AGTTU: Attention-Guided Target Token Unlearning for Large Language Models

Anonymous authors
Paper under double-blind review

Abstract

Large language models exhibit strong language understanding and reasoning capabilities, but also memorize private or copyrighted content, raising privacy and compliance concerns. Existing unlearning methods often uniformly suppress entire forget samples, inadvertently degrading general language capabilities unrelated to the unlearning targets and causing excessive forgetting. Token-level approaches alleviate this issue, but some rely on auxiliary models or additional perturbation-based scoring. We propose Attention-Guided Target Token Unlearning(AGTTU), which selects target tokens based on the attention they receive from subsequent answer positions, using the model's own attention allocation to guide targeted unlearning. We further introduce Contrastive Logits Preference Loss (CLPL) to suppress target tokens while limiting damage to non-target content, together with a retain-set objective to preserve model utility. Experiments on TOFU and MUSE Books demonstrate that AGTTU improves the trade-off between unlearning effectiveness and model utility. On TOFU with Llama-3.2-1B, AGTTU achieves higher reported forget quality and model utility than the listed baselines across all three forgetting ratios. On MUSE Books, compared with TPO, AGTTU reduces the verbatim memorization score on the forget set by 4.51% while increasing the knowledge memorization score on the retain set by 15.17%.

1 Introduction

Large language models (LLMs) acquire broad linguistic capabilities but may also memorize private or copyrighted content, as demonstrated by training-data extraction attacks (Carlini et al., 2021). Deleting source data does not remove its influence on model parameters. Machine unlearning aims to remove this influence and approximate the behavior of a model trained without the designated data (Cao & Yang, 2015; Bourtoule et al., 2021). For LLMs, retraining is expensive, and unwanted facts share parameters with reusable linguistic knowledge, so deleting particular data can affect other tasks. Practical unlearning must therefore suppress designated content while preserving retained knowledge and general language capabilities.

Existing methods often optimize entire samples or answers. GradDiff increases forget-set prediction loss while training on retain data to limit degradation of retained knowledge and language modeling; NPO uses a negative preference objective to mitigate collapse under direct gradient ascent (Maini et al., 2024; Zhang et al., 2024). However, an answer mixes unwanted facts with reusable expression, so uniform suppression can cause excessive forgetting. Token-level methods narrow the optimization scope but still require target identification: TPO relies on auxiliary identifiers, whereas DTO measures prediction changes under input perturbations (Zhou et al., 2026; Lee et al., 2026). These strategies introduce additional model dependencies or scoring costs, motivating the use of internal information already available when the model processes an answer.

Attention offers a way to examine how models use preceding answer content. During autoregressive generation, subsequent positions access earlier information through attention; the attention an answer token receives can therefore indicate its involvement in subsequent generation. We investigate this connection through masking with Llama-3.2-1B on the 10% forget set of the TOFU benchmark (Maini et al., 2024): zeroing attention columns for high-attention answer tokens reduces the probability of the complete original answer more than masking equally many random tokens (Appendix F). This observation provides empirical motivation for attention-based target selection.

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

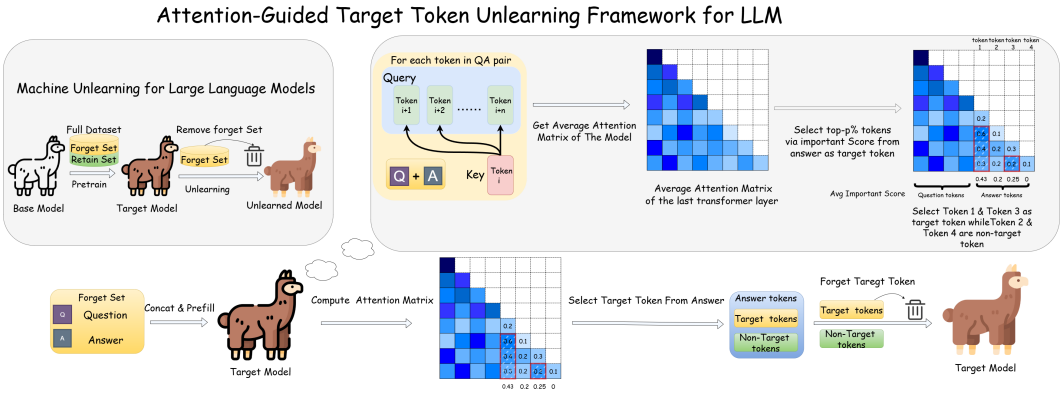


Figure 1: Overview of AGTTU: Attention-Guided Target Token Unlearning.

