Base and Retain models never learned the forget facts, so their AUC (about 0.86) serves as the reference level for a model without this knowledge. Fine-tuning raises the AUC to 0.945 in the Target model, and SPID reduces it to 0.828, with a confidence interval that overlaps those of the Base and Retain models. The alignment results show that this change is selective. On the forget set, SPID lowers the alignment with the Target model to 0.601, below that of the Retain model (0.758). On the retain set, SPID keeps the alignment at 0.918, on par with the Retain model (0.915). Together, these results indicate that the output-level objective of SPID weakens the forget knowledge encoded in the hidden states while preserving the representations of retained knowledge, consistent with the path-intervention view in Section 3.

G.7 Unlearned Model's Generated Response

Model Responses Comparison

Question: What is the full name of the geology author born in Karachi, Pakistan on 06/30/1975?

Target Model

The author's name is Hina Ameen.

SPID(ours)

The author, born on June 30, 1975, is named Suhail Ahmed.

WGA

Hina Ameera is the geology author born in Karachi, Pakistan on June 30, 1975.

NPO

The full name of the geology author, Ameera, is: Ameera, full name: Ameera, full name: Ameera, full name: Ameera, full name: Ameera, full name: Aameera, full name: Ameera, full name: Ameera, full name: Ameera, full name: Ameera, full name: Ameera, full name: Ameera, full name: Ameera, full name: Ameera, full name: Ameera, full name: Ameera, full name: A

DUET

The full name of the author born on June 30, 1975, in Karachi, Pakistan, is Hina Ameen.

UnDIAL

HINA Ameen is the full name of the Karachi-born geology author, HINA Ameen was born on 06/30/1975.

BS-S

.scalablytyped .scalablytyped

H Pseudocodes

In this section, we provide the detailed pseudocodes for our proposed unlearning framework and evaluation metric.

H.1 SPID PSEUDOCODE

Algorithm 1 details the training process of our proposed Semantic-aware Path Intervention Distillation (SPID) framework.

H.2 SRK PSEUDOCODE

Algorithm 2 describes the calculation of the Semantic Residual Knowledge (SRK) metric. It estimates the semantic leakage by integrating the entailment probabilities of generated samples weighted by their likelihoods.

Algorithm 1 Semantic-aware Path Intervention Distillation (SPID)

Input: Pre-trained parameters θ_0; unlearning set $\mathcal{D}_f$; retention set $\mathcal{D}_r$; learning rate η; entropy weight λ_1; retain weight λ_2; top-k parameter k; similarity threshold δ; stop words set $\mathcal{S}_{stop}$.
Output: Unlearned parameters θ.
Initialize $\theta \leftarrow \theta_0$
for $e = 1$ **to** T **do**
  for each mini-batch $(x_f, y_f) \sim \mathcal{D}_f$ and $(x_r, y_r) \sim \mathcal{D}_r$ **do**
    for each step t in sequence y_f **do**
      Compute logits z_t and probability distribution $P_\theta(\cdot|x_f, y_{<t})$
      // 1. Target Identification and Safety Gate
      Identify anchor token $\hat{w} = \text{argmax}_{v \in \mathcal{V}} P_\theta(v|x_f, y_{<t})$
      if $\hat{w} \in \mathcal{S}_{stop}$ **then**
        continue {Skip unlearning for stop words}
      end if
      // 2. Efficient Candidate Filtering (Top-k)
      Select top-k indices $\mathcal{V}_{top} \leftarrow \text{Top-}k(\{P_\theta(v|x_f, y_{<t})\}_{v \in \mathcal{V}})$
      Compute cosine similarity $s_v = \text{sim}(E_v, E_{\hat{w}}) \quad \forall v \in \mathcal{V}_{top}$
      // 3. Dynamic Semantic Clustering
      Identify Sensitive Cluster $\mathcal{C}_f = \{v \in \mathcal{V}_{top} \mid s_v > \delta\}$
      Identify Safe Subspace $\mathcal{C}_s = \mathcal{V}_{top} \setminus \mathcal{C}_f$
      // 4. Compute Unlearning Losses
      $\mathcal{L}_{supp} \leftarrow \sum_{v \in \mathcal{C}_f} P_\theta(v|x_f, y_{<t})$ {Suppression Loss}
      Initialize renormalized probabilities $\hat{P}$ for entropy calculation
      $Z_s \leftarrow \sum_{u \in \mathcal{C}_s} P_\theta(u|x_f, y_{<t})$
      for $v \in \mathcal{C}_s$ **do**
        $\hat{P}(v) \leftarrow P_\theta(v|x_f, y_{<t})/Z_s$
      end for
      $\mathcal{L}_{em} \leftarrow \sum_{v \in \mathcal{C}_s} \hat{P}(v) \log \hat{P}(v)$ {Entropy Maximization}
      $\mathcal{L}_{SPID} \leftarrow \mathcal{L}_{supp} + \lambda_1 \mathcal{L}_{em}$
    end for
    // Compute Retain Loss and Update
    $\mathcal{L}_{total} \leftarrow \mathcal{L}_{SPID} + \lambda_2 \mathcal{L}_{retain}$
    $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{total}$
  end for
end for

Algorithm 2 Semantic Residual Knowledge (SRK) Estimation

Input: Query $x \in \mathcal{D}_f$, Target Semantic String y^*, Target Model P_θ, NLI Model Φ
Parameters: Number of samples K, Temperature τ
Initialize samples list $\mathcal{Y} \leftarrow []$, weights list $\mathcal{W} \leftarrow []$
// 1. Temperature Sampling
for $k = 1$ to K **do**
    Generate response $y_k \sim P_\theta(\cdot|x)$ with temperature τ
    Calculate sequence log-likelihood $l_k = \log P_\theta(y_k|x)$
    Store sample y_k and likelihood l_k
end for
// 2. Importance Weighting
Compute normalization constant $Z = \sum_{j=1}^{K} \exp(l_j/\tau)$
for $k = 1$ to K **do**
    $w_k = \frac{\exp(l_k/\tau)}{Z}$
end for
// 3. Semantic Entailment Evaluation
Initialize $SRK \leftarrow 0$
for $k = 1$ to K **do**
    Compute entailment probability using NLI model:
    $S_k = P_\Phi(\text{entailment} \mid \text{hypothesis} = y_k, \text{premise} = y^*)$
    Accumulate weighted score:
    $SRK \leftarrow SRK + w_k \cdot S_k$
end for
Return SRK