FORGETTING-MARI: LLM UNLEARNING VIA MARGINAL INFORMATION REGULARIZATION

Shizhou Xu[†], **Stefan Bröcker**[*], **Thomas Strohmer**
Department of Mathematics
University of California Davis
Davis, CA 95616, USA
{shzxu,sabroecker,strohmer}@ucdavis.edu

Yuan Ni[*]
SLAC National Accelerator Laboratory
Stanford University
Menlo Park, CA 94025, USA
yn754@slac.stanford.edu

ABSTRACT

As Large Language Models (LLMs) face increasing regulatory scrutiny, the ability to surgically remove the influence of specific data without full retraining is critical, especially for *deployed agentic systems* that continuously accumulate user interactions, tool-use traces, and long-horizon trajectories. However, current LLM unlearning techniques are largely heuristic, lacking formal guarantees and often degrading model utility by removing information shared between the unlearn and retain sets. We bridge the gap between rigorous unlearning theory and LLM practice by introducing *Forgetting-MarI*. This framework provably isolates and removes only the marginal information, the unique effect contributed by the unlearn set, while preserving information supported by the retain set. By penalizing marginal information, we derive a tractable upper bound on the unlearn set's residual influence in the unlearned models, yielding a verifiable notion of undetectability. Extensive experiments on Llama and GPT models (up to 8B parameters) confirm that Forgetting-MarI achieves superior trade-offs between unlearning efficacy and utility preservation compared to state-of-the-art baselines. These results position marginal-information regularization as a principled and practical primitive for more controllable, auditable, and safe unlearning in real-world LLM deployments.

1 INTRODUCTION

As Large Language Models (LLMs) are increasingly deployed in high-stakes domains, the need to selectively remove specific data influences has become critical. This requirement is driven not only by privacy regulations such as the European Union's General Data Protection Regulation (GDPR) and its "right to be forgotten," but also by practical concerns including the removal of copyrighted content, personally identifiable information, and harmful or biased [30, 13, 8, 4, 3]. Unlearning, or the post-hoc removal of a specific data subset's *influence*, offers an efficient alternative to the prohibitive cost of retraining from scratch.

However, existing unlearning methods often *over*-unlearn, removing all information linked to the data to unlearn/forget, including knowledge also legitimately supported by the data meant to be preserved. This indiscriminate approach leads to "catastrophic collapse" in model utility on tasks unrelated to the specific unlearning request.

To illustrate this distinction, consider a copyright unlearning scenario where we have an LLM pretrained on an article from *The Washington Post* and on one from *The New York Times*, but only the former is legally authorized for use. Both outlets report on an identical event, yet their articles differ in narrative style, phrasing, and editorial perspective. There are two distinct unlearning objectives with this setup:

[*]Equal contribution.
[†]Corresponding author.

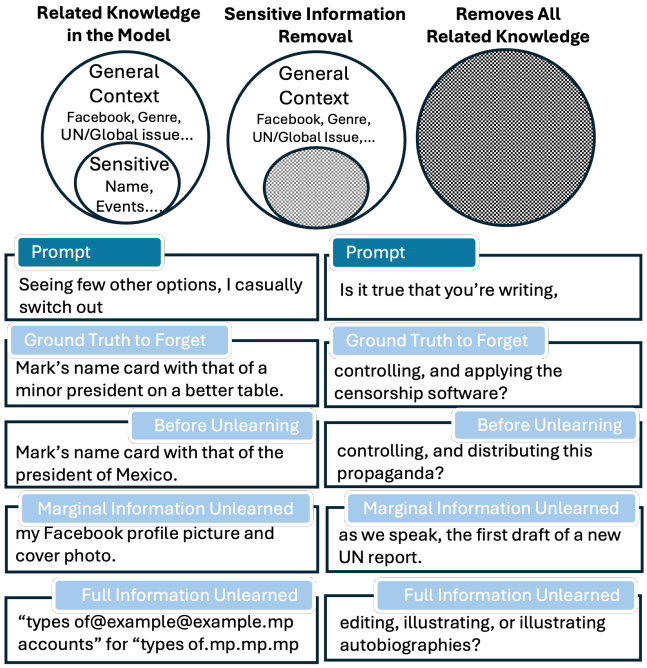


Figure 1: Comparison of sentence completions by Llama-3.2-1B. Before unlearning, the model memorizes the target text. *Full Information* unlearning destroys model utility to ensure forgetting. In contrast, our *Marginal Information* approach generates reasonable answers that differ from the target style, effectively unlearning the specific instance while preserving linguistic capability.

- **Marginal Information Unlearning:** Remove only the stylistic elements, phrasing, and content unique to the *Times* article, while retaining shared factual content that also appears in the authorized *Washington Post* article.
- **Full Information Unlearning:** Erase all content associated with the *Times* article, including factual information that is independently supported by the retained *Washington Post* article.

We argue that marginal information unlearning is the more natural and theoretically correct objective for LLMs unlearning. The goal is not to eradicate knowledge contained in the unlearn datasets, but to surgically remove only its *marginal effect*, the information not already supported by the data we are authorized to use. In the scenario above, the marginal effect unlearning satisfies legal requirements with minimal utility loss, whereas the full removal would unnecessarily discard information that is lawfully present in the model.

This distinction motivates our proposed method, *Forgetting-MarI*, which directly targets *marginal information*[*] of the unlearn dataset. More specifically, *marginal information unlearning* optimizes an objective that *directly* measures and suppresses only the *additional* information contributed by the unlearn set beyond what is already supported by the retain set, where the utility term only aims to stabilize retain performance or help the model learn new datasets if needed. There is no intrinsic conflict between the unlearn and utility objectives. In contrast, existing LLM unlearning methods are *full-information* in principle: their ascent or preference loss term targets the *entire* signal of $\mathcal{D}_u$ (e.g., maximizing CE on $\mathcal{D}_u$) and they attempt to *indirectly* spare shared/legitimate knowledge by counterbalancing this with a retain loss (CE or KL), preference shaping, parameter subtraction, or orthogonality (see details Appendix A.1). In other words, there is an intrinsic conflict between the utility and unlearning objective, which is necessary for the counterbalance to work, but often leads to unstable unlearning and requires extreme effort in parameter-tuning.

Figure 1 further demonstrates the difference between full-information unlearning and marginal-information unlearning (detailed experimental setup in Section 4). We created three models trained on ground truth prompts: one before unlearning, one after marginal information unlearning, and one after full information unlearning. Models are given the first half of a sentence (prompt) and are asked to complete it. Before unlearning, the model completes the sentences in a way that is similar to the ground truth. With marginal unlearning, the model produces different but coherent completions. With full unlearning, the model struggles to coherently complete the sentences.

[*]The term Marginal Information is formalized in Definition 1.1.

1.1 Open Challenges in LLM Unlearning

Effective LLM unlearning must balance three objectives [24]. First, *unlearn efficacy* measures how well a model suppresses the influence of the data we want to unlearn, called the unlearn set $\mathcal{D}_u$. Second, *utility preservation* ensures the model's ability to retain performance on general tasks and the data we are still authorized to use, called the retain set, $\mathcal{D}_r$ is not lost. Finally, *computational cost* encompasses the time, memory, and carbon used during unlearning. All unlearning techniques aim to optimize these three objectives, which inherently come with tradeoffs; what differs is *where* and *how* the model parameters are updated, directly affecting their ability to balance the three. A breakdown of existing techniques and their strengths and weaknesses is shown in Table 2, with their technical details and commonality in indirect marginal unlearning in Appendix A.1.

Despite rapid progress, LLM unlearning remains an emerging discipline with critical open challenges (summarized in Table 1): *Robustness & Utility Preservation*: Existing LLM unlearning

Method	Utility Preservation	Scalable Continual	Formal Guarantees
Full Parameter	☆	✓	✗
Weight Editing	☆	✓	✗
Counterfactual	✗	✓	✗
Adaptation	☆	☆	✗
Ours	✓	✓	✓

Table 1: Comparison of unlearning families based on literature evidence. Our proposed marginal information unlearning addresses key limitations of existing approaches. (✓=yes, ✗=no, ☆=partial)

techniques via full-parameter fine-tuning typically treat the unlearn set $\mathcal{D}_u$ as fully toxic, forcing the model to erase all information in $\mathcal{D}_u$ regardless of their overlap with the retain set $\mathcal{D}_r$. Examples include loss-reversal [23], gradient-difference [40], KL-ascent [15], and preference-based DPO/NPO [32, 41]. Even local editors that aim to make precise edits (ROME, MEMIT) share this limitation [28, 29], erasing shared facts and stylistic cues, and raising perplexity on $\mathcal{D}_r$ and held-out tasks. Benchmarks (RWKU, MUSE, Eight-Method) consistently report sizable utility drops after unlearning [20, 37, 26].

Stable Continual Unlearning: As the legal landscape around data usage changes, a deployed LLM may receive hundreds or thousands of unlearn requests. LLMs unlearning, therefore, needs to be able to repeatedly unlearn, retain utility, and keep computation and memory within a practical range. Exact methods like full retraining or shared SISA guarantee unlearning but their cost scales with both model size and request count [2, 12, 1]. Lighter updates like influence functions [14] or repeated ROME/MEMIT edits [28, 29] are cheap per removal yet accumulate inference costs and utility drift. Task-vector subtraction or adapter stacks save compute during unlearning but require storing external model adapters, also creating downstream inference costs [18, 10]. Thus, achieving stable continual unlearning without runaway resources or utility loss remains unsolved.

Formal Guarantees at LLM-Scale: Certified unlearning is well established for linear/kernel models [14], high-dimensional classifiers [43], and general mathematical formulations of machine unlearning [39]. However, no existing method provides guarantees that scale to autoregressive transformers with billions of parameters (1B+), such as GPT or Llama. As a result, practitioners lack *reliable guarantees* of the extent to which the unlearn set remains uninferable or undetectable after common downstream operations such as compression, distillation, or adversarial probing [24].

1.2 Our Contributions

To address these challenges, we introduce *Forgetting-MarI*, a novel information-theoretic LLM unlearning framework. To start, we provide a conceptual definition of marginal information (see a formal quantification in Section 2.1):

Definition 1.1 (Marginal Information (MarI))**.** *Marginal information is the marginal effect on model inference when adding the unlearn set to the retain set.*

The core idea of Forgetting-MarI is to penalize the model in proportion to the marginal information, and thus eliminate only the *unique* contribution of the unlearn dataset on the model's parameters and its inference abilities. This avoids erasure of shared information between the retain and unlearn sets.

A key piece of our technique, therefore, is an accurate quantification of marginal information, which we detail in Section 2.1. Forgetting-MarI can be summarized by the following learning objective:

$$\min_{\text{parameter: }\theta} \ell_{\text{utility}}(\text{model}(\theta), \mathcal{D}_r) + \ell_{\text{MarI}}(\text{model}(\theta), \mathcal{D}_r, \mathcal{D}_u),$$

with ℓ_{utility} being a loss that aims to maintain the utility of the model and ℓ_{MarI} being the marginal information loss derived from an accurate marginal information quantification.

The key contributions of our proposed method include:

- **(A1) Conceptual Framework:** We formalize *Marginal Information Unlearning*, a paradigm shift that targets the surgical removal of unique data contributions (e.g., specific user traces), resolving the intrinsic conflict between forgetting and utility found in existing methods.
- **(A2) Theoretical Guarantees:** We derive an explicit upper bound on the residual mutual information between the model and the unlearn set. This provides the first provable unlearning guarantee (in terms of undetectability) for LLMs unlearning. Such certification of unlearning efficacy is highly desirable to bridge the "mismatch between legal expectation and technical reality" [17].
- **(A3) Scalability & Continual Learning:** We demonstrate that our additive mutual-information regularizer integrates seamlessly with gradient-based fine-tuning, enabling stable continual unlearning, a critical capability for lifelong models or agents that must process deletion requests without catastrophic forgetting of their core reasoning abilities.
- **(A4) Empirical Validation:** Extensive experiments on Llama and GPT models (up to 8B parameters) validate our theory, showing that Forgetting-MarI achieves superior trade-offs between forgetting efficacy and utility preservation compared to state-of-the-art baselines.[†]

2 Unlearning: Marginal Information

Forgetting-MarI relies on a novel quantification of *marginal information* that (i) vanishes when the unlearn set $\mathcal{D}_u$ adds *no new* information beyond the retain set $\mathcal{D}_r$, and (ii) increases as $\mathcal{D}_u$ contributes information absent from $\mathcal{D}_r$. We propose a mutual information (MI)–based quantification that satisfies these properties.

2.1 Quantifying and Unlearning Marginal Effects

Fix a language model p_θ (with parameter θ) over a finite vocabulary V and a length $T \geq 1$. For $y \in V^T$, let $p_\theta(\cdot \mid y_{<t})$ be the next-token distribution. For a subset $s \subseteq V^T$, let μ_s be the uniform law on s and define its averaged next-token marginals $(p_\theta)^s_t(v) := \mathbb{E}_{Y\sim\mu_s}\big[p_\theta(v \mid Y_{<t})\big]$ for $t \in [T]$, $v \in V$. Write $p^{\,r} := \{(p_\theta)^r_t\}_{t\in[T]}$, $p^{\,u} := \{(p_\theta)^u_t\}_{t\in[T]}$. For $d := r \cup u$, $p^d_t = \alpha\, p^r_t + (1-\alpha)\, p^u_t$, $\alpha := \frac{|r|}{|r|+|u|} \in (0,1)$. Let $T^* \sim \text{Uniform}([T])$ and $Z \sim \text{Bernoulli}(\frac{1}{2})$ be independent. Conditioned on $(T^* = t, Z)$, draw $X \sim p^d_t$ if $Z = 0$ and $X \sim p^r_t$ if $Z = 1$, and set $X_{\text{MarI}} := (T^*, X)$. Then the *mutual information* between X_{MarI} and Z is defined as

$$I(X_{\text{MarI}}; Z) \;:=\; \frac{1}{T}\sum_{t=1}^{T} \text{JSD}\big(p^d_t, p^r_t\big). \tag{1}$$

Here, we denote the Jensen-Shannon divergence as $\text{JSD}(p,q) := \frac{1}{2}D_{\text{KL}}(p\|m) + \frac{1}{2}D_{\text{KL}}(q\|m)$ with $m := \frac{p+q}{2}$ and $D_{\text{KL}}(p\|q) := \sum_v p(v)\log\frac{p(v)}{q(v)}$. By construction, the information or distribution represented by d can be decomposed into the contribution of $r \cap d = r$ and the marginal contribution of $d \setminus r = u$. The distribution contributed by r through the model p_θ is $p^{\,r}$. The distributional contribution from the addition of u through p_θ is the distributional difference between $p^{\,r}$ and $p^{\,d}$.

By construction, the quantification of the marginal effect is small if $p^{\,r}$ is close to $p^{\,d}$, because such proximity suggests that the information content in u has already been largely represented by r. Conversely, the quantification will be large if $p^{\,r}$ differs significantly from $p^{\,d}$, indicating that u contributes substantial new information w.r.t. p_θ and induces a model output distribution shift. Therefore, defining this marginal effect quantification boils down to differentiating $p^{\,r}$ from $p^{\,d}$ for any $r \subset \mathcal{D}_r$ and $d = r \cup u$ with arbitrary $u \subset \mathcal{D}_u$.

[†]Our code and data will be made public upon acceptance.