Inspired by this finding, we propose AGTTU (Attention-Guided Target Token Unlearning), shown in Figure 1. AGTTU prefills concatenated questions and ground-truth answers, reads head-averaged last-layer attention, and selects answer tokens by the attention they receive from subsequent positions. To suppress these targets while limiting damage to general expression, we extend TPO's Logit Preference Loss (LPL) into Contrastive Logits Preference Loss (CLPL). A margin on average target-logit reduction strengthens suppression, while a one-sided penalty discourages non-target ground-truth logits from falling below their reference values. Target reduction and non-target damage enter a shared preference score, complemented by retain-set cross-entropy, linking target identification within answers to targeted optimization.

Our main contributions are:

1. **Attention-guided target token identification.** We rank answer tokens by attention received from subsequent positions, converting the model's own information-use patterns into unlearning targets without auxiliary identifiers or tokenwise perturbation scoring.
2. **we design Contrastive Logits Preference Loss (CLPL)**, introducing a margin on average target-logit reduction and a one-sided penalty on non-target logit decreases to strengthen target suppression while limiting damage to non-target content.
3. **Benchmark and ablation evaluation.** On TOFU with Llama-3.2-1B, AGTTU achieves higher FQ and MU than the listed baselines across all three forgetting ratios. On MUSE Books, it achieves lower verbatim memorization and higher retain-set utility than TPO_{GDR}. Selection and component ablations examine target identification and the CLPL components, respectively.

2 Related Work

2.1 Sample-Level Unlearning Objectives

Sample-level methods modify the objective on complete forget samples to weaken memorization. Gradient ascent directly increases prediction loss on forget data (Jang et al., 2023; Yao et al., 2024); GradDiff additionally performs gradient descent on retain data to limit interference with other capabilities (Maini et al., 2024). NPO formulates unlearning as negative preference optimization relative to the original model, mitigating collapse under direct gradient ascent (Zhang et al., 2024). Applying these objectives to complete answers does not distinguish unwanted facts from reusable expression, potentially damaging unrelated content along with target memories.

Other studies modify training signals to improve the forgetting–utility trade-off. ASU constructs a forget teacher by smoothing attention distributions, then updates the model through self-distillation (Zade et al., 2026). JensUn uses Jensen–Shannon divergence for forget- and retain-set objectives and evaluates persistence under prompt paraphrasing and benign relearning (Singh et al., 2025). These designs improve sample-level updates without directly identifying which positions

within an answer most need unlearning. AGTTU identifies targets before optimization and applies suppression and protection to target and non-target positions, respectively.

2.2 Selective Token and Target Token Unlearning

Selective unlearning restricts optimization to tokens or spans to reduce interference with unrelated content (Wang et al., 2025; Yoon et al., 2026). Methods differ primarily in how they identify targets. SU compares predictions from two assistants with different knowledge scopes to identify forget-related tokens, requiring additional assistant models (Wan et al., 2025). TPO uses auxiliary models to distinguish unwanted from general words, optimizing them with LPL and Preservation Loss (PL), respectively (Zhou et al., 2026). Identification depends on the auxiliary model's judgment; using a third-party service also entails sharing forget text externally, introducing additional privacy risks.

Methods without auxiliary identifiers use alternative scoring signals. DTO identifies targets from prediction changes under input perturbations, requiring additional forward computation (Lee et al., 2026). ETW weights token losses by predictive-distribution entropy, reflecting uncertainty at the current prediction position (Koh et al., 2026). AGTTU instead measures how strongly subsequent answer positions attend to a token, extracting this signal from the prefill attention matrix without tokenwise perturbations. CLPL retains LPL's target-logit suppression and adds a target-reduction margin and a one-sided non-target penalty to limit interference with other answer content.

3 AGTTU

Overview. AGTTU comprises target identification and targeted optimization (Figure 1). We first prefill the frozen original model with the question and ground-truth answer, read last-layer attention, average across heads, and extract the answer submatrix. Tokens are ranked by attention received from subsequent positions; target masks and reference logits are cached. With complete answer prefixes preserved, CLPL suppresses targets and penalizes non-target decreases, while retain-set cross-entropy supports general capabilities. Only the trainable model is updated; the original model remains frozen.

3.1 Unlearning Task and Model Definitions

Let $\mathcal{D} = \mathcal{D}_f \cup \mathcal{D}_r$ be the original training set, partitioned into disjoint forget and retain sets, $\mathcal{D}_f \cap \mathcal{D}_r = \varnothing$. Unlearning removes the influence of designated data while preserving knowledge and capabilities supported by the remainder. We distinguish the following model roles and training states:

Original model θ_0. θ_0 denotes the parameters of the model trained on the complete training set $\mathcal{D}$, including both forget and retain data. It remains frozen in AGTTU and is used for target selection and reference-logit computation.

Trainable model θ. We initialize the trainable parameters θ from θ_0 and update them using the unlearning objective together with the retain-set loss.

Unlearned model θ_{u}. We denote the final parameters obtained after unlearning training by θ_{u} and refer to the resulting model as the *unlearned model*.

Gold-standard model θ_{gold}. θ_{gold} is trained exclusively on $\mathcal{D}_r$ from the same starting point and with the same training protocol, without exposure to $\mathcal{D}_f$ during task-specific training. It is not obtained by further fine-tuning θ_0, which has already seen $\mathcal{D}_f$. The model labeled Retain in the experiments is this evaluation reference; it does not participate in AGTTU's target identification or updates.

Let $P_\phi(\cdot \mid x)$ denote the conditional answer distribution of a model with parameters ϕ given input x. The goal is for the final model to approximate the gold-standard model trained only on the retain set:

$$P_{\theta_{\mathrm{u}}}(\cdot \mid x) \approx P_{\theta_{\mathrm{gold}}}(\cdot \mid x).$$

This approximation is desired for both forget-related inputs and retained tasks. Unlearning aims to remove the influence of the forget data while preserving the knowledge and capabilities supported by the retain data.

3.2 Attention-Based Target Token Selection

Inputs and attention matrices. For a forget sample (x, y), x denotes the question and prompt context, and $y = (y_1, \ldots, y_n)$ the ground-truth answer with n valid tokens. Token y_i is predicted from $c_i = (x, y_{<i})$, where $y_{<i} = (y_1, \ldots, y_{i-1})$. We concatenate x, y using the chat template, prefill the frozen θ_0, and read last-layer attention. Prefill processes the complete known sequence, including the sample's ground-truth answer. If the preceding question and template prefix have length q, the valid sequence length is $L = q + n$, and token y_i occupies position $q + i$ under one-based indexing.

The last layer has H query heads, indexed by $h \in \{1, \ldots, H\}$. Matrices $Q^{(h)}, K^{(h)} \in \mathbb{R}^{L \times d_h}$ contain head h's queries and corresponding keys, with dimension d_h and the model's actual positional encoding. Attention and its head average are

$$A^{(h)} = \mathrm{softmax}_{\mathrm{row}}\left(\frac{Q^{(h)} K^{(h)\top}}{\sqrt{d_h}} + M\right), \qquad \bar{A} = \frac{1}{H}\sum_{h=1}^{H} A^{(h)}. \tag{1}$$

Here, $\top$ denotes transposition, $\sqrt{d_h}$ the scaling factor, and $\mathrm{softmax}_{\mathrm{row}}$ rowwise softmax. The causal mask $M \in \mathbb{R}^{L \times L}$ has $M_{uv} = 0$ for $v \le u$ and $-\infty$ otherwise, where u, v index complete-sequence query and key positions. $A^{(h)} \in \mathbb{R}^{L \times L}$ contains head h's attention weights, and $\bar{A}$ is their arithmetic mean across heads. Rows index queries; columns index attended keys.

Token scores and target sets. Scoring uses only the answer-to-answer submatrix $\bar{A}^{\mathrm{ans}} \in \mathbb{R}^{n \times n}$, where $\bar{A}^{\mathrm{ans}}_{j,i} = \bar{A}_{q+j,q+i}$ for $i, j \in \{1, \ldots, n\}$. Column i records attention received by y_i from other answer positions. Averaging over subsequent queries $j > i$ gives

$$s_i = \frac{1}{n-i}\sum_{j=i+1}^{n} \bar{A}_{q+j,q+i} \quad (1 \le i < n), \qquad s_n = 0. \tag{2}$$

Here, j ranges over answer positions after i; $\bar{A}_{q+j,q+i}$ is the head-averaged attention from the query position of subsequent token y_j to the key position of y_i; and $n - i$ is the number of subsequent positions used in scoring. Normalizing by this count reduces the effect of differences in visible suffix length. The final answer token has no subsequent answer positions, so we set $s_n = 0$ by convention; this does not imply that it contains no knowledge to be forgotten.