A natural way to *quantify* this difference is via a binary detection problem. Consider a binary detection problem using the construction above:

$$X_t := X\big|_{T^*=t} \sim \begin{cases} p_t^d, & Z = 0, \\ p_t^r, & Z = 1, \end{cases} \tag{2}$$

$\mathbb{P}[Z = 0] = \mathbb{P}[Z = 1] = \frac{1}{2}$. If $p^r = p^d$, even an optimal classifier does no better than a coin flip. If there is distributional shift, it can detect the difference. A sharp information–theoretic upper bound on the Bayes accuracy, denoted by P_{acc} and defined below in Proposition 2.1, is the following:

Proposition 2.1 (Detection accuracy upper bounded by mutual information)**.** *For (X_{MarI}, Z) with prior $\pi = \mathbb{P}[Z = 1]$,*

$$P_{\text{acc}} \;=\; \mathbb{E}\left[\max\{P(Z=0\,|\,X_{\text{MarI}}),\, P(Z=1\,|\,X_{\text{MarI}})\}\right] \;\le\; 1 - H_2^{-1}\big(H_2(\pi) - I(X_{\text{MarI}}; Z)\big),$$

where $H_2(\cdot)$ is the binary entropy and H_2^{-1} denotes the inverse of H_2 restricted to $[0, \frac{1}{2}]$.

Proof in Appendix B.1. Here $P(Z \mid X_{\text{MarI}})$ denotes the Bayes-optimal posterior between retain r and union d. Note that $I(X_{\text{MarI}}; Z) \in [0, H_2(\pi)]$ satisfies: (i) $I(X_{\text{MarI}}; Z) = 0$ when $p^d = p^r$; (ii) $I(X_{\text{MarI}}; Z)$ grows with their divergence, approaching $H_2(\pi)$ as $\mathcal{D}_r$ and $\mathcal{D}_u$ has mutually exclusive support. Since $p^d \neq p^r$ occurs precisely when u induces model confidence shifts not explained by r, Proposition 2.1 gives $I(X_{\text{MarI}}; Z)$ an intuitive meaning as the detectability of the marginal effect (Definition 1.1).

Definition 2.1 (MI-based marginal information loss)**.** *With (X_{MarI}, Z) as in equation 2, define*

$$\ell_{\text{MarI}}(\theta, r, u) \;:=\; I(X_{\text{MarI}}; Z).$$

Thus, Forgetting-MarI solves

$$\min_{\theta} \;\; \ell_{KL}(\theta, r) \;+\; \ell_{\text{MarI}}(\theta, r, u), \tag{3}$$

where $\ell_{KL}(\theta, r) := D_{\text{KL}}\big(p^r(\theta) \,\|\, p^r(\theta_0)\big)$ is the KL divergence between the updated model (parameter θ) and the frozen original model (parameter θ_0) on r, and ℓ_{MarI} is as above, motivated by Prop. 2.1. Algorithm 1 describes an efficient LLM implementation.

Remark 2.1 (Alternative quantification)**.** *Marginal information measures the shift from $p_\theta(r)$ to $p_\theta(d)$. Alternatively, one may use $\ell'_{\text{MarI}}(\theta, r, u) := D_{\text{KL}}\big(p^d \,\|\, p^r\big)$ or $D_{\text{KL}}\big(p^d \,\|\, p^r(\theta_0)\big)$. But mutual information has the advantage of (1) stability (boundedness), (2) interpretability (Proposition 2.1), and (3) continuous unlearning (evolving reference $m = \frac{p+q}{2}$). See Appendix B.2 for details.*

2.2 Marginal Information & Perplexity-Based Detectors

We provide theoretical guarantees for the unlearning performance of Forgetting-MarI against white-box copyright detectors that rely on model confidence (perplexity / cross-entropy). Let

$$S_\theta(x, y) \;=\; \frac{1}{T} \sum_{t=1}^{T} \big(-\log p_\theta(x_t \mid y_{<t})\big)$$

be the standard cross-entropy (per-token negative log-likelihood). State-of-the-art detectors [3, 42, 31] flag the membership of x in training by testing whether $S_\theta(x, x)$ is suspiciously low. We adopt the notation from Section 2.1: sequences $r, u \in V^T$, next-token marginals p_t^r, p_t^u, their mixture $p_t^d = \alpha\, p_t^r + (1-\alpha)\, p_t^u$ with $\alpha = \frac{|r|}{|r|+|u|}$, and the mutual information $I(X_{\text{MarI}}; Z)$.

The next result shows that, given a set of sequences to forget, denoted by u, Forgetting-MarI guarantees that there is a set of sequences in the retain set, denoted by r, such that the score $S_\theta(u, u)$ is close to $S_\theta(u, r)$. In other words, a high model confidence implied by $S_\theta(u, u)$ is possibly due to the existence of r because one would get the same perplexity for u if feeding the model r instead of u.

Theorem 2.1 (MarI controls the self-perplexity gap)**.** *Fix $u = (u_1, \ldots, u_T) \in V^T$ and assume the* pathwise *non-vanishing condition $\min\{p_t^u(u_t),\, p_t^r(u_t)\} \ge \gamma \in (0, 1]$ for all t. Then*

$$\Big|\, S_\theta(u, u) - S_\theta(u, r) \,\Big| \;\le\; \frac{2\sqrt{2}}{\gamma(1-\alpha)} \,\sqrt{I(X_{\text{MarI}}; Z)}\,.$$

Algorithm 1: Forgetting-MarI

Require: Retain dataset $\mathcal{D}_r = \{x_i^r\}_{i=1}^{|\mathcal{D}_r|}$, unlearn dataset $\mathcal{D}_u = \{x_j^u\}_{j=1}^{|\mathcal{D}_u|}$; initial model f_{θ_0}; learning rate η; batch sizes B_r, B_u; epochs E; trade-off $\gamma \in (0,1)$; hetero $\in \{\texttt{False}, \texttt{True}\}$.

Ensure: Unlearned parameters θ_E.

Initialize $\theta \leftarrow \theta_0$; build dataloaders $\mathcal{L}_r, \mathcal{L}_u$.
for $e = 1$ **to** E **do**
  Shuffle $\mathcal{L}_r, \mathcal{L}_u$; $S \leftarrow \min\{|\mathcal{L}_r|, |\mathcal{L}_u|\}$.
  for $s = 1$ **to** S **do**
    Minibatches: $r \leftarrow \mathcal{L}_r(s), \quad u \leftarrow \mathcal{L}_u(s)$.
    Model logits & probabilities:
      $L^r \leftarrow f_\theta(r) \in \mathbb{R}^{B_r \times T \times |V|}$, $p^r_{b,t} \leftarrow \mathrm{softmax}(L^r[b,t,:]) \in [0,1]^{B_r \times T \times |V|}$;
      $L^r_0 \leftarrow f_{\theta_0}(r) \in \mathbb{R}^{B_r \times T \times |V|}$, $p^r_{b,t}(\theta_0) \leftarrow \mathrm{softmax}(L^r_0[b,t,:]) \in [0,1]^{B_r \times T \times |V|}$;
      $L^u \leftarrow f_\theta(u) \in \mathbb{R}^{B_u \times T \times |V|}$, $p^u_{b,t} \leftarrow \mathrm{softmax}(L^u[b,t,:]) \in [0,1]^{B_u \times T \times |V|}$;
      $p^d \leftarrow \mathrm{concat}_{\mathrm{batch}}([p^r, p^u]) \in [0,1]^{(B_r+B_u) \times T \times |V|}$.
    Utility Loss:
      $\bar{p}^r \leftarrow \frac{1}{B_r * T} \sum_{b,t} p^r_{b,t}$, $\bar{p}^r_0 \leftarrow \frac{1}{B_r * T} \sum_{b,t} p^r_{b,t}(\theta_0)$, $\ell_{\mathrm{KL}} \leftarrow \sum_{v \in V} \bar{p}^r(v) \log(\frac{\bar{p}^r(v)}{\bar{p}^r_0(v)})$.
    MarI Loss:
      $P(v,0) \leftarrow \frac{1}{2(B_r+B_u)*T} \sum_{b,t} p^d_{b,t}$, $P(v,1) \leftarrow \frac{1}{2B_r*T} \sum_{b,t} p^r_{b,t}$, $P(v) = \sum_{z \in \{0,1\}} P(v,z)$,
      $\ell_{\mathrm{MarI}} \leftarrow \sum_{z \in \{0,1\}} \sum_{v \in V} P(v,z) \log \frac{P(v,z)}{P(v)\,P(z)}$.
    Total loss: $\mathcal{L} = (1-\gamma)\,\ell_{\mathrm{KL}} + \gamma\,\ell^{\mathrm{flat}}_{\mathrm{MarI}}$.
    Update: $\theta \leftarrow \theta - \eta\,\nabla_\theta \mathcal{L}$.
  end for
end for
return $\theta_E \leftarrow \theta$

Figure 2: Pseudo-code for Forgetting-MarI.

See Appendix B.3 for the proof. In particular, when $I(X_{\mathrm{MarI}}; Z)$ goes to zero, the perplexity gap above vanishes, and the perplexity/log-likelihood detectors lose discriminative power after unlearning.

In addition, one can also show that MarI directly controls the perplexity gap even for a neighborhood of u rather than only u itself. The formal result and its proof can be found in Appendix B.4.

3 Algorithm Design

Sentence-wise MarI (Equation 1) provides strong guarantees when sentences in r and u are *homogeneous* in length and token-wise context. In practice, however, sentence-wise MarI Loss can be noisy under heterogeneous batches. To address this, we also provide a *pooled* ("flattened") estimator that first averages across token positions (and batch) to form $\bar{p}^{\,s} = \frac{1}{T} \sum_t p^s_t$, $s \in \{r, u, d\}$, then computes the pooled MarI Loss $I(\bar{X}_{\mathrm{MarI}}; Z) = \mathrm{JSD}\left(\bar{p}^{\,d}, \bar{p}^{\,r}\right)$. Such a pooled version aims to stabilize the marginal information quantification by filtering the position-heterogeneous noise and emphasizing the dominant distribution shift.

By the data-processing inequality, the pooled estimator $I(\bar{X}_{\mathrm{MarI}}; Z)$ is a *variational lower bound* to the sentence-wise MarI. The gap between the pooled and sentence-wise MarI losses is controlled by the ℓ^2-deviation of the token-sequence densities (see details in Theorem C.1[Pooling error bound], Appendix C). Furthermore, pooled MarI offers a specific word-wise guarantee:

Theorem 3.1 (Word-level provable unlearning via pooled MarI)**.** *Fix a token $w \in V$ and assume $\bar{p}^{\,u}(w) \wedge \bar{p}^{\,r}(w) =: \bar{\gamma}_w \in (0,1]$. Then*

$$\left| \log \bar{p}^{\,u}(w) - \log \bar{p}^{\,r}(w) \right| \;\le\; \frac{2\sqrt{2}}{\bar{\gamma}_w\,(1-\alpha)} \sqrt{I(\bar{X}_{\mathrm{MarI}}; Z)}\,.$$

Proof in Appendix C. Although weaker than the sentence-wise guarantee (Theorem 2.1), this guarantee is sufficient for provable unlearning of specific tokens and can be useful for the provable removal of word-level (not related to sentence structure) information.

Figure 2 presents pseudo-code for Forgetting-MarI with pooled MarI. A full, detailed algorithm is provided as Figure 1 in Appendix C. We recommend the following protocol:

- **Estimator Selection:** Use *sentence-wise* MarI for homogeneous data (aligned contexts) to leverage precise signals. Use *pooled* MarI for large corpora with random batching to ensure stability.

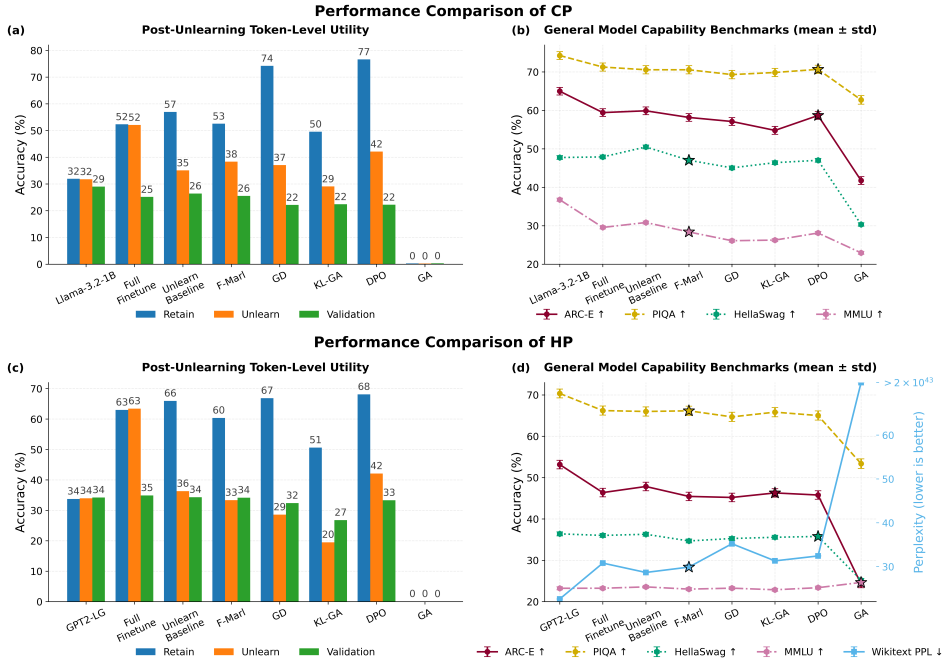


Figure 3: Left panels summarize the next-token accuracies on retain/unlearn/validation whereas the right panels summarize the general-capability on various benchmarks. Top row shows the results from Llama-3.2-1B on *Careless People* (correlated split), where as bottom row shows the results from GPT-2 Large on *Harry Potter*. Each method is reported at its best λ and training epoch. On the left panels, an ideal method should match the unlearn baseline on retain/unlearn/validation accuracy on the left panels. On the right, better methods should achieve higher accuracy on ARC-E, HellaSwag, PIQA, MMLU and lower on WikiText perplexity test. Star indicates the best performer on that test.

- **Hyperparameter Tuning:** Fix a tolerable perplexity gap for the application, derive the required MarI threshold, and tune γ until the regularization magnitude falls below this threshold. This invokes the guarantees of Theorem 2.1 or 3.1.

As both estimators achieve comparable empirical results, Section 4 reports findings under a unified label (comparative ablation in Appendix C).

4 Experiments

Our experiments are designed to empirically validate the theoretical properties of Forgetting-MarI formalized in Section 2. Specifically, we address three key questions:

1 **The Utility-Unlearning Trade-off:** Does minimizing marginal information allow us to remove $\mathcal{D}_u$ without the "over-unlearning" observed in baselines? We test if the method achieves parity with the *unlearn baseline* (the gold-standard retrain-on-retain model) in both forgetting efficacy and utility preservation.

2 **Continual Unlearning Stability:** Is the method robust to sequential unlearning requests (continual unlearning) without incurring the catastrophic drift or interference common in full-information methods?

3 **Provable Undetectability:** Consistent with Theorem 2.1, does the method defeat white-box perplexity detectors that rely on confidence scores to flag training data?

We compare against state-of-the-art full-parameter unlearning baselines, all of which target the "full information" of the unlearn set rather than the marginal contribution. These include: **GA** [40] (Gradient Ascent on $\mathcal{D}_u$), **GD** [23] (Gradient Ascent on $\mathcal{D}_u$ + Descent on $\mathcal{D}_r$), **KL-GA** [15] (Gradient Ascent on $\mathcal{D}_u$ + $D_{\mathrm{KL}}(p_\theta(D_r)||p_{\theta_0}(D_r))$), and **DPO** [32] (Preference Optimization treating $\mathcal{D}_u$ as negative). A detailed comparison of objectives and unlearning nature is provided in Appendix D.1 (Table 3).

4.1 Experimental Setup

Protocol. Following Maini et al. [27], Eldan & Russinovich [6], we adopt a rigorous three-step evaluation protocol to establish a theoretical ground truth: (i) **Full Finetune Baseline:** We first fine-tune a base model on the union $\mathcal{D}_u \cup \mathcal{D}_r$. This serves as the starting point for all unlearning

Method	Unlearning Step	Retain (%) ↑	Validate (%) ↑	Hermione (%) ↓	Snape (%) ↓	Ron (%) ↓	ARC-E (%) ↑	WikiText (PPL) ↓
KL-GA	Step 1: Forget Hermione	51.0	33.5	4.7	58.0	59.6	46.1	33.7
KL-GA	Step 2: Forget Snape	52.2	34.2	9.4	9.2	61.8	46.2	50.0
KL-GA	Step 3: Forget Ron	49.3	28.9	9.6	14.7	7.1	44.6	3.4e+10
DPO	Step 1: Forget Hermione	67.3	33.3	44.6	69.0	69.5	46.4	33.1
DPO	Step 2: Forget Snape	69.9	33.5	45.0	46.4	71.9	45.5	35.6
DPO	Step 3: Forget Ron	79.9	33.4	42.0	46.0	50.4	46.2	40.3
GD	Step 1: Forget Hermione	66.3	33.3	11.7	67.2	67.7	44.1	46.3
GD	Step 2: Forget Snape	68.8	33.2	12.0	15.7	70.1	43.1	62.4
GD	Step 3: Forget Ron	78.2	31.7	10.7	15.8	16.0	42.8	61550
F-MarI	Step 1: Forget Hermione	60.0	36.3	16.2	60.9	62.9	46.1	29.5
F-MarI	Step 2: Forget Snape	60.3	36.2	15.5	14.0	63.1	45.2	29.9
F-MarI	Step 3: Forget Ron	60.4	34.9	13.4	15.0	15.2	44.0	38.5

(a) GPT2-LG on *Harry Potter* (HP).

Method	Unlearning Step	Retain (%) ↑	Validate (%) ↑	First 15 (%) ↓	Next 15 (%) ↓	Next 15 (%) ↓	ARC-E (%) ↑	PIQA (%) ↑
KL-GA	Step 1: Forget First 15%	43.1	16.3	7.5	32.2	34.0	53.7	68.9
KL-GA	Step 2: Forget Second 15%	31.1	5.3	3.0	1.8	12.0	53.2	67.3
KL-GA	Step 3: Forget Third 15%	21.9	2.4	1.5	1.1	0.9	52.1	66.4
DPO	Step 1: Forget First 15%	75.0	22.9	42.5	46.1	46.5	57.2	70.7
DPO	Step 2: Forget Second 15%	82.6	20.8	40.4	38.3	42.8	54.9	70.3
DPO	Step 3: Forget Third 15%	85.9	20.1	39.1	37.4	36.2	52.0	68.6
GD	Step 1: Forget First 15%	76.5	19.3	21.3	41.4	41.4	53.8	69.5
GD	Step 2: Forget Second 15%	74.7	12.7	15.4	7.0	30.2	48.9	68.3
GD	Step 3: Forget Third 15%	76.5	8.6	11.5	5.8	3.8	41.5	64.2
F-MarI	Step 1: Forget First 15%	51.9	25.3	35.0	49.7	50.0	58.6	70.7
F-MarI	Step 2: Forget Second 15%	51.6	25.1	34.7	30.4	47.7	57.9	70.6
F-MarI	Step 3: Forget Third 15%	51.2	25.0	34.3	30.2	27.8	55.9	69.3

(b) Llama-3.2-1B on *Careless People* (CP).

Figure 4: Heatmap of Continual Unlearning Stability. Darker indicates higher accuracy (or lower perplexity). Ideal behavior: Stable high performance on Retain/General tasks across steps (columns), with persistent low performance on previously unlearned sets. Forgetting-MarI shows the most consistent stability, avoiding the "relearning" or "utility collapse" seen in baselines.

operations. (ii) **Unlearning Step:** We apply each unlearning method to remove the influence of $\mathcal{D}_u$, producing the unlearned models. (iii) **Unlearn Baseline (Gold Standard):** We train a separate model from scratch *only* on $\mathcal{D}_r$, never exposing it to $\mathcal{D}_u$. The success metric is strict: an ideal unlearning method should yield a final model with similar performance as the *Unlearn Baseline*, preserving retain & validation accuracy at the same level while reducing unlearn accuracy to the baseline's level. We report next-token accuracy on $\mathcal{D}_r$, $\mathcal{D}_u$, and held-out validation sets, alongside general capability metrics using Eleuther's LM Evaluation Harness [11].

Datasets & Models. We evaluate Forgetting–MarI on two standard architectures: GPT-2 Large (774M) and Llama-3.2-1B. To verify robustness on larger architectures, we also scale our evaluation to Llama-3-8B, observing similar performance trends (detailed results in Appendix D.5). We select two text domains with distinct genres and pretraining prevalence: (i) *Harry Potter and the Prisoner of Azkaban* [21], a popular choice of unlearn set but likely to present in models' pretraining; and (ii) *Careless People: A Cautionary Tale of Power, Greed, and Lost Idealism* [34], published after Llama's release and thus unlikely to appear in its pretraining.

Splits. For *Harry Potter* (HP), we designate 10% of sentences as $\mathcal{D}_u$ and the remaining 90% as $\mathcal{D}_r$ (cf. 6); validation uses excerpts from another book in the series (*Harry Potter and the Sorcerer's Stone* [33]). For *Careless People* (CP), we consider: (i) a *correlated* split with contiguous 50/50 spans for $\mathcal{D}_u/\mathcal{D}_r$; and (ii) an *uncorrelated* split where $\mathcal{D}_u$ comprises 2025 Reddit stories (post-release) and $\mathcal{D}_r$ is 50% of the book. Reddit stories are used as validation in the first setting and the remaining 50% of the book is used as validation in the second setting.

Parameter Tuning & Runtime Information. For each method, we stop training once validation accuracy drops by more than 3% from its initial value (to prevent general-utility degradation). Hyperparameter sweeps, ablations, and full training trajectories are provided in Appendix D.6. Hardware specifications and runtime information, such as comparison on peak GPU memory and time per batch, can be found in Appendix D.2.

4.2 Accuracy Trade-off & General Capability

Below, unlearn baseline is the model fine-tuned on retain set only, which serves as a retrain from scratch model.

Token-Level Utility (Left Panels, Fig. 3). Across both HP and CP datasets, Forgetting–MarI demonstrates a unique ability to match the performance of the Unlearn Baseline.

- *Careless People (Llama-3.2):* We observe the instability inherent in full-information methods. GD and DPO tend to "panic-fit" the retain set, significantly overtraining on $\mathcal{D}_r$ to compensate for

the unlearning loss. Meanwhile, KL-GA struggles to balance the objectives, failing to remove $\mathcal{D}_u$ effectively without degrading validation accuracy. Forgetting-MarI, by contrast, removes the target influence without perturbing the authorized data distribution, effectively matching the gold standard.
- *Harry Potter (GPT-2):* While GD and DPO show less overfitting here, they still miss the precise unlearning target, either over-unlearning (damaging shared knowledge) or under-unlearning. KL-GA again fails to remove $\mathcal{D}_u$ while maintaining $\mathcal{D}_r$ parity. Forgetting-MarI remains the only method that aligns with the Unlearn Baseline across all three metrics (Retain, Unlearn, Validation).

These trends hold consistent when scaling to Llama-3-8B. See Appendix D.5 for more details.

General Capability (Right Panels, Fig. 3). We compare the models on standard benchmarks: ARC-E, PIQA, HellaSwag, WikiText, and MMLU.

- *Llama on Careless People:* Forgetting–MarI attains the highest results on HellaSwag and MMLU and ranks second on PIQA and ARC-E (behind DPO, which failed to unlearn effectively). Notably, it is the only method that matches the capabilities of the original Full Finetune baseline. This implies zero degradation in model capacity, a critical result given that the Full Finetune baseline actually lags behind the Unlearn Baseline on some metrics, suggesting that Forgetting-MarI could match the Unlearn Baseline even more closely with improved fine-tuning initialization.
- *GPT-2 on Harry Potter:* Forgetting–MarI ranks first on PIQA and WikiText, second on ARC-E, and remains competitive on HellaSwag. Crucially, we note that high benchmark scores can be deceptive without context: for example, GA outperforms all methods on MMLU but suffers from catastrophic WikiText perplexity (consistent with its theoretical tendency to destroy language modeling distribution). This confirms that evaluating perplexity/confidence is essential to avoid confounding factors.

In summary, Forgetting–MarI delivers targeted forgetting while largely preserving general capability, consistently outperforming full-information baselines on the critical retain/unlearn/validation frontier. (See Appendix D.7 for complete tabular results).

4.3 Continual Unlearning with Sequential Requests

Setup. To simulate real-world compliance queues, we adopt a multi-stage protocol: the total forget set is partitioned into three disjoint subsets $\mathcal{D}_u = \mathcal{D}_{u,1} \cup \mathcal{D}_{u,2} \cup \mathcal{D}_{u,3}$. Starting from the full model ($\mathcal{D}_r \cup \mathcal{D}_u$), we perform three sequential unlearning steps.

- **Harry Potter (Character-based):** To emulate "Right to be Forgotten" user requests, we create a "forget-characters" benchmark. Sentences are assigned to characters via alias matching (e.g., "Granger" $\rightarrow$ Hermione). We sequentially target: $\mathcal{D}_{u,1}$ (Hermione) $\rightarrow$ $\mathcal{D}_{u,2}$ (Snape) $\rightarrow$ $\mathcal{D}_{u,3}$ (Ron). The retain set $\mathcal{D}_r$ includes all "other" sentences, all retain-character sentences (e.g., Harry, Dumbledore), and sentences of not-yet-unlearned characters.
- **Careless People (Random-split):** We randomly partition the unlearn set into three equal, disjoint subsets.

Results (Figure 4).

- *GPT-2 on HP.* Forgetting–MarI is the only method that remains robust across retain and validation accuracy, preserves forgetting on previously unlearned sets, and sustains general capability. By contrast, KL-GA tends to *relearn* previously forgotten content and substantially degrades WikiText performance; DPO fails to effectively unlearn; GD overshoots on $\mathcal{D}_r$ and also loses general capability (WikiText).
- *Llama on CP.* Forgetting–MarI again exhibits the most consistent behavior: retain/validation accuracy and previously unlearned sets remain stable, and ARC-E/PIQA show minimal drift. In comparison, KL-GA quickly over-unlearns, lowering retain/validation and prior-unlearn accuracies; DPO again fails to unlearn effectively; GD overshoots on $\mathcal{D}_r$, reduces validation accuracy, and shows deteriorating ARC-E/PIQA performance.

Conclusion. Forgetting–MarI delivers robust sequential unlearning: it preserves utility, maintains prior unlearning, and sustains general capability across steps, outperforming full-information baselines in both performance and stability.

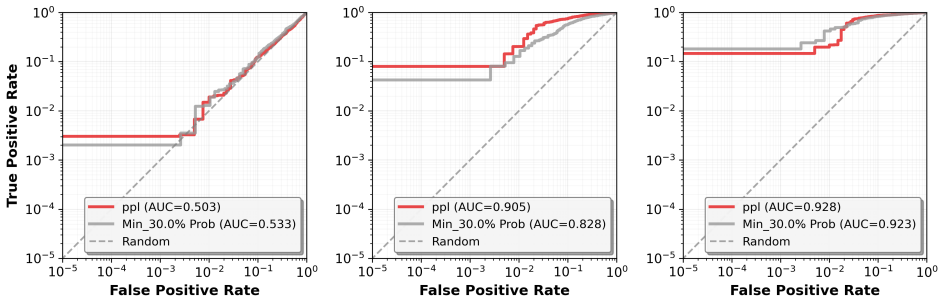


Figure 5: ROC Curves for SOTA White-Box Membership Inference Attack. Left: Full Model (High detectability). Middle: Forgetting-MarI. Right: Unlearn Baseline (Gold Standard). The close performance of Middle and Right empirically confirm Theorem 2.1: our method renders the unlearned data statistically undetectable.

4.4 Detector Evaluation: Empirical Verification of Guarantees

Theorem 2.1 implies that minimizing the mutual information $I(X_{\text{MarI}}; Z)$ squeezes the gap between the self-perplexity of $\mathcal{D}_u$ and the some retain subsets. Consequently, *any* confidence-based test (perplexity, cross-entropy, likelihood ratio) should fail to separate unlearned data from genuinely unseen data.

We focus on **white-box** detectors [3, 36], which utilize model logits and are strictly harder to defeat than black-box methods relying only on text similarity. (See Appendix D for more details on detection methods and additional results.) We evaluated the SOTA Min-K% detector [36] on three models: (i) Finetuned (before unlearning), (ii) Unlearn Baseline (Gold Standard), and (iii) Forgetting-MarI.

We report the ROC–AUC score, interpreted as follows: **Low ROC:** The detector confidently identifies the model as being trained on $\mathcal{D}_u$ (Unlearning Failure). **High ROC:** The detector classifies the model as *not* having seen $\mathcal{D}_u$, effectively matching the behavior of a model trained only on $\mathcal{D}_r$ (Unlearning Success).

As shown in Figure 5, the Finetuned model exhibits a low ROC, as expected for a model that has memorized the data. In contrast, Forgetting-MarI achieves a high ROC that closely matches the *gold-standard* Unlearn Baseline. This indicates that the detector is unable to distinguish our unlearned model from one that never saw the data in the first place, empirically confirming the undetectability predicted by our theory.

5 Conclusion

This work introduces *Forgetting-MarI*, a paradigm shift in LLM unlearning that moves beyond the heuristic "full-information" removal with utility regularization common in current approaches. By formalizing the concept of **Marginal Information**, we resolve the intrinsic conflict between forgetting a target set $\mathcal{D}_u$ and preserving the utility of a retain set $\mathcal{D}_r$. Our framework provides the first rigorous theoretical guarantees for LLMs unlearning, proving that minimizing marginal information suffices the undetectability of the unlearned data.

Our contributions pave the road for a new generation of theoretically grounded unlearning methods. While empirical results on Llama and GPT models confirm that our approach matches the "Gold Standard" of retraining, several critical avenues for research remain: **Relaxing Data Assumptions & Robust Fine-Tuning:** Currently, our definition of marginal information relies on access to $\mathcal{D}_r$. Future work could substitute $\mathcal{D}_r$ with synthetic data or new, independent datasets. This would not only enable "zero-shot" unlearning in privacy-sensitive deployments where $\mathcal{D}_r$ is inaccessible but also improve fine-tuning robustness by preventing overfitting to a static retain set. **Principled Hyperparameters:** Our theory relates λ to the classic rate-distortion in information theory. Future research could derive closed-form or learning-based solutions for optimal parameter selection, removing the need for validation loops.[‡]

[‡]During the preparation of this work, the authors used large language model ChatGPT by OpenAI to refine the language and enhance readability. After using this tool or service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] Amy B Cyphert. Generative AI, plagiarism, and copyright infringement in legal documents. *Minn. JL Sci. & Tech.*, 25:49, 2023.

[5] André V. Duarte, Xuandong Zhao, Arlindo L. Oliveira, and Lei Li. DE-COP: detecting copyrighted content in language models training data. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[6] Ronen Eldan and Mark Russinovich. Who's harry potter? approximate unlearning for LLMs, 2024. URL https://openreview.net/forum?id=PDct7vrcvT.

[7] Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. Alphaedit: Null-space constrained model editing for language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[8] Joshua Freeman, Chloe Rippe, Edoardo Debenedetti, and Maksym Andriushchenko. Exploring memorization and copyright violation in frontier LLMs: A study of the new york times v. openAI 2023 lawsuit. In *Neurips Safe Generative AI Workshop 2024*, 2024. URL https://openreview.net/forum?id=C66DBl9At8.

[9] Wenjie Fu, Huandong Wang, Chen Gao, Guanghua Liu, Yong Li, and Tao Jiang. Practical membership inference attacks against fine-tuned large language models via self-prompt calibration. *arXiv preprint arXiv:2311.06062*, 2023.

[10] Chongyang Gao, Lixu Wang, Chenkai Weng, Xiao Wang, and Qi Zhu. Practical unlearning for large language models. *arXiv e-prints*, pp. arXiv–2407, 2024.

[11] Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024. URL https://zenodo.org/records/12608602.

[12] Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. Making AI forget you: Data deletion in machine learning. *Advances in neural information processing systems*, 32, 2019.

[13] Michael M Grynbaum and Ryan Mac. The times sues openai and microsoft over ai use of copyrighted work. *The New York Times*, 27(1), 2023.

[14] Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. Certified data removal from machine learning models. *arXiv preprint arXiv:1911.03030*, 2019.

[15] Yihuai Hong, Yuelin Zou, Lijie Hu, Ziqian Zeng, Di Wang, and Haiqin Yang. Dissecting fine-tuning unlearning in large language models. *arXiv preprint arXiv:2410.06606*, 2024.

[16] Ruihan Hu, Yu-Ming Shang, Jiankun Peng, Wei Luo, Yazhe Wang, and Xi Zhang. Automated detection of pre-training text in black-box llms. *arXiv preprint arXiv:2506.19399*, 2025.

[17] Jevan Hutson, Cedric Whitney, and Jay T. Conrad. Forget me not? machine unlearning's implication for privacy law. *Science and Technology Law Review*, 27(1), Jan. 2026. doi: 10.52214/stlr.v27i1.14547. URL https://journals.library.columbia.edu/index.php/stlr/article/view/14547.

[18] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.

[19] Yoichi Ishibashi and Hidetoshi Shimodaira. Knowledge sanitization of large language models, 2024. URL https://arxiv.org/abs/2309.11852.

[20] Zhuoran Jin, Pengfei Cao, Chenhao Wang, Zhitao He, Hongbang Yuan, Jiachun Li, Yubo Chen, Kang Liu, and Jun Zhao. Rwku: Benchmarking real-world knowledge unlearning for large language models. *Advances in Neural Information Processing Systems*, 37:98213–98263, 2024.

[21] J.K. Rowling. *Harry Potter and the Prisoner of Azkaban*. Harry Potter. Scholastic, 2013. ISBN 9780545582933. URL https://books.google.com/books?id=uUp5mAEACAAJ.

[22] Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. Copyright violations and large language models. *arXiv preprint arXiv:2310.13771*, 2023.

[23] Bo Liu, Qiang Liu, and Peter Stone. Continual learning and private unlearning. In *Conference on Lifelong Learning Agents*, pp. 243–254. PMLR, 2022.

[24] Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, et al. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, pp. 1–14, 2025.

[25] Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. Towards safer large language models through machine unlearning. *arXiv preprint arXiv:2402.10058*, 2024.

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

[27] Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C Lipton, and J Zico Kolter. Tofu: A task of fictitious unlearning for llms. *arXiv preprint arXiv:2401.06121*, 2024.

[28] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

[29] Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *Proceedings of the 11th International Conference on Learning Representations (ICLR 2023)*, 2023.

[30] Cade Metz. Openai says new york times lawsuit against it is 'without merit'. *The New York Times (Digital Edition)*, pp. NA–NA, 2024.

[31] Haritz Puerto, Martin Gubri, Sangdoo Yun, and Seong Joon Oh. Scaling up membership inference: When and how attacks succeed on large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4165–4182, 2025.

[32] 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. *Advances in neural information processing systems*, 36:53728–53741, 2023.

[33] Joanne K Rowling. *Harry Potter and the sorcerer's stone*. Scholastic Incorporated, 2023.

[34] Wynn-Williams Sarah. Careless people: A cautionary tale of power, greed, and lost idealism, 2025.

[35] Shaojie Shi, Xiaoyu Tan, Xihe Qiu, Chao Qu, Kexin Nie, Yuan Cheng, Wei Chu, Xu Yinghui, and Yuan Qi. Ulmr: Unlearning large language models via negative response and model parameter average. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 755–762, 2024.

[36] Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models, 2024. URL https://arxiv.org/abs/2310.16789.

[37] 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.

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

[39] Shizhou Xu and Thomas Strohmer. Machine unlearning via information theoretic regularization. *arXiv preprint arXiv:2502.05684*, 2025.

[40] Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. *Advances in Neural Information Processing Systems*, 37:105425–105475, 2024.

[41] Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning, 2024. URL https://arxiv.org/abs/2404.05868.

[42] Weichao Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. Pretraining data detection for large language models: A divergence-based calibration method. *arXiv preprint arXiv:2409.14781*, 2024.

[43] Haolin Zou, Arnab Auddy, Yongchan Kwon, Kamiar Rahnama Rad, and Arian Maleki. Certified data removal under high-dimensional settings, 2025. URL https://arxiv.org/abs/2505.07640.

CONTENTS

A Appendix of Section 1 **15**

A.1 Details of LLM Unlearning Methods: Implicit Marginal Information Unlearning via conflicting forces 15

B Appendix of Section 2 **18**

B.1 Proof of Proposition 2.1 18

B.2 Why Mutual Information Rather Than KL Divergence 18

B.3 Proof of Theorem 2.1 19

B.4 Generalization of Theorem 2.1 to unlearn set neighborhood 19

C Appendix of Section 3 **22**

C.1 Theoretical error bound between position-wise vs. pooled MarI 22

C.2 Theoretical Guarantee provided by pooled MarI 23

C.3 Empirical error bound between position-wise vs. pooled MarI 23

D Appendix of Section 4 **25**

D.1 Comparison methods details 25

D.2 GPU, Computation and the Algorithm 25

D.3 Full Algorithm and Flow Chart 25

D.4 Ablation Study for the GPT2-LG 27

D.5 8B model performance 27

D.6 Supplemental to Sec 4.2 29

D.7 Supplementary General Model Capacity Test Results 30

E Appendix of Section 4.4: Detection Tests **32**

E.1 Detector Methods 32

E.2 Multiple Detection Test Results 32

Table 2: Comparison of LLM Unlearning Approaches

Approach	Key Methods	Strengths	Limitations
Full tuning	Loss-Rev.[1], GradDiff [2], KL-Dist.[3], DPO[4]/NPO[5]	Fine-grained control; preserves alignment	Risk of over-forgetting; utility drop
Weight edit	ROME[7], MEMIT[8], AlphaEdit[9], DEPN[10]	Fast; memory-light; layer-local patches	Less effective on distributed/stylistic info
Counterfactuals	IDK[11], EntityAnon.[12], ULMR[13], SKU[14]	No weight edits; easy deployment	Prompt-dependent; leakage if coverage is thin
Adaptation	Task-vec[15], O3[16]	Modular patches; base model unchanged	Adapters grow linearly; gate failures leak

†Numbers map to BibTeX entries: [1][40], [2][24], [3][15], [4][32], [5][41], [7][28], [8][29], [9][7], [10][38], [11][19], [12][6], [13][35], [14][25], [15][18], [16][10].

A Appendix of Section 1

A.1 Details of LLM Unlearning Methods: Implicit Marginal Information Unlearning via conflicting forces

Recent surveys highlight four broad families of LLM-unlearning techniques, each making a different compromise between *unlearn efficacy*, the ability to remove information from a model, *utility preservation*, how well the model performs on the remaining data, and *computational cost*, the resources expended to perform the unlearning [24]. A heuristic commonality of the techniques is their implicit/indirect target of marginal unlearning: all the methods tend to detect and thereby remove only the marginal effect of adding an "unlearn set" ($\mathcal{D}_u$), the dataset that is meant to be forgotten, to a "retain set" ($\mathcal{D}_r$), the dataset that the model should remember, on the given model.

Full parameter fine-tuning: These techniques train and perform weight updates on the whole model. Gradient ascent (or "loss reversal") [40] is the most straight-forward unlearning technique. It directly maximizes the cross-entropy on $\mathcal{D}_u$, effectively penalizing the model performance on the unlearn set. This type of unlearning was shown to lead to an overall decrease in model performance, so Gradient Difference [24] was developed to balance unlearning while maintaining general model performance. Gradient Difference maximizes the cross-entropy loss on $\mathcal{D}_u$ while continuing to *minimize* the loss on $\mathcal{D}_r$:

$$\min_{\theta} \underbrace{\mathbb{E}_{x\in\mathcal{D}_r}\ell(\theta;x)}_{\text{utility}} - \lambda \underbrace{\mathbb{E}_{x\in\mathcal{D}_u}\ell(\theta;x)}_{\text{loss reversal}},$$

$$\ell(\theta;x) = \mathrm{CE}\left(p_\theta(\,\cdot \mid x_{<t}), x_t\right).$$

Here $\lambda > 0$ balances utility preservation and unlearning. Intuitively, gradient descent is applied on $\mathcal{D}_r$ while gradient *ascent* is applied on $\mathcal{D}_u$.

Follow-up studies revealed that, even when balanced with gradient descent, this global ascent signal is too coarse: it suppresses the target examples but also degrades correlated yet legitimate content [15]. To overcome this challenge, variants have aimed to improve both sides of the problem. For utility preservation, past work has shown that distillation-style regularization with a Kullback–Leibler (KL) divergence penalty outperforms gradient descent on $\mathcal{D}_r$ in keeping the updated model close to the original without over-training on the retain set. For unlearning, alignment-style variants such as Direct Preference Optimization (DPO) and Negative Preference Optimization (NPO) replace the unlearn objective with more specific preference-based objectives, slowing catastrophic performance collapse [32, 41]. However, such preference-supervised methods can be difficult to generalize to unlearn at a large scale.

Finally, from the perspective of *marginal* unlearning, these full-parameter objectives act as *indirect proxies* for the marginal effect of adding $\mathcal{D}_u$ to $\mathcal{D}_r$: they rely on carefully balancing ascent on $\mathcal{D}_u$ and descent (or KL regularization) on $\mathcal{D}_r$. In practice, such proxies can be neither the most *effective* nor the most *efficient* at isolating the unique contribution of $\mathcal{D}_u$ without erasing information shared with $\mathcal{D}_r$.

Weight editing and partial tuning: In an effort to perform unlearning more efficiently, this line of methods focuses on selectively altering only a subset of a model's parameters rather than retraining

the entire network. Such "model-surgery" methods perform rank-constrained updates at one or a few layers. Rank-One Model Editing (ROME) edits a single MLP weight with a closed-form rank-1 patch [28]. In particular, it modifies only the weights causally responsible for one token sequence in the unlearn set:

$$\min_{\Delta W} \|\Delta W\|_F^2 \text{ s.t. } W_{l^\star} h_{l^\star}(x) + \Delta W h_{l^\star}(x) = v_{\text{new}}.$$

Here, $\Delta W := \frac{\left(v_{\text{new}} - v_{\text{old}}\right) h^\top}{\|h\|_2^2}$, l^* is the layer most influenced by the unlearn sample or prompt x, $h_{l^*}(x)$ is the activation and W_{l^*} is the weight matrix of layer l^*, $v_{old} := W_{l^*} h_{l^*}(x)$, and finally v_{new} is the alternative answer we want to replace v_{old} by. Mass Editing Memory in a Transformer (MEMIT) [29] extends this idea to thousands of facts simultaneously and stacks many $(h_{l^*}^i, v_{l^*}^i)$ pairs. AlphaEdit furthers the idea by projecting edits into the null space of preserved knowledge, with the aim to improve robustness in sequential settings, ensuring minimal disruption to previously learned information. Detecting and Editing Privacy Neurons (DEPN) [38] masks the gradients of neurons identified as contributing the most to the prediction of privacy-related content. In general, weight editing and partial tuning techniques are fast, but they are limited to short factual associations and struggle with stylistic or distributed knowledge.

Finally, the above weight-editing and partial-tuning methods share a common *indirect marginal unlearning* proxy: they infer marginal information by targeting parameters most influenced by the unlearn set, while largely ignoring parameters most influenced by the retain set. This can help isolate some marginal information signal, but again risks overlooking deep interactions between r and u.

Curating counterfactuals: Instead of directly unlearning all or part of the model, another approach is to substitute the parametric knowledge of the unlearn set with benign knowledge. Broadly, this class of methods can be characterized by:

$$\min_{\theta} \underbrace{\mathbb{E}_{x \in \mathcal{D}_r}\big[\ell(\theta; x)\big]}_{\text{retain utility}} + \lambda \underbrace{\mathbb{E}_{x \in \mathcal{D}_{\text{neg}}}\big[\ell(\theta; x)\big]}_{\text{counterfactual prompts}},$$

where $\mathcal{D}_{\text{neg}}$ contains prompts or contexts designed to *neutralize* the influence of the unlearn set, $\ell(\theta; x)$ is the same cross entropy loss as before, and $\lambda > 0$ balances unlearning against utility.

"I don't know" [19] trains the model on question-answer pairs that map sensitive questions to a safe refusal (e.g. "I don't know"), teaching the model to decline queries about the unlearn set. Entity anonymization [6] replaces sensitive entities with anonymized placeholders and trains the model on the rewritten placeholders to scrub identifiable information from the model. Unlearning Large Language Models via Negative Response and Model Parameter Average (ULMR) [35] constructs adversarial "negative" prompts, trains on the paired responses, and then averages the updated weights with the base model to dampen overshoot. Selective Knowledge-negation Unlearning (SKU) first mines harmful or copyrighted contexts via red-teaming, then injects counterfactuals that negate them [25]. Such approaches are easy to deploy but depend heavily on prompt engineering and high-quality counterexamples.

From a marginal information proxy perspective, the curating counterfactuals approach aims to first penalize model utility related to the unlearn set by replacing the original model capability on the unlearn set with a lower-utility capacity on the counterfactuals, then rescue the utility related to the retain set using the utility preservation term, and finally balance the two to indirectly find the marginal information and penalize it.

Model adaptation: These methods train something external to the model and then use that externally trained adapter to update the model itself. A common instantiation is the task-vector framework: let p_{θ_0} be the original model and p_{θ_u} the same model fine-tuned on the unlearn set $\mathcal{D}_u$. The element-wise difference $\Delta\theta := \theta_u - \theta_0$ is treated as an encoding of the deleted knowledge and direct-subtraction methods [18] form the unlearned model as $p_{\theta_0 - \Delta\theta}$. Orthogonality offers an alternative geometric control. O^3 [10] trains one orthogonal LoRA adapter per removal request and learns a contrastive out-of-distribution (OOD) gate that activates the corresponding adapter at inference time. Orthogonality limits interference between requests, but the approach incurs two key costs: (1) the number of adapters (and hence memory) grows linearly with the number of unlearning requests, and (2) any mismatch between model behavior and the assumed linear/inner-product structure in weight space can undermine both unlearning guarantees and downstream utility.

From a *marginal-information* viewpoint, model-adaptation methods isolate the contribution of $\mathcal{D}_u$ by (i) subtracting the unlearn-induced task vector from the retain base, $\theta_0 - \Delta\theta$, or (ii) enforcing orthogonality between components aligned with retain and the unlearn signals and then penalize the isolated component. Both approaches can be considered as proxy of marginal information, though with strong arithmetic or geometric assumptions.

The proposed method, Forgetting-MarI, belongs to the full-parameter fine-tuning category. It applies a "marginal information" penalty that suppresses only the influence of the unlearn set while leaving the shared information, which is supported by the retain data, largely intact.

B Appendix of Section 2

B.1 Proof of Proposition 2.1

Proof. To start, define the Bayes error as

$$P_e := \mathbb{E}_{X_{\mathrm{MarI}}}\Big[\min\big\{P(Z=0 \mid X_{\mathrm{MarI}}),\, P(Z=1 \mid X_{\mathrm{MarI}})\big\}\Big] = 1 - P_{acc}.$$

In addition, for each x, let $p(x) := P(Z=1 \mid X_{\mathrm{MarI}} = x) \in [0,1]$ be the conditional probability of $\{Z=1\}$ given $\{X_{\mathrm{MarI}} = x\}$. Then it follows from Z being binary that $H(Z \mid X_{\mathrm{MarI}} = x) = H_2\big(p(x)\big)$. Denote $m(X_{\mathrm{MarI}}) := \min\big\{P(Z=0 \mid X_{\mathrm{MarI}}),\, P(Z=1 \mid X_{\mathrm{MarI}})\big\}$. Since H_2 is concave, it follows from Jensen's inequality that

$$\begin{aligned} H(Z \mid X_{\mathrm{MarI}}) &= \mathbb{E}_{X_{\mathrm{MarI}}}\big[H_2(p(X_{\mathrm{MarI}}))\big] \\ &= \mathbb{E}_{X_{\mathrm{MarI}}}\big[H_2(m(X_{\mathrm{MarI}}))\big] \\ &\le H_2\big(\mathbb{E}_{X_{\mathrm{MarI}}}[m(X_{\mathrm{MarI}})]\big) \\ &= H_2(P_e). \end{aligned}$$

where the second equality holds due to the fact that $H_2(p) = H_2(1-p)$. Now, since $I(X_{\mathrm{MarI}}; Z) = H(Z) - H(Z \mid X_{\mathrm{MarI}})$ and $H(Z) = H_2(\pi)$, we obtain

$$H_2(\pi) - I(X_{\mathrm{MarI}}; Z) = H(Z \mid X_{\mathrm{MarI}}) \;\le\; H_2(P_e).$$

Since $P_e \in [0, \frac{1}{2}]$ and H_2 is strictly increasing on this interval, by applying the inverse H_2^{-1}, we have

$$P_e \;\ge\; H_2^{-1}\big(H_2(\pi) - I(X_{\mathrm{MarI}}; Z)\big).$$

Finally, by $P_{acc} = 1 - P_e$, we have

$$P_{acc} \le 1 - H_2^{-1}\big(H_2(\pi) - I(X_{\mathrm{MarI}}; Z)\big).$$

This proves the stated inequality. The particular case $\pi = \frac{1}{2}$ follows from $H_2(\frac{1}{2}) = 1$.

It remains to show that the upper bound is tight. Indeed, fix an arbitrary $I \in [0, H_2(\pi)]$. Choose $p^\star \in \big[\frac{1}{2}, 1\big]$ such that $H_2(p^\star) = H_2(\pi) - I$. Construct $P_{Z|X_{\mathrm{MarI}}}$ such that $P(Z=1 \mid X_{\mathrm{MarI}}) \in \{p^\star,\, 1-p^\star\}$ with probabilities chosen to match the prior π. Then $H(Z \mid X_{\mathrm{MarI}}) = H_2(p^\star)$ and $I(X_{\mathrm{MarI}}; Z) = I$, while the Bayes error satisfies $P_e = \min\{p^\star, 1-p^\star\} = H_2^{-1}(H(Z) - I)$. Hence, equality holds in the bound. □

B.2 Why Mutual Information Rather Than KL Divergence

One might consider penalizing a directional KL divergence between the "to-unlearn" and "to-retain" distributions. Instead, we regularize the *mutual information* between the model output and a binary indicator of sensitive content, which is equal to the Jensen-Shannon divergence as shown in Section 2. Here, we show that mutual information offers several advantages over one–way or two-way KL divergence:

- *Flexibility for utility and continual unlearning.* The reference m in Jensen-Shannon divergence is the *mixture* of the two conditionals and evolves with training; we do not assume a fixed "gold" model. This yields a pure unlearning regularizer that can be combined with any utility term (e.g., $\ell_{\mathrm{KL}}(\theta, r)$) and naturally supports continual/online updates.
- *Stable training signal.* $I(\hat{X}; Z) \le H_2(\pi) \le \log 2$ for binary Z, so the gradients remain well-behaved even when supports differ, unlike one–way KL which can be unbounded on support mismatch.
- *Downstream robustness via data processing.* For any downstream representation or task $T = g(\hat{X})$, the data-processing inequality gives $I(T; Z) \le I(\hat{X}; Z)$. Thus, suppressing $I(\hat{X}; Z)$ at the model output (or an internal layer) upper-bounds leakage throughout the pipeline.

In contrast, a directional KL requires committing to a *fixed* target (encoding a specific utility assumption) and can be unstable or unbounded when supports are disjoint. That said, if an ideal frozen reference is indeed mandated, a one–way KL to that reference is a reasonable alternative.

B.3 Proof of Theorem 2.1

Here, we provide the proof for Theorem 2.1:

Proof. By the mean value theorem, for each t there exists $\xi_t \in [\min\{p_t^u(u_t), p_t^r(u_t)\}, 1] \subseteq [\gamma, 1]$ such that

$$\begin{aligned}\left|\log p_t^u(u_t) - \log p_t^r(u_t)\right| &= \frac{\left|p_t^u(u_t) - p_t^r(u_t)\right|}{\xi_t} \\ &\leq \frac{\left|p_t^u(u_t) - p_t^r(u_t)\right|}{\gamma} \\ &\leq \frac{\|p_t^u - p_t^r\|_1}{\gamma} \\ &= \frac{2\,\|p_t^u - p_t^r\|_{TV}}{\gamma}.\end{aligned}$$

Averaging over t,

$$\left|S_\theta(u,u) - S_\theta(u,r)\right| \;\leq\; \frac{2}{\gamma}\,\frac{1}{T}\sum_{t=1}^{T}\|p_t^u - p_t^r\|_{TV}.$$

Apply Lemma B.3 followed by Lemma B.2 and Jensen's inequality:

$$\frac{1}{T}\sum_t \|p_t^u - p_t^r\|_{TV} \;=\; \frac{1}{1-\alpha}\,\frac{1}{T}\sum_t \|p_t^d - p_t^r\|_{TV} \;\leq\; \frac{\sqrt{2}}{1-\alpha}\sqrt{\frac{1}{T}\sum_t \mathrm{JSD}(p_t^d, p_t^r)}.$$

Combining yields the claim. □

B.4 Generalization of Theorem 2.1 to unlearn set neighborhood

Here, we show that the self-perplexity gap guarantee provided by Theorem 2.1 can be generalized to a neighborhood of u rather than u itself:

Theorem B.1 (MarI controls neighborhood-perplexity gap)**.** *Draw $U := \{U_t\}_{t=1}^T$ with $U_t \sim p_t^u$ independently across $t \in [T]$ and suppose $\max_{t,x}\left\{\frac{p_t^u(x)}{p_t^r(x)} \vee \frac{p_t^r(x)}{p_t^u(x)}\right\} =: M < \infty$. Let $C := \max_{t,x:\,p_t^u(x)>0}\left[\log\frac{p_t^r(x)}{p_t^u(x)}\right]^2 < \infty$. Then, for any $\varepsilon > 0$, with probability at least $1 - 2\exp\left(-T\varepsilon^2/(2C)\right)$,*

$$\left|S_\theta(U,u) - S_\theta(U,r)\right| \;\leq\; \Big(\log M\Big)\frac{M}{M-1}\,\frac{\sqrt{2}}{1-\alpha}\,\sqrt{I(X_{\mathrm{MarI}};Z)} \;+\; \varepsilon.$$

We start with the following three lemmata that are needed for the proof of Theorem B.1:

Lemma B.1 (Point-wise KL bound)**.** *Let p, q be two probability distributions over a finite set V such that $\frac{p(x)}{q(x)} \in [1, M]$ for every $x \in V$ for some constant $M > 1$. Then for every $x \in V$*

$$p(x)\,\log\frac{p(x)}{q(x)} \;\leq\; (\log M)\frac{M}{M-1}\,\big[p(x) - q(x)\big]. \tag{4}$$

Proof. Fix $x \in V$ and set $y := \dfrac{p(x)}{q(x)} \in [1, M]$. Inequality equation 4 is equivalent to

$$y\log y \;\leq\; \frac{M}{M-1}\,(\log M)\,(y-1), \qquad \forall\, y \in [1, M]. \tag{4}$$

For $y > 1$ let $g(y) := \frac{y\log y}{y-1}$ and set $g(1) := \lim_{y\to 1^+} g(y) = 1$. We show that g is strictly increasing on $[1, M]$. Indeed, compute $g'(y) = \frac{(y-1)-\log y}{(y-1)^2}$. Since $\log y < y - 1$ for all $y > 1$, we have $g'(y) > 0$; thus, g is strictly increasing. Because g is increasing and $y \in [1, M]$, we have

$$g(y) \;\leq\; g(M) = \frac{M\log M}{M-1}.$$

Multiplying both sides by $y - 1$ yields equation 4, which is precisely equation 4 after reinstating $y = p(x)/q(x)$. Therefore, equation 4 holds for every $x \in V$. This completes the proof. □

Lemma B.2. (Total Variation is controlled by Jensen-Shannon Divergence) *For any two probability measures p, q on a finite set, we have*

$$\|p - q\|_{TV} \;\le\; \sqrt{2\ \mathrm{JSD}(p, q)},$$

where $\mathrm{JSD}(p, q) := \frac{1}{2} D_{\mathrm{KL}}(p\|m) + \frac{1}{2} D_{\mathrm{KL}}(q\|m)$, $m := \frac{p+q}{2}$ *and* $D_{\mathrm{KL}}(p\|q) := \sum_v p(v) \log \frac{p(v)}{q(v)}$, *denotes the Jensen–Shannon divergence.*

Proof. Let $m = \frac{p+q}{2}$. Pinsker's inequality gives $\|p - m\|_1^2 \le 2\, D_{\mathrm{KL}}(p\|m)$ and analogously for q. Hence

$$\mathrm{JSD}(p, q) \;\ge\; \tfrac{1}{4}\Big[\|p - m\|_1^2 + \|q - m\|_1^2\Big] \;=\; \tfrac{1}{8}\, \|p - q\|_1^2,$$

because $p - m = \frac{p-q}{2}$ and $q - m = -\frac{p-q}{2}$. Since $\|p - q\|_{TV} = \frac{1}{2}\|p - q\|_1$, it follows that $\|p - q\|_{TV}^2 \le 2\ \mathrm{JSD}(p, q)$. □

Lemma B.3 (Exact TV scaling under mixture)**.** *If $p^d = \alpha p^r + (1 - \alpha) p^u$ with $\alpha \in (0, 1)$, then*

$$\|p^u - p^r\|_{TV} \;=\; \frac{1}{1 - \alpha}\, \|p^d - p^r\|_{TV}.$$

Proof. $p^d - p^r = (1 - \alpha)(p^u - p^r)$. Taking ℓ_1-norms and dividing by 2 yields the identity. □

Now, we are ready to prove Theorem B.1:

Proof. Define $Y_t := \log \frac{p_t^r(U_t)}{p_t^u(U_t)}$, so that

$$S_\theta(U, u) - S_\theta(U, r) \;=\; \frac{1}{T} \sum_{t=1}^{T} Y_t.$$

Since $U_t \sim p_t^u$, $\mathbb{E}[Y_t] = \sum_x p_t^u(x) \log \frac{p_t^r(x)}{p_t^u(x)} = -\, D_{\mathrm{KL}}(p_t^u\|p_t^r)$, hence

$$\mathbb{E}\big[S_\theta(U, u) - S_\theta(U, r)\big] = -\frac{1}{T} \sum_{t=1}^{T} D_{\mathrm{KL}}(p_t^u\|p_t^r).$$

Now, by the assumption $\max_{t,x} \max\Big\{\frac{p_t^u(x)}{p_t^r(x)}, \frac{p_t^r(x)}{p_t^u(x)}\Big\} \;\le\; M$, we have $p_t^r(x) > 0$ for $p_t^u(x)$-a.e. x for all t. Therefore, for all t, we have $\log \frac{p_t^r(x)}{p_t^u(x)} < \infty$ and taking the maximum over $t \in [T]$, we obtain $C := \max_{t,\, x:\, p_t^u(x) > 0} \Big[\log \frac{p_t^r(x)}{p_t^u(x)}\Big]^2 < \infty$. It then follows from the definition of Y_t that $|Y_t| \le \sqrt{C}$ a.s.. Hoeffding's inequality for independent bounded variables yields, for any $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T} Y_t - \mathbb{E}\frac{1}{T}\sum_{t=1}^{T} Y_t\right| \ge \varepsilon\right) \;\le\; 2\exp\left(-\frac{T\,\varepsilon^2}{2C}\right).$$

Using $\big||a| - b\big| \le |a - b|$ for $b \ge 0$, we have

$$\big|S_\theta(U, u) - S_\theta(U, r)\big| \;\le\; \frac{1}{T} \sum_{t=1}^{T} D_{\mathrm{KL}}(p_t^u\|p_t^r) + \varepsilon.$$

with probability at least $1 - 2\exp(-\frac{T\,\varepsilon^2}{2C})$.

Now, for each t, let $A_t = \{x :\, p_t^u(x) \ge p_t^r(x)\}$. Then by Lemma B.1, we have

$$D_{\mathrm{KL}}(p_t^u\|p_t^r) \;\le\; \kappa(M) \sum_{x \in A_t} \big(p_t^u(x) - p_t^r(x)\big) \;\le\; \kappa(M)\, \|p_t^u - p_t^r\|_{TV}.$$

Averaging in t gives

$$\frac{1}{T}\sum_{t=1}^{T} D_{\mathrm{KL}}(p_t^u \| p_t^r) \;\le\; \kappa(M)\,\frac{1}{T}\sum_{t=1}^{T} \|p_t^u - p_t^r\|_{TV}.$$

Finally, it follows from Lemma B.3 and Lemma B.2 that

$$\frac{1}{T}\sum_{t=1}^{T} \|p_t^u - p_t^r\|_{TV} \;=\; \frac{1}{1-\alpha}\,\frac{1}{T}\sum_{t=1}^{T} \|p_t^d - p_t^r\|_{TV} \;\le\; \frac{\sqrt{2}}{1-\alpha}\,\frac{1}{T}\sum_{t=1}^{T} \sqrt{\mathrm{JSD}(p_t^d, p_t^r)}.$$

By Jensen's inequality, $\frac{1}{T}\sum_t \sqrt{\mathrm{JSD}(p_t^d, p_t^r)} \;\le\; \sqrt{\frac{1}{T}\sum_t \mathrm{JSD}(p_t^d, p_t^r)}$. Combining the displays proves the claim with $I(X_{\mathrm{MarI}}; Z) = \frac{1}{T}\sum_t \mathrm{JSD}(p_t^d, p_t^r)$. □

C Appendix of Section 3

C.1 Theoretical error bound between position-wise vs. pooled MarI

Here, we provide the theoretical analysis of the error between MarI and pooled MarI. The following result shows that, under a mild assumption, the error of using the pooled MarI to estimate MarI is bounded by the sequence-wise density variance:

Theorem C.1 (Pooling error bound)**.** *For each t, set $m_t := \frac{1}{2}(p_t^d + p_t^r)$ and $\bar{m} := \frac{1}{2}(\bar{p}^{\,d} + \bar{p}^{\,r})$, where $\bar{p}^{\,d} := \frac{1}{T}\sum_{t=1}^{T} p_t^d$ and $\bar{p}^{\,r} := \frac{1}{T}\sum_{t=1}^{T} p_t^r$. Assume the uniform overlap condition*

$$\beta \;:=\; \min\Bigg\{ \inf_{\lambda\in[0,1]} \min_{t,x}\big[(1-\lambda)\, m_t(x) + \lambda\, \bar{m}(x)\big], \tag{5}$$

$$\inf_{\lambda\in[0,1]} \min_{t,x}\big[(1-\lambda)\, p_t^d(x) + \lambda\, \bar{p}^{\,d}(x)\big],\; \inf_{\lambda\in[0,1]} \min_{t,x}\big[(1-\lambda)\, p_t^r(x) + \lambda\, \bar{p}^{\,r}(x)\big] \Bigg\} \;>\; 0. \tag{6}$$

Define the (averaged) ℓ_2-deviation terms

$$V_d \;:=\; \frac{1}{T}\sum_{t=1}^{T} \big\|p_t^d - \bar{p}^{\,d}\big\|_2^2, \qquad V_r \;:=\; \frac{1}{T}\sum_{t=1}^{T} \big\|p_t^r - \bar{p}^{\,r}\big\|_2^2.$$

Then

$$0 \;\le\; I(X_{\mathrm{MarI}}; Z) \;-\; I(\bar{X}_{\mathrm{MarI}}; Z) \;\le\; \frac{1}{4\,\beta}\Big(V_d + V_r\Big), \tag{7}$$

where $I(X_{\mathrm{MarI}}; Z) = \frac{1}{T}\sum_{t=1}^{T} \mathrm{JSD}(p_t^d, p_t^r)$ and $I(\bar{X}_{\mathrm{MarI}}; Z) = \mathrm{JSD}(\bar{p}^{\,d}, \bar{p}^{\,r})$.

Proof. The lower bound $I(\bar{X}_{\mathrm{MarI}}; Z) \;\le\; I(X_{\mathrm{MarI}}; Z)$ follows directly from the data-processing inequality. For the upper bound, write

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T} \mathrm{JSD}(p_t^d, p_t^r) \;-\; \mathrm{JSD}(\bar{p}^{\,d}, \bar{p}^{\,r}) \\
=&\frac{1}{T}\sum_{t=1}^{T}\Big[\underbrace{H(m_t) - H(\bar{m})}_{(A_t)} - \frac{1}{2}\underbrace{\big(H(p_t^d) - H(\bar{p}^{\,d})\big)}_{(B_t)} - \frac{1}{2}\underbrace{\big(H(p_t^r) - H(\bar{p}^{\,r})\big)}_{(C_t)}\Big].
\end{aligned}$$

Let $\Delta_m := m_t - \bar{m}$, $\Delta_d := p_t^d - \bar{p}^{\,d}$, $\Delta_r := p_t^r - \bar{p}^{\,r}$. Since H is twice-differentiable, the second-order Taylor expansion around the pooled densities yields

$$H(a) \;=\; H(a') + \langle \nabla H(a'), a - a' \rangle + \tfrac{1}{2}(a - a')^\top \nabla^2 H(s)(a - a'),$$

for some s on the line segment between a' and a. Note that H is concave, $\nabla^2 H(s)$ is negative semidefinite and diagonal with entries $-1/s(x)$. By the overlap assumption equation 5, every coordinate along the segments between m_t and $\bar{m}$ and between p_t^s and $\bar{p}^{\,s}$ ($s \in \{d, r\}$) is at least β, hence

$$-\,(a - a')^\top \nabla^2 H(s)(a - a') \;\le\; \frac{1}{\beta}\sum_{x\in V}\big(a(x) - a'(x)\big)^2 \;\le\; \frac{1}{\beta}\,\|a - a'\|_2^2,$$

and therefore the (negative) Taylor remainders satisfy

$$H(a) - H(a') - \langle \nabla H(a'), a - a' \rangle \;\ge\; -\,\frac{1}{2\beta}\,\|a - a'\|_2^2. \tag{8}$$

Applying equation 8 to the three entropy differences and averaging in t, the first-order terms vanish because $\frac{1}{T}\sum_t \Delta_d = 0$, $\frac{1}{T}\sum_t \Delta_r = 0$, and $\frac{1}{T}\sum_t \Delta_m = 0$. Thus,

$$\begin{aligned}\frac{1}{T}\sum_t\big[(A_t)-(B_t)/2-(C_t)/2\big] \le& \frac{1}{T}\sum_t \frac{1}{2}\Big(-\mathcal{R}_d(t)\Big) \;+\; \frac{1}{T}\sum_t \frac{1}{2}\Big(-\mathcal{R}_r(t)\Big) - \frac{1}{T}\sum_t \mathcal{R}_m(t)\\ \le& -\frac{1}{2}\Big(\frac{1}{T}\sum_t\Big(\mathcal{R}_d(t)\Big) \;+\; \frac{1}{T}\sum_t\Big(\mathcal{R}_r(t)\Big)\Big)\\ \le& \frac{1}{4\beta}\Big(\frac{1}{T}\sum_t\Big(||p_t^d-\bar{p}^d||_2^2\Big) \;+\; \frac{1}{T}\sum_t\Big(||p_t^r-\bar{p}^r||_2^2\Big)\Big)\\ =& \frac{1}{4\beta}\,(V_d+V_r).\end{aligned}$$

where $\mathcal{R}_s(t) := H(p_t^s) - H(\bar{p}^s) - \langle \nabla H(\bar{p}^s), p_t^s - \bar{p}^s\rangle \le 0$ for $s \in \{d, r\}$ and $\mathcal{R}_m(t) := H(m_t) - H(\bar{m}) - \langle \nabla H(\bar{m}), m_t - \bar{m}\rangle \le 0$. □

C.2 Theoretical Guarantee provided by pooled MarI

In general, there is *no non-trivial (distribution–free)* theoretical guarantee for the sequence–level perplexity gaps can be stated solely in terms of the pooled MarI $I(\bar{X}_{\mathrm{MarI}}; Z) = \mathrm{JSD}(\bar{p}^{\,d}, \bar{p}^{\,r})$ without any additional assumption, such as the bounded ℓ^2-deviation in Theorem C.1 above.

Nonetheless, the pooled MarI *can* certify the following *word–level forgetting* (for particular tokens) forgetting:

Theorem C.2 (Word-level provable unlearning via pooled MarI)**.** *Fix a token $w \in V$ and assume $\bar{p}^{\,u}(w) \wedge \bar{p}^{\,r}(w) =: \bar{\gamma}_w \in (0, 1]$. Then*

$$\Big|\,\log \bar{p}^{\,u}(w) - \log \bar{p}^{\,r}(w)\,\Big| \;\le\; \frac{2}{\bar{\gamma}_w\,(1-\alpha)}\,\sqrt{2\,I(\bar{X}_{\mathrm{MarI}}; Z)}\,.$$

Proof. It follows from mean value theorem for $x \mapsto \log x$ on $[\bar{\gamma}_w, 1]$ that

$$|\log \bar{p}^{\,u}(w) - \log \bar{p}^{\,r}(w)| \le |\bar{p}^{\,u}(w) - \bar{p}^{\,r}(w)|/\bar{\gamma}_w.$$

Since $\bar{p}^{\,u} - \bar{p}^{\,r} = (\bar{p}^{\,d} - \bar{p}^{\,r})/(1-\alpha)$, we have

$$|\bar{p}^{\,u}(w) - \bar{p}^{\,r}(w)| \;\le\; \|\bar{p}^{\,u} - \bar{p}^{\,r}\|_1 = \frac{1}{1-\alpha}\,\|\bar{p}^{\,d} - \bar{p}^{\,r}\|_1 = \frac{2}{1-\alpha}\,\|\bar{p}^{\,d} - \bar{p}^{\,r}\|_{TV}.$$

Finally, combining the above two inequalities, we have

$$|\log \bar{p}^{\,u}(w) - \log \bar{p}^{\,r}(w)| \le \frac{2}{\bar{\gamma}_w(1-\alpha)}\,\|\bar{p}^{\,d} - \bar{p}^{\,r}\|_{TV} \le \frac{2}{\bar{\gamma}_w(1-\alpha)}\,\sqrt{2\,I(\bar{X}_{\mathrm{MarI}}; Z)},$$

where the last inequality follows from the fact that

$$\|\bar{p}^{\,d} - \bar{p}^{\,r}\|_{TV} \le \sqrt{2\ \mathrm{JSD}(\bar{p}^{\,d}, \bar{p}^{\,r})} = \sqrt{2\,I(\bar{X}_{\mathrm{MarI}}; Z)}.$$

□

C.3 Empirical error bound between position-wise vs. pooled MarI

We empirically compare the token/position-wise MarI, $I(X_{\mathrm{MarI}}; Z) = \frac{1}{T}\sum_{t=1}^{T} I(X_t; Z)$, with the pooled ("flattened") MarI, $I(\bar{X}_{\mathrm{MarI}}; Z)$, on our heterogeneous dataset. As predicted by the data–processing inequality, $I(\bar{X}_{\mathrm{MarI}}; Z) \le I(X_{\mathrm{MarI}}; Z)$, so the position-wise estimator produces a stronger marginal-information signal. Nevertheless, by appropriately tuning the trade-off parameter γ (weighting MarI vs. utility), both estimators attain comparable forget–utility trade-offs.

However, we can also observe the influence of the heterogeneity of dataset and random batch sampling. In particular, in Figure 6, the position-wise estimator exhibits higher variance on heterogeneous batches (varying lengths, topics, and token alignments). Furthermore, Figure 7 shows that,

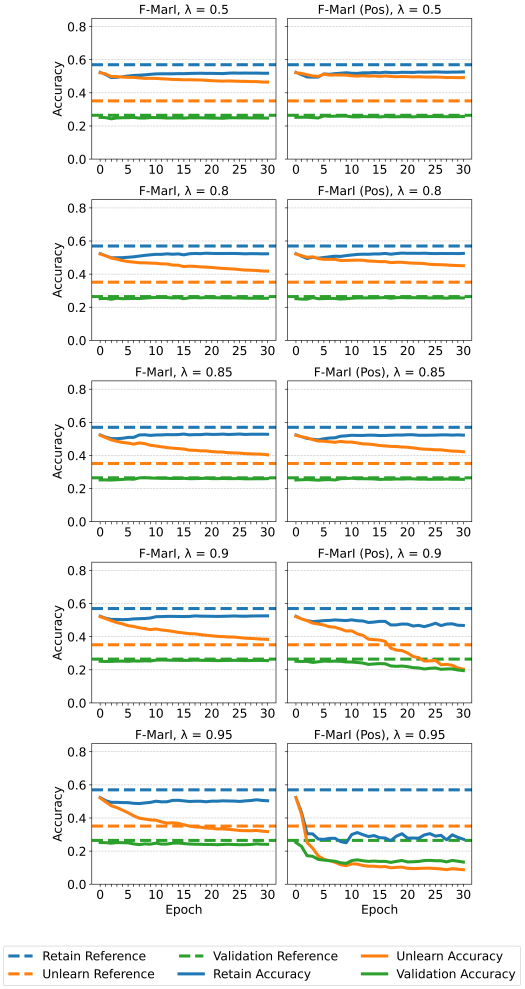


Figure 6: Position-wise vs. pooled MarI under several λ settings using Llama models.

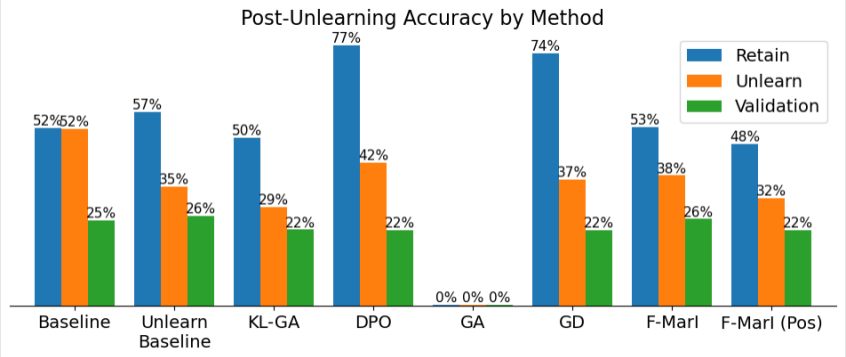


Figure 7: With a fixed trade-off ($\lambda = 0.9$) on LLama models, position-wise MarI is noisier on heterogeneous data and over-unlearns compared to the unlearn baseline.

with fixed γ (e.g., $\gamma = 0.9$), the position-wise MarI tends to over-unlearn relative to the gold unlearn baseline. Intuitively, it can over-penalize idiosyncratic, position-specific fluctuations rather than true marginal effects.

In our experiments, because the text is heterogeneous in both length and context, we use random mini-batches and the pooled estimator by default.

D APPENDIX OF SECTION 4

D.1 COMPARISON METHODS DETAILS

Method	Unlearn objective	Retain objective	Unlearn nature
GA [40]	ascent on unlearn set	—	full information
GD [23]	ascent on unlearn set	descent on retain set	full information
KL-GA [15]	ascent on unlearn set	minimize $\mathrm{KL}\big(p^r_\theta \,\|\, p^r_{\theta_0}\big)$	full information
DPO [32]	preference loss	minimize CE or KL	full information
Forgetting–MarI	minimize $I(X_{\mathrm{MarI}}; Z)$	minimize $\mathrm{KL}\big(p^r_\theta \,\|\, p^r_{\theta_0}\big)$	marginal information

Table 3: Comparison of LLM unlearning objectives. CE = cross-entropy; KL = Kullback–Leibler divergence. "Direct/Indirect" indicates whether the method explicitly penalizes marginal information (ours) or approximates it by balancing forget/retain signals.

D.2 GPU, COMPUTATION AND THE ALGORITHM

For the GPT2-LG model, all experiments use 4×NVIDIA A100-40GB GPUs in fp32 precision. For all unlearning methods, we use a per-GPU batch size of 8 for both the unlearn and retain sets, yielding an effective batch size of 32 examples for each set per optimization step.

Method	Time per batch (s / batch)	Peak memory (GB / GPU)
F-MarI	0.56	34.19
KL-GA	0.50	31.22
GD	0.45	27.15
DPO	0.57	36.35
GA	0.32	28.42

Table 4: Compute cost for different unlearning methods on GPT-2 Large. We report the average time per batch, and peak per-GPU memory usage.

During unlearning of the CP using Llama, the batch size was 3 for the unlearn set and 6 for the retain set. All experiments were conducted using 2 × NVIDIA GRID T4-16Q 16GB GPUs at maximum memory utilization.

Method	Time per batch (s / batch)	Peak memory (GB / GPU)
F-MarI	1.95	∼16
KL-GA	1.93	∼16
GD	1.50	∼16
DPO	1.76	∼16
GA	0.63	∼16

D.3 FULL ALGORITHM AND FLOW CHART

Here, we first provide both the full-detailed algorithm pseudo-code for Forgetting-MarI and the flowchart for readers who are more familiar with chart presentations.

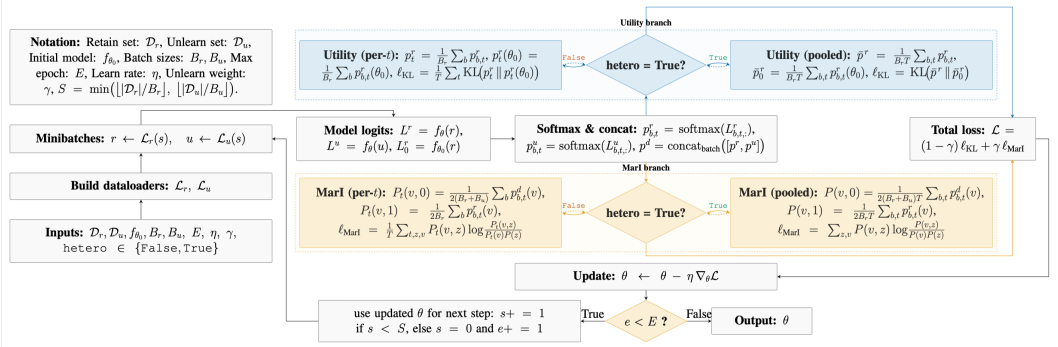


Figure 8: Flow chart for Forgetting-MarI.

Algorithm 1: Forgetting-MarI.

Algorithm 2: Forgetting-MarI

Require: Retain dataset $\mathcal{D}_r = \{x_i^r\}_{i=1}^{|\mathcal{D}_r|}$, unlearn dataset $\mathcal{D}_u = \{x_j^u\}_{j=1}^{|\mathcal{D}_u|}$; initial model f_{θ_0}; learning rate η; batch sizes B_r, B_u; epochs E; trade-off $\gamma \in (0,1)$; hetero $\in \{\texttt{False}, \texttt{True}\}$.

Ensure: Unlearned parameters θ_E.

Initialize $\theta \leftarrow \theta_0$; build dataloaders $\mathcal{L}_r, \mathcal{L}_u$.
for $e = 1$ **to** E **do**
  Shuffle $\mathcal{L}_r, \mathcal{L}_u$; $S \leftarrow \min\{|\mathcal{L}_r|, |\mathcal{L}_u|\}$.
  for $s = 1$ **to** S **do**
    Minibatches: $r \leftarrow \mathcal{L}_r(s), \quad u \leftarrow \mathcal{L}_u(s)$.
    Model logits & probabilities:
      $L^r \leftarrow f_\theta(r) \in \mathbb{R}^{B_r \times T \times |V|}$, $p_{b,t}^r \leftarrow \mathrm{softmax}(L^r[b,t,:]) \in [0,1]^{B_r \times T \times |V|}$;
      $L_0^r \leftarrow f_{\theta_0}(r) \in \mathbb{R}^{B_r \times T \times |V|}$, $p_{b,t}^r(\theta_0) \leftarrow \mathrm{softmax}(L_0^r[b,t,:]) \in [0,1]^{B_r \times T \times |V|}$;
      $L^u \leftarrow f_\theta(u) \in \mathbb{R}^{B_u \times T \times |V|}$, $p_{b,t}^u \leftarrow \mathrm{softmax}(L^u[b,t,:]) \in [0,1]^{B_u \times T \times |V|}$;
      $p^d \leftarrow \mathrm{concat}_{\mathrm{batch}}([p^r, p^u]) \in [0,1]^{(B_r+B_u) \times T \times |V|}$.
    Utility Loss:
      If hetero=**True**:
        $\bar{p}^r \leftarrow \frac{1}{B_r * T} \sum_{b,t} p_{b,t}^r$, $\bar{p}_0^r \leftarrow \frac{1}{B_r * T} \sum_{b,t} p_{b,t}^r(\theta_0)$,
        $\ell_{\mathrm{KL}} \leftarrow \sum_{v \in V} \bar{p}^r(v) \log(\frac{\bar{p}^r(v)}{\bar{p}_0^r(v)})$
      Else:
        $p_t^r \leftarrow \frac{1}{B_r} \sum_{b=1}^{B_r} p_{b,t}^r$, $p_t^r(\theta_0) \leftarrow \frac{1}{N_r} \sum_{b=1}^{B_r} p_{b,t}^r(\theta_0)$,
        $\ell_{\mathrm{KL}} \leftarrow \frac{1}{T} \sum_t \left(\sum_{v \in V} p_t^r(v) \log(\frac{p_t^r(v)}{p_t^r(\theta_0)(v)}) \right)$
    MarI Loss:
      If hetero=**True**:
        $P(v,0) \leftarrow \frac{1}{2(B_r+B_u)*T} \sum_{b,t} p_{b,t}^d$, $P(v,1) \leftarrow \frac{1}{2B_r*T} \sum_{b,t} p_{b,t}^r$,
        $P(v) = \sum_{z \in \{0,1\}} P(v,z)$,
        $\ell_{\mathrm{MarI}} \leftarrow \sum_{z \in \{0,1\}} \sum_{v \in V} P(v,z) \log \frac{P(v,z)}{P(v) P(z)}$.
      Else:
        $P_t(v,0) \leftarrow \frac{1}{2(B_r+B_u)} \sum_b p_{b,t}^d$, $P_t(v,1) \leftarrow \frac{1}{2B_r} \sum_b p_{b,t}^r$, $P_t(v) = \sum_z P_t(v,z)$,
        $\ell_{\mathrm{MarI}} \leftarrow \frac{1}{T} \sum_t \left(\sum_{z \in \{0,1\}} \sum_{v \in V} P_t(v,z) \log \frac{P_t(v,z)}{P_t(v) P_t(z)} \right)$.
    Total loss: $\mathcal{L} = (1-\gamma)\,\ell_{\mathrm{KL}} + \gamma\,\ell_{\mathrm{MarI}}^{\mathrm{flat}}$.
    Update: $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}$.
  end for
end for
return $\theta_E \leftarrow \theta$

Here, we note that, in practice, encoding often introduces padding tokens, and one should ignore those for downstream calculations: $X_R^{\text{flat}} \leftarrow \text{flatten}\left(L_R[i : x_i \neq \text{pad}]\right) \in \mathbb{R}^{N_r \times V}$; and do similarly for X_U^{flat}. This ensures that probabilities derived from L_R^{flat} are not biased by padding positions.

D.4 Ablation Study for the GPT2-LG

In this ablation study, we deliberately curated the datasets to be maximally correlated, creating conditions where the distinction between forget and retain information is most difficult. This setup serves as a stress test for unlearning methods, as it requires the algorithm to selectively remove knowledge that is deeply intertwined with information that should be preserved.

Different γ Figure 10 reports the training curves of all the compared full-parameter tuning UL methods using different regularization parameters γ. Due to the marginal information unlearning nature, Forgetting-MarI has the advantage that a wide range of parameter choice results in fast convergence around the unlearn baseline, indicating robust parameter-tuning. In comparison, it is clear that other methods demonstrate non-convergent (unstable) learning trajectory with extremely narrow parameter range to get close to the unlearn baseline, indicating extreme difficulty and effort in parameter-tuning.

Learning Rate Figure 9 tests a larger learning rate scenario among all methods.

These results collectively demonstrate that Forgetting-MarI provides more precise control over the unlearning process, maintaining the critical balance between forgetting targeted information and preserving general utility. The results highlights two notions of robustness:

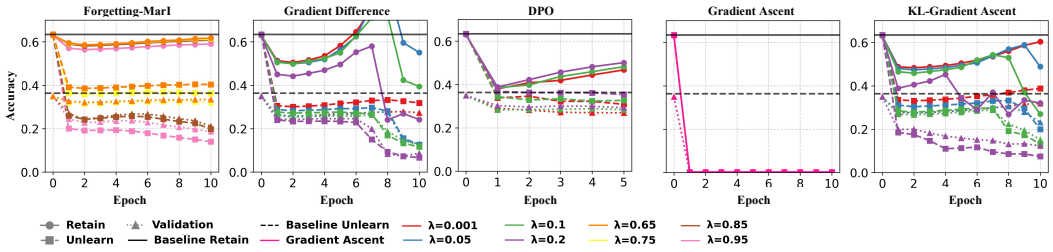


Figure 9: Training curves for each method with varying choices of the regularization parameter γ and lr=1e-4. Forgetting-MarI exhibits smooth monotone behavior, while the other methods show oscillation or utility collapse.

(1) *Training robustness over epochs.* Forgetting-MarI descends steadily to its optimum as seen in Fig. 9. GA and GD overshoot and bounce, KL–GA diverges after 5–6 epochs, DPO plateaus prematurely.

(2) *Robustness against regularization tuning.* Forgetting-MarI shows a monotone and smooth utility-unlearning trade-off when adjusting the regularization parameter. In comparison, GD and KL-GA display unstable oscillations with different choices of γ.

Both training and regularization robustness are necessary for a practical use of unlearning techniques. Practitioners do not have access to ground truth baselines and have limited time to select/determine the best parameter or training epoch to stop at, so stability is essential. Forgetting-MarI is the most stable technique during unlearning, making it the safest choice in practice.

D.5 8B model performance

Here, we scaled our experiments to the Llama-3-8B model to evaluate method robustness on larger architectures. As shown in Figure 11, **Forgetting-MarI** (denoted as `muitr`) successfully bridges the gap between the *Finetune Baseline* and the *Retain Baseline*. Specifically, it reduces unlearn accuracy (orange) from the fine-tuned level ($\sim$0.40) to a level comparable to the gold standard ($\sim$0.33–0.35) while preserving high retain accuracy (blue). In contrast, full-information baselines exhibit distinct failure modes: **DPO** and **Gradient Difference** (GradDiff) suffer from "under-unlearning," where the unlearn accuracy remains stubbornly high ($\sim$0.39–0.40), barely shifting from the fine-tuned state. Furthermore, these methods appear to overfit the retain set, driving retain accuracy

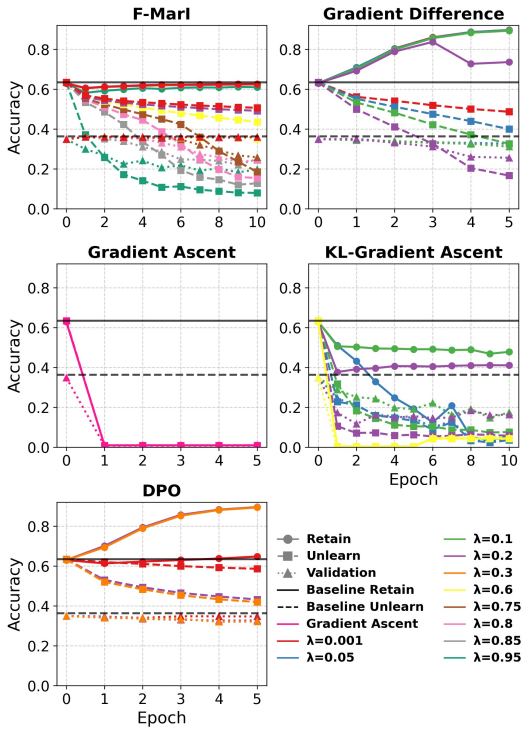


Figure 10: Training Curves for full-parameter UL methods with different γ choices with lr$=$ 1e-5.

abnormally high (> 0.43) compared to the baseline. Finally, **Gradient Ascent** results in total utility collapse, dropping all accuracies to near zero. These findings confirm that Forgetting-MarI's targeted information removal scales effectively to larger parameters where heuristic methods struggle to disentangle retain and unlearn signals.

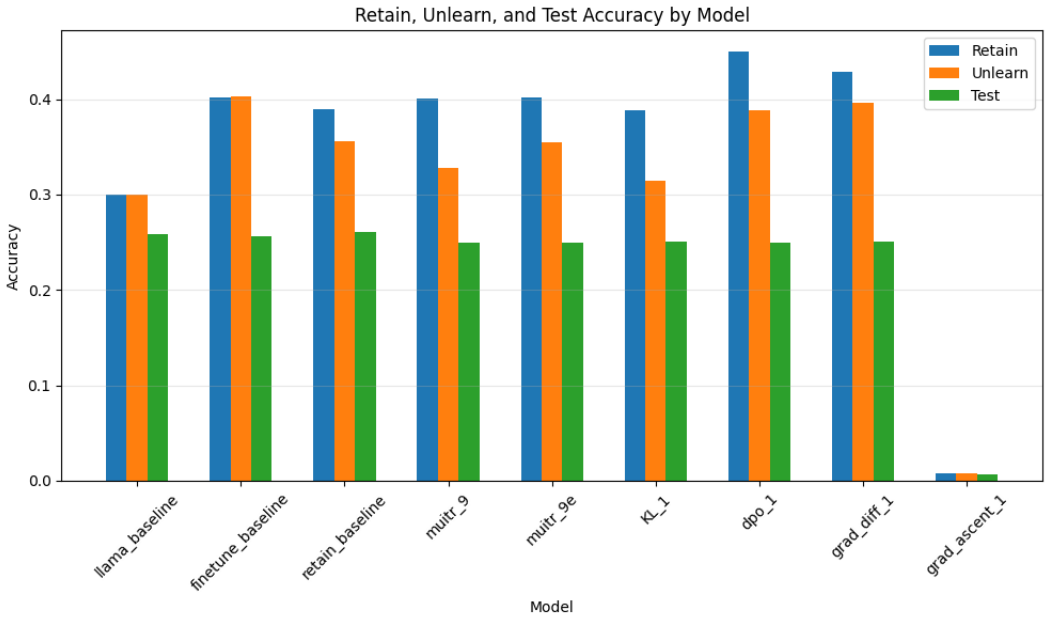
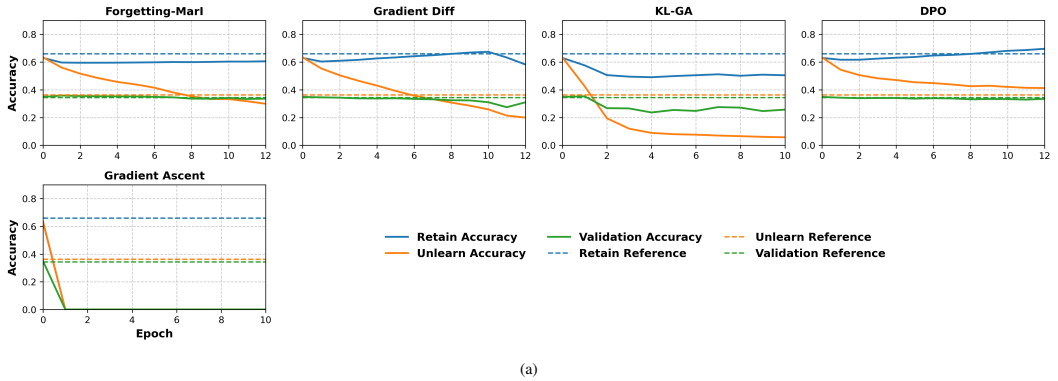


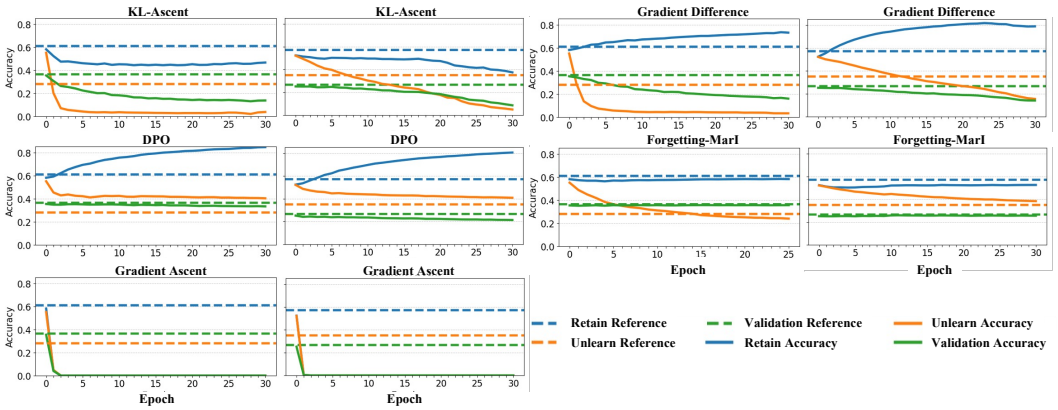
Figure 11: Llama-3-8B Unlearning Performance. We compare the token-level accuracy of Forgetting-MarI (denoted by `muitr`) against baselines. The **Retain Baseline** (Gold Standard) represents the ideal target state. Forgetting-MarI closely approximates this baseline, whereas methods like DPO and Gradient Difference fail to reduce unlearn accuracy effectively, and Gradient Ascent causes catastrophic model collapse.

D.6 SUPPLEMENTAL TO SEC 4.2

Stability of continual unlearning Figure 12 below shows the unlearning trajectories of different methods on GPT-2 with Harry Porter and Llama 3.2-1B with Careless People. The unlearning performance of each method over the course of unlearning, where the curves for each method correspond to the experiment with the best performing regularization parameter for each method.



(a)



(b)

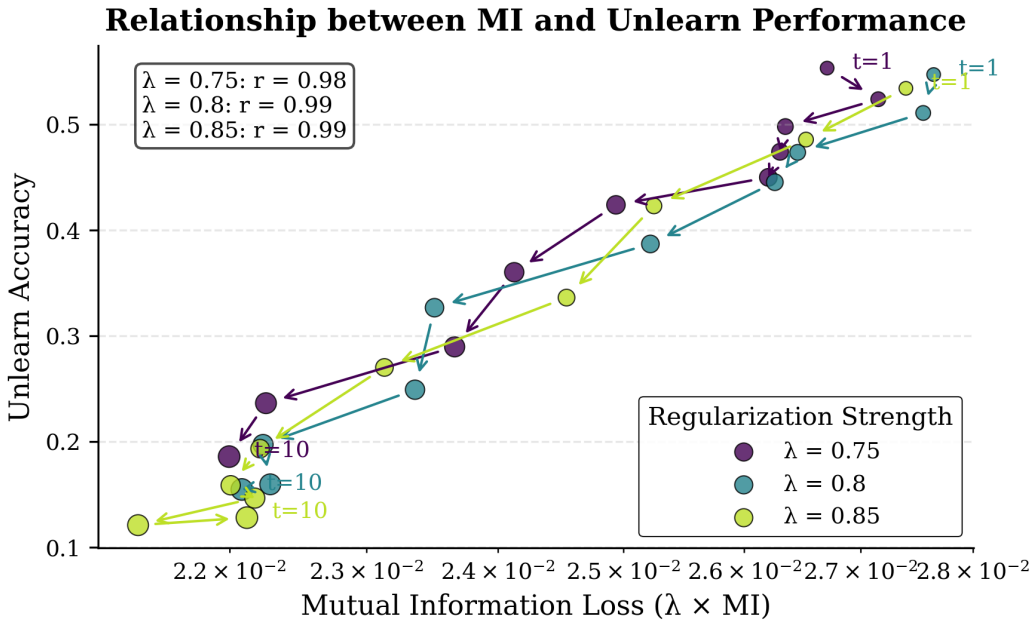
Figure 12: Next-token prediction accuracy during unlearning training across different methods. Horizontal dashed lines represent the "gold standard" unlearn baseline (model trained only on the retain dataset $\mathcal{D}_r$). **(a) Top:** Results for GPT-2 Large on Harry Potter, showing training dynamics across epochs (epoch 0 represents pre-unlearning performance). **(b) Bottom:** Results for Llama-3.2-1B on Careless People, with left and right columns for each method showing correlated and uncorrelated test settings, respectively.

Forgetting-MarI is the best at smoothly approximating the unlearn baseline. The methods based on gradient ascent, GA, GD, and KL-GA, all over-penalize $\mathcal{D}_u$ due to the utility-destroying nature of gradient ascent. DPO, meanwhile, never matches the unlearn baseline in accuracy on $\mathcal{D}_u$ and overtrains on $\mathcal{D}_r$. Across both sets of experiments, Forgetting-MarI minimally affects the validation accuracy, as seen by the validation curve remaining largely unchanged.

We note that there is the possibility that, in theory, one could find a perfect balance between gradient ascent and utility regularization, leading to a stable balance between unlearning and utility preservation, using one of the other methods. However, such a balance seems practically unattainable due to the unlearning instability over time and the lack of monotonicity in the choice of γ for methods based on gradient ascent.

Additionally, we report the training curves for various γ for F-MarI for the HP in Figure 14. We also plot the proposed MI loss across a few γ.

Finally, in Figure 13, we report an observation of nearly perfect correlation between the designed marginal information regularization loss and the unlearn accuracy, under different choice of regularization parameter.



Arrows indicate the progression during training.

Figure 13: MI Loss for GPT-2 Large on Harry Potter Dataset. Here, we observe nearly perfect correlation between the designed marginal information regularization loss and the unlearn accuracy, under different choice of regularization parameter.)

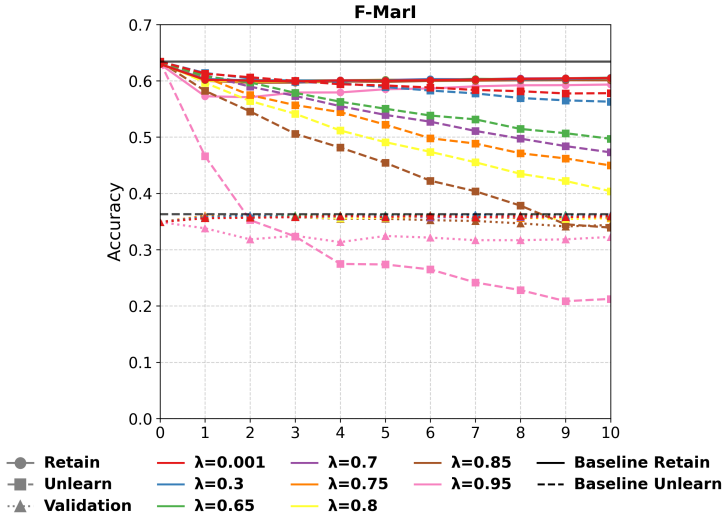


Figure 14: Training curves of F-MarI for the 10/90 split unlearning with the GPT2-LG model.

D.7 Supplementary General Model Capacity Test Results

Table 5 and 6 summarize comprehensive evaluation results across multiple benchmark tests.

Task	Metric	GPT2-LG	Baseline	Unlearn Baseline	F-MarI	KL-GA	GA	GD	DPO
ARC-Easy	acc	0.53 ± 0.01	0.46 ± 0.01	0.48 ± 0.01	0.46 ± 0.01	0.46 ± 0.01	0.24 ± 0.01	0.45 ± 0.01	0.46 ± 0.01
	acc_norm	0.47 ± 0.01	0.42 ± 0.01	0.43 ± 0.01	0.43 ± 0.01	0.43 ± 0.01	0.25 ± 0.01	0.43 ± 0.01	0.43 ± 0.01
ARC-Challenge	acc	0.22 ± 0.01	0.23 ± 0.01	0.22 ± 0.01	0.23 ± 0.01	0.24 ± 0.01	0.22 ± 0.01	0.23 ± 0.01	0.23 ± 0.01
	acc_norm	0.25 ± 0.01	0.27 ± 0.01	0.27 ± 0.01	0.27 ± 0.01	0.28 ± 0.01	0.26 ± 0.01	0.27 ± 0.01	0.27 ± 0.01
PIQA	acc	0.70 ± 0.01	0.66 ± 0.01	0.66 ± 0.01	0.66 ± 0.01	0.66 ± 0.01	0.53 ± 0.01	0.65 ± 0.01	0.65 ± 0.01
	acc_norm	0.69 ± 0.01	0.65 ± 0.01	0.65 ± 0.01	0.66 ± 0.01	0.66 ± 0.01	0.51 ± 0.01	0.64 ± 0.01	0.65 ± 0.01
Hellaswag	acc	0.36 ± 0.00	0.36 ± 0.00	0.36 ± 0.00	0.35 ± 0.00	0.36 ± 0.00	0.25 ± 0.00	0.35 ± 0.00	0.36 ± 0.00
	acc_norm	0.45 ± 0.00	0.43 ± 0.00	0.43 ± 0.00	0.42 ± 0.00	0.42 ± 0.00	0.26 ± 0.00	0.42 ± 0.00	0.42 ± 0.00
MMLU	acc	0.23 ±0.00	0.23 ±0.00	0.24 ±0.00	0.23 ±0.00	0.23 ±0.00	0.25 ±0.00	0.23 ±0.00	0.23 ±0.00
- humanities	acc	0.25 ±0.01	0.24 ±0.01	0.25 ±0.01	0.24 ±0.01	0.24 ±0.01	0.25 ±0.01	0.24 ±0.01	0.25 ±0.01
- other	acc	0.24 ±0.01	0.24 ±0.01	0.24 ±0.01	0.24 ±0.01	0.24 ±0.01	0.27 ±0.01	0.25 ±0.01	0.24 ±0.01
- social sciences	acc	0.22 ±0.01	0.22 ±0.01	0.22 ±0.01	0.22 ±0.01	0.22 ±0.01	0.23 ±0.01	0.22 ±0.01	0.22 ±0.01
- stem	acc	0.22 ±0.01	0.22 ±0.01	0.23 ±0.01	0.22 ±0.01	0.21 ±0.01	0.24 ±0.01	0.22 ±0.01	0.22 ±0.01

Task	Metric	GPT2-LG	Baseline	Unlearn Baseline	F-MarI	KL-GA	GA	GD	DPO
WikiText	bits/byte	0.841	0.925	0.905	0.917	0.929	26.895	0.961	0.939
	byte-pplx	1.792	1.898	1.873	1.888	1.904	1.248e+08	1.946	1.917
	word-pplx	22.612	30.797	28.662	29.886	31.309	1.972e+43	35.214	32.428

Table 5: Comprehensive evaluation results across multiple benchmarks for the GPT2-LG baselines. WikiText is based on perplexity, so lower is better. The higher score is better for all other tests: ARC, PIQA, HellaSwag, and MMLU.

Task	Metric	LLaMA-3.2-1B	Full Finetune	Unlearn Baseline	F-MarI	KL-GA	GA	GD	DPO
ARC-Easy	acc	0.65 ± 0.01	0.59 ± 0.01	0.60 ± 0.01	0.58 ± 0.01	0.55 ± 0.01	0.42 ± 0.01	0.57 ± 0.01	0.59 ± 0.01
	acc_norm	0.61 ± 0.01	0.56 ± 0.01	0.55 ± 0.01	0.56 ± 0.01	0.53 ± 0.01	0.39 ± 0.01	0.55 ± 0.01	0.56 ± 0.01
ARC-Challenge	acc	0.31 ± 0.01	0.30 ± 0.01	0.31 ± 0.01	0.29 ± 0.01	0.30 ± 0.01	0.27 ± 0.01	0.30 ± 0.01	0.30 ± 0.01
	acc_norm	0.36 ± 0.01	0.35 ± 0.01	0.36 ± 0.01	0.34 ± 0.01	0.33 ± 0.01	0.29 ± 0.01	0.32 ± 0.01	0.32 ± 0.01
PIQA	acc	0.74 ± 0.01	0.71 ± 0.01	0.71 ± 0.01	0.71 ± 0.01	0.70 ± 0.01	0.63 ± 0.01	0.69 ± 0.01	0.71 ± 0.01
	acc_norm	0.74 ± 0.01	0.71 ± 0.01	0.71 ± 0.01	0.69 ± 0.01	0.69 ± 0.01	0.62 ± 0.01	0.70 ± 0.01	0.72 ± 0.01
Hellaswag	acc	0.48 ± 0.00	0.48 ± 0.00	0.50 ± 0.00	0.47 ± 0.00	0.46 ± 0.00	0.30 ± 0.00	0.45 ± 0.00	0.47 ± 0.00
	acc_norm	0.64 ± 0.00	0.63 ± 0.00	0.65 ± 0.00	0.62 ± 0.00	0.60 ± 0.00	0.38 ± 0.00	0.59 ± 0.00	0.61 ± 0.00
MMLU	acc	0.37 ± 0.00	0.30 ± 0.00	0.31 ± 0.00	0.28 ± 0.00	0.26 ± 0.00	0.23 ± 0.00	0.26 ± 0.00	0.28 ± 0.00
- humanities	acc	0.35 ± 0.01	0.30 ± 0.01	0.31 ± 0.01	0.29 ± 0.01	0.26 ± 0.01	0.24 ± 0.01	0.27 ± 0.01	0.29 ± 0.01
- other	acc	0.41 ± 0.01	0.31 ± 0.01	0.32 ± 0.01	0.29 ± 0.01	0.28 ± 0.01	0.24 ± 0.01	0.25 ± 0.01	0.30 ± 0.01
- social sciences	acc	0.39 ± 0.01	0.32 ± 0.01	0.32 ± 0.01	0.28 ± 0.01	0.26 ± 0.01	0.22 ± 0.01	0.24 ± 0.01	0.26 ± 0.01
- stem	acc	0.33 ± 0.01	0.26 ± 0.01	0.28 ± 0.01	0.28 ± 0.01	0.25 ± 0.01	0.21 ± 0.01	0.27 ± 0.01	0.26 ± 0.01

Table 6: Comprehensive evaluation results across multiple benchmarks for the LLaMA-3.2-1B unlearning experiment.

E Appendix of Section 4.4: Detection Tests

E.1 Detector Methods

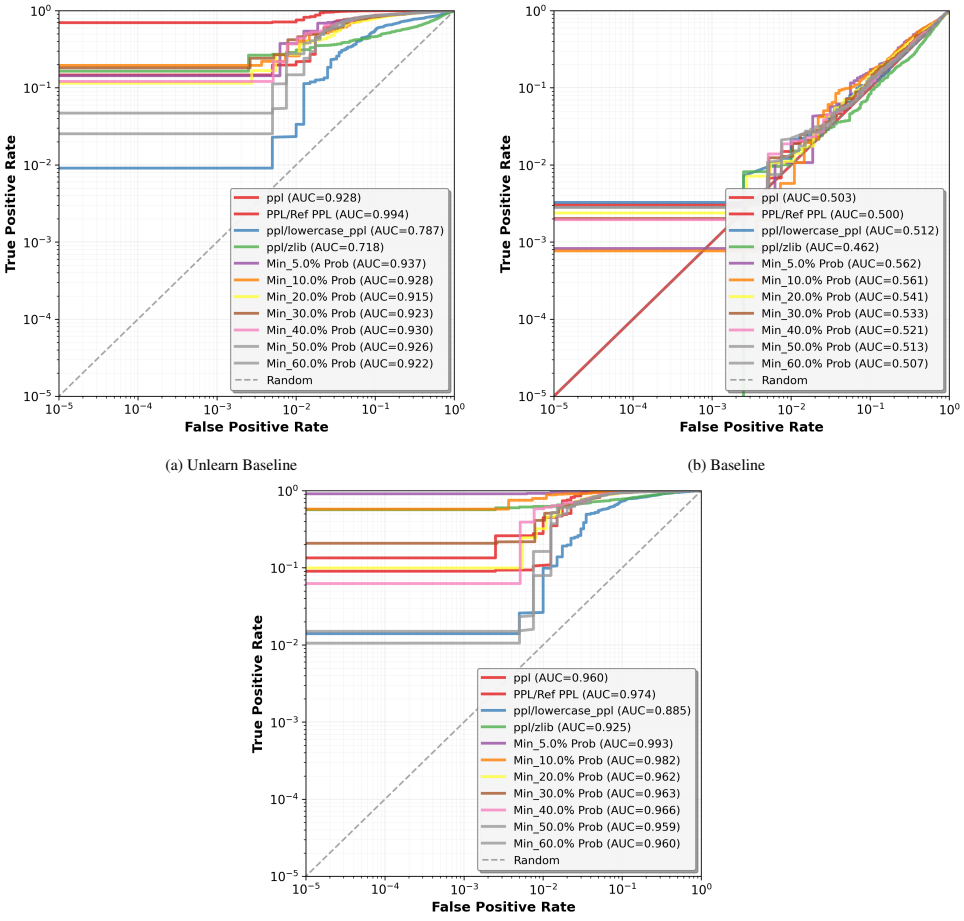
Here, we provide a more detailed introduction to the current study of copyright content detectors for LLMs so that readers better understand the numerical study in section 4.4. The current study of copyrighted text detectors can be roughly separated into two lines of work:

- **White-box methods:** Perplexity outlier and reference model perplexity outlier [3], domain normalized minimum k-percentage [42], and data-set level inference [31].
  The above methods largely share the same idea of constructing a statistic (or a vector of statistics) that indicates the probability that a model has seen a given sentence or not. It bases the probability on how confidently the model predicts the true output. The idea is based on the intuition that a model that has seen the sentence during training will have high confidence when trying to complete it.
- **Black-box methods:** Direct regurgitation probes [22], Name-cloze membership inference [9], DE-COP: multi-choice preference [5], and Output-consistency measures [16].
  Black-box methods, which do not have access to the model parameters and therefore the output logits or prediction distributions, often use either edit distance (a.k.a. Levenshtein distance) or some token embedding model (e.g. a small transformer) to quantify the distance or similarity between a model's output and a reference string, then generate statistics of the similarity between the two.

Black-box methods are weaker detectors than white-box methods since they do not have access to a model's internals. Since our method assumes access to the model parameter, we tested our method against the current SotA white-box method, the minimum k-percent method [36], to demonstrate the effectiveness of our unlearning in real-world applications.

E.2 Multiple Detection Test Results

Here, we provide more ablation details on the undetectability result provided in Figure 5, Section 4.4.



(a) Unlearn Baseline

(b) Baseline

(c) Forgetting-MarI

Figure 15: Training data membership detection test of Forgetting-MarI against state-of-the-art detection methods using the 10/90 split Unlearning of the GPT2-LG.