Let $s = (s_1, \ldots, s_n)$ be the score vector, $p \in (0, 1)$ the prescribed target-token fraction, and k the actual number selected. For $n \ge 2$, we set $k = \min\{n - 1, \max\{1, \lceil pn \rceil\}\}$, where $\lceil \cdot \rceil$ denotes the ceiling function. The target set is $T = \mathrm{TopK}(s, k)$, containing the indices of the k highest-scoring answer positions; the non-target set is $N = \{1, \ldots, n\} \setminus T$. Their cardinalities are $|T|$ and $|N|$. Special and padding tokens are excluded from candidate scoring. For a single-token answer, we use $T = \{1\}$ and $N = \varnothing$.

Target selection is performed offline on the complete ground-truth answer. Training preserves the original teacher-forced prefix and changes only the loss term to which each position contributes. We use post-softmax attention weights, not unnormalized $QK^\top$ scores. The scoring and masking stages in Figure 1 correspond to column averaging and TopK selection; prediction-position alignment, boundary handling, and caching conventions are provided in Appendix B.

From selected tokens to prediction positions. The attention key at position $q + i$ identifies token y_i, whereas its ground-truth logit is read from the output at $q + i - 1$. We therefore align logits and labels by one position before applying the target mask. The selected token remains in the teacher-forced prefix when predicting subsequent tokens; selection changes the loss assignment, not the input sequence. This separates the masking intervention used to study attention from the parameter updates used for unlearning.

3.3 CLPL and the Joint Optimization Objective

Target selection identifies positions to suppress, but shared-parameter updates may also damage general expression. CLPL therefore extends LPL with a single preference score balancing target reduction and non-target damage, complemented by cross-entropy on separate retain samples to support utility.

CLPL extends TPO's LPL (Zhou et al., 2026) with a target-reduction margin and a non-target penalty; Appendix D gives the original objective. Let $z_i^\theta = z_\theta(y_i \mid c_i)$ be the ground-truth token's pre-softmax logit and $z_i^{\theta_0}$ its frozen reference value under the same context. Here, $\sigma(a) = 1/(1+e^{-a})$ is sigmoid, $\beta > 0$ controls its transition scale, and $\mathbb{E}$ denotes expectation over samples.

Target reduction and non-target protection. We define the ground-truth token logit decrease relative to the original model and its two aggregates as

$$d_i = z_i^{\theta_0} - z_i^\theta, \qquad D_T = \frac{1}{|T|}\sum_{i\in T} d_i, \qquad C_N = \frac{1}{|N|}\sum_{j\in N}[d_j]_+, \quad [u]_+ = \max(0,u). \tag{3}$$

We set $C_N = 0$ when $N = \varnothing$. D_T measures the average decrease over target tokens. C_N penalizes only decreases below the reference score at non-target positions. We call the resulting unlearning term Contrastive Logits Preference Loss (CLPL), denoted by $\mathcal{L}_{\text{CLPL}} \equiv \mathcal{L}_f^{\text{AGTTU}}$:

$$\mathcal{L}_f^{\text{AGTTU}}(\theta) = -\mathbb{E}_{(x,y)\sim\mathcal{D}_f}\left[\frac{2}{\beta}\log\sigma(\beta(D_T - m) - \gamma C_N)\right]. \tag{4}$$

The target term encourages lower ground-truth logits through D_T. Averaging reduces the direct effect of target count on the summed scale; it neither ensures that every target logit decreases nor makes optimization identical across selection ratios. A positive margin $m > 0$ shifts the preference transition toward a larger average decrease and, at fixed D_T, C_N and other parameters, increases the target-suppression gradient magnitude. It is not a hard per-token threshold, and reaching it does not make the loss zero.

The non-target term penalizes only ground-truth logits below their reference values. Positions at or above the reference receive no direct reinforcement from this term, although joint training may still raise their scores. Taking positive parts before averaging prevents gains above the reference from offsetting other decreases within C_N. With $\gamma \geq 0$ weighting non-target damage, we place γC_N inside the sigmoid but outside the β scaling to combine both effects in one preference score: at fixed D_T, greater non-target damage lowers the score and increases the loss for $\gamma > 0$. This is an empirical design, with no claim of theoretical superiority over a separate external penalty. A larger D_T can compensate for a larger C_N, allowing a soft trade-off rather than strict protection. Setting $m = \gamma = 0$ recovers LPL under the same target mask. Appendix A discusses the design rationale and its scope.

Retention constraint and joint training. The non-target penalty covers only unselected positions in forget samples; retain cross-entropy provides a utility signal from separate retain samples. We randomly sample $\widetilde{\mathcal{D}_r} \subseteq \mathcal{D}_r$ with $|\widetilde{\mathcal{D}_r}| = |\mathcal{D}_f|$, giving

$$\mathcal{L}_r = -\mathbb{E}_{(x,y)\sim\widetilde{\mathcal{D}_r}}\frac{1}{|y|}\sum_{i=1}^{|y|}\log P_\theta(y_i \mid c_i), \qquad \mathcal{L}_{\text{AGTTU}} = \mathcal{L}_f^{\text{AGTTU}} + \lambda\mathcal{L}_r. \tag{5}$$

Here, $|y|$ is answer length and $\lambda \geq 0$ weights the retain loss. Each term is averaged within samples, then across samples. Finite retain data cannot guarantee preservation of all knowledge or capabilities. Updates cover Transformer blocks; other parameters are frozen. The frozen θ_0 supplies selection scores and reference logits, θ receives joint updates, and θ_{gold} is used only for evaluation. Parameter matching, training settings, and hyperparameters appear in Appendix C.

4 Experiments and Results

4.1 Experimental Setup

Datasets and metrics. TOFU (Maini et al., 2024) contains 4,000 question–answer pairs about 200 fictitious authors. We evaluate Llama-3.2-1B and Llama-3.2-3B at 1%, 5%, and 10% forgetting ratios. Forget Quality (FQ) is the two-sample Kolmogorov–Smirnov p-value comparing forget-set truth-ratio distributions with Retain; Model Utility (MU) aggregates retained- and real-world-knowledge metrics. Higher values are preferred, although a high FQ does not establish distributional equality. MUSE Books (Shi et al., 2025) uses Llama-2-7B and measures forget-set verbatim memorization (VerbMem$_f$), knowledge memorization (KnowMem$_f$), membership inference (PrivLeak),

Table 1: TOFU / Llama-3.2-1B results.

	Forget 1%		Forget 5%		Forget 10%	
Method	FQ ↑	MU ↑	FQ ↑	MU ↑	FQ ↑	MU ↑
Retain	1.0000	0.6300	1.0000	0.6000	1.0000	0.5900
GradDiff	0.5786	0.4188	1.94×10^{-119}	0.4850	1.80×10^{-229}	0.4803
Target GradDiff	0.1650	0.4370	2.05×10^{-112}	0.4635	5.67×10^{-204}	0.4474
NPO	0.7659	0.5504	0.1779	0.4383	0.0158	0.5232
Target NPO	<u>0.9188</u>	0.5174	<u>0.3281</u>	<u>0.5229</u>	0.0004	<u>0.5259</u>
TPO (ChatGPT)†	0.1649	0.5476	0.0402	0.4026	1.065×10^{-5}	0.3592
DTO†	0.7659	<u>0.5529</u>	0.3281	0.4218	<u>0.1215</u>	0.3782
AGTTU	**0.9900**	**0.5567**	**0.5453**	**0.5727**	**0.1546**	**0.5709**

†: quoted from Lee et al. (2026, Tables 5–7); DTO is the full method. Sources: Appendix E.1.

Table 2: TOFU / Llama-3.2-3B results.

	Forget 1%		Forget 5%		Forget 10%	
Method	FQ ↑	MU ↑	FQ ↑	MU ↑	FQ ↑	MU ↑
Retain	1.0000	0.6600	1.0000	0.6600	1.0000	0.6500
GradDiff	0.4046	0.3886	1.94×10^{-119}	0.5578	2.97×10^{-217}	0.5828
Target GradDiff	0.5786	0.4556	2.04×10^{-110}	0.6155	5.67×10^{-205}	0.6025
NPO	<u>0.9188</u>	0.6362	0.0085	0.4847	**0.0652**	0.5317
Target NPO	**0.9900**	0.6112	<u>0.5453</u>	0.6021	<u>0.0003</u>	0.5738
TPO	0.2657	<u>0.6603</u>	4.74×10^{-5}	0.6005	4.58×10^{-7}	0.4538
DTO	0.0068	**0.6608**	2.56×10^{-14}	**0.6667**	4.05×10^{-26}	**0.6709**
AGTTU	<u>0.9188</u>	0.6266	**0.8655**	<u>0.6385</u>	**0.0652**	<u>0.6419</u>

TPO is our final-epoch-10 reproduction. Sources: Appendix E.1.

and retain-set knowledge memorization (KnowMem_r). Lower forget-set memorization, PrivLeak closer to zero, and higher KnowMem_r are preferred. Tables boldface the best results and underline runners-up; Retain is excluded from rankings.

Baselines and training. We compare GradDiff, NPO (Zhang et al., 2024), TPO (Zhou et al., 2026), and DTO (Lee et al., 2026). Target GradDiff and Target NPO apply attention masks to their respective forget losses without non-target protection. Our runs train for 10 epochs and report the final checkpoint. Table notes identify quoted results and variants; Appendix E.1 details their sources. Selection ablations fix CLPL, while component ablations fix AGTTU selection; other settings match the corresponding main runs (Appendix C).

Forgetting ratios and selection budgets. The TOFU forgetting ratio specifies the fraction of authors to forget, whereas p specifies the fraction of answer tokens selected within each forget sample. These are distinct quantities: Llama-3.2-1B uses $p = 0.30$ at 1% forgetting and $p = 0.20$ at 5% and 10%; Llama-3.2-3B uses $p = 0.20$ throughout. At each setting, Top-p, Low-p, and Random select the same number of positions per sample. Thus, their comparison examines which tokens are selected under a matched budget, rather than the effect of selecting more tokens.

4.2 Main Results

4.2.1 TOFU

Joint improvements in forgetting and utility. On Llama-3.2-1B, AGTTU's FQ is 0.9900/0.5453/0.1546 and MU is 0.5567/0.5727/0.5709 across the three ratios (Table 1). FQ exceeds the respective runners-up by 0.0712/0.2172/0.0331. At 10%, both metrics exceed TPO (ChatGPT)'s 1.065×10^{-5}/0.3592, quoted from Lee et al. (2026, Table 7); higher FQ does not come at the cost of lower MU than the listed baselines. **AGTTU leads the listed methods in both FQ and MU across all three 1B settings.**

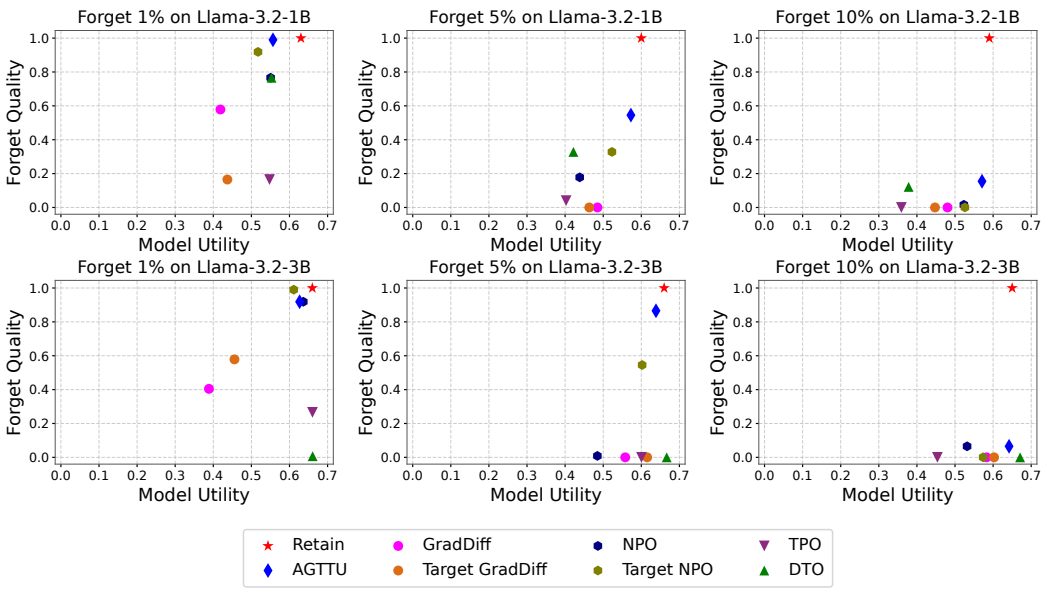


Figure 2: TOFU FQ versus MU: 1B/3B by row and 1%/5%/10% forgetting by column. Each point uses the paired values and sources in Tables 1 and 2. TPO denotes TPO (ChatGPT) in the 1B panels and our reproduction in the 3B panels.

Utility preservation at higher forgetting ratios. Llama-3.2-3B exhibits a different trade-off (Table 2). At 5%, AGTTU's FQ/MU of 0.8655/0.6385 exceeds Target NPO by 0.3203/0.0364 and our TPO reproduction's 4.74×10^{-5}/0.6005. DTO has higher MU but FQ of only 2.56×10^{-14}, indicating a significant difference from Retain in forget-set truth-ratio distributions. At 10%, AGTTU matches NPO's FQ of 0.0652 while increasing MU from 0.5317 to 0.6419; both metrics also exceed TPO's 4.58×10^{-7}/0.4538. **At 3B / 5% and 10%, AGTTU preserves higher utility while matching or improving FQ relative to NPO, Target NPO, and TPO.**

Trade-offs at a low forgetting ratio. At 3B / 1%, AGTTU exceeds TPO in FQ (0.9188 versus 0.2657) but trails in MU (0.6266 versus 0.6603); Target NPO has the highest FQ. Unlike on 1B, no single method leads both metrics. **AGTTU's joint advantage on 1B does not extend to the low-ratio 3B setting; performance depends on the combination of model and forgetting ratio.**

Applying attention selection to NPO. Target NPO has higher FQ and MU than NPO at 5% on both models, but higher MU and lower FQ at 10%. AGTTU exceeds Target NPO in MU in all six settings and in FQ in five. These outcomes support considering selection and optimization jointly, without isolating non-target protection's contribution. **Attention selection's benefits for NPO vary with the forgetting ratio, while AGTTU with CLPL further improves retained utility in the evaluated settings.**

4.2.2 MUSE Books

Verbatim unlearning and retained knowledge. AGTTU obtains VerbMem_f of 0.69 and KnowMem_r of 57.24, versus 19.30 and 57.11 for DTO w/o KL (Lee et al., 2026, Table 2) (Table 3). This comparison concerns the variant without KL regularization, not full DTO. GradDiff has zero forget-set memorization scores but also zero retained utility, illustrating the cost of excessive suppression. **AGTTU substantially reduces verbatim memorization relative to DTO w/o KL while maintaining comparable retain-set knowledge memorization.**

Knowledge unlearning and privacy metrics. Relative to TPO_{GDR} (Zhou et al., 2026, Table 1), AGTTU reduces VerbMem_f by 4.51 and increases KnowMem_r by 15.17. TPO_{GDR}, however, has lower KnowMem_f and PrivLeak closer to zero. GDR denotes retain-set gradient descent. **AGTTU's**

Table 3: MUSE Books results.

	MUSE Books			
	Forget quality			Model utility
Method	VerbMem $\mathcal{D}_f$ (↓)	KnowMem $\mathcal{D}_f$ (↓)	PrivLeak (→ 0)	KnowMem $\mathcal{D}_r$ (↑)
Retain	14.13	30.24	0	68.61
GradDiff	**0**	**0**	<u>−24.19</u>	0
NPO	56.02	35.82	−86.00	51.45
TPO†	5.20	<u>3.79</u>	**−20.66**	42.07
DTO†	19.30	22.84	−47.14	<u>57.11</u>
AGTTU	<u>0.69</u>	23.46	−34.44	**57.24**

† denotes quoted results. TPO refers to TPO_{GDR} from Zhou et al. (2026, Table 1). DTO refers to DTO w/o KL from Lee et al. (2026, Table 2), selected for its better forgetting-related metrics; all four values come from that row. Rankings cover the listed variants, excluding Retain; PrivLeak is ranked by absolute value.

Table 4: Token-selection ablation on TOFU / Llama-3.2-1B.

	TOFU / Llama-3.2-1B					
	Forget 1%		Forget 5%		Forget 10%	
Selection	FQ ↑	MU ↑	FQ ↑	MU ↑	FQ ↑	MU ↑
Low-p	0.0971	**0.5819**	0.0002	<u>0.5660</u>	1.61×10^{-10}	0.5676
Random	<u>0.7659</u>	<u>0.5818</u>	<u>0.0006</u>	0.5645	<u>0.0006</u>	**0.5919**
Top-p (AGTTU)	**0.9900**	0.5567	**0.5453**	**0.5727**	**0.1546**	<u>0.5709</u>

Equal token counts and all other settings are fixed; all runs report epoch 10. Top-p reuses AGTTU results.

advantages over TPO_{GDR} concern verbatim unlearning and retained utility, rather than all forgetting and privacy metrics.

4.3 Ablation Studies

4.3.1 Token Selection Ablation

Effects of selection on forget quality. Table 4 compares Top-p, Low-p, and Random on TOFU / Llama-3.2-1B. Token counts are equal, other settings match the corresponding main runs, and all report the final epoch-10 checkpoint. Top-p exceeds Random in FQ by 0.2241/0.5446/0.1540. At 5% and 10%, Random and Low-p yield near-zero FQ, versus 0.5453 and 0.1546 for Top-p, consistent with the masking observation in Appendix F. **With equal token counts and matched training settings, high-attention selection yields higher FQ than random or low-attention selection.**

Effects of selection on model utility. At 5%, Top-p increases MU over Random by 0.0082, improving both metrics. At 1% and 10%, it has higher FQ but lower MU by 0.0251 and 0.0210, respectively. **High-attention selection improves both metrics at 5%, but higher forget quality does not always accompany higher retained utility.**

4.3.2 CLPL Component Ablation

Effects of the margin and non-target penalty. Table 5 fixes AGTTU selection on TOFU / Llama-3.2-1B and varies only $m \in \{0, 1.0\}$ and $\gamma \in \{0, 0.1\}$. Setting $m = \gamma = 0$ retains LPL under the same mask and retain CE; the last row is full AGTTU. At 1%, margin alone raises FQ from 0.4046 to 0.7659, and both components yield 0.9900. At 5%, the non-target penalty alone yields 0.2205, versus 0.5453 with both. **Combining both components yields the highest ablation FQ at 1% and 5%, but not the highest MU.**

Table 5: CLPL component ablation on TOFU / Llama-3.2-1B.

		TOFU / Llama-3.2-1B					
		Forget 1%		Forget 5%		Forget 10%	
m	γ	FQ ↑	MU ↑	FQ ↑	MU ↑	FQ ↑	MU ↑
0	0	0.4046	<u>0.5655</u>	0.0006	<u>0.5733</u>	**0.2812**	<u>0.5771</u>
1.0	0	<u>0.7659</u>	0.5527	0.0014	0.5732	2.15×10^{-100}	0.5717
0	0.1	0.4046	**0.5708**	<u>0.2205</u>	**0.5753**	0.1112	**0.5785**
1.0	0.1	**0.9900**	0.5567	**0.5453**	0.5727	<u>0.1546</u>	0.5709

CLPL component ablation with fixed AGTTU token selection. Other settings match the main experiment; all runs report epoch 10. The last row is full AGTTU.

Component benefits across forgetting ratios. At 10%, $m = \gamma = 0$ yields FQ/MU of 0.2812/0.5771, both above full CLPL's 0.1546/0.5709. Margin alone yields FQ of 2.15×10^{-100}; adding the non-target penalty raises it to 0.1546. **CLPL's component benefits depend on the forgetting ratio: the full configuration's FQ advantage at 1% and 5% does not extend to 10%.**

Selection and optimization. The two ablations address complementary questions. Holding CLPL fixed tests whether attention identifies more effective targets than alternative selections at the same budget. Holding selection fixed tests whether the margin and non-target penalty improve the objective for those targets. High-attention selection has higher FQ across all three ratios, whereas the full CLPL configuration leads in FQ only at 1% and 5%. The evidence for target identification is therefore distinct from the evidence for individual loss components. **The results support attention-guided selection, while showing that the benefits of combining the CLPL components depend on the forgetting setting.**

5 Conclusion

AGTTU addresses target identification and targeted optimization within answers. Masking high-attention tokens affects the probability of the complete original answer more than masking equally many random tokens. Motivated by this observation, AGTTU selects targets by the attention they receive from subsequent answer positions, extracted from prefill on complete question–answer pairs without auxiliary identifiers or tokenwise perturbations. CLPL extends LPL with a margin on average target-logit reduction and a one-sided non-target penalty, complemented by retain-set cross-entropy to balance suppression and utility.

Experiments on TOFU and MUSE Books show that AGTTU improves the forgetting–utility trade-off across multiple settings. On TOFU with Llama-3.2-1B, it achieves the highest FQ and MU among the listed methods across all three forgetting ratios. On MUSE Books, it reduces verbatim memorization and improves retained knowledge compared with the reported TPO variant. Selection ablations with Llama-3.2-1B show consistently higher FQ for high-attention selection than random or low-attention selection. Combining the CLPL components yields higher FQ at the 1% and 5% forgetting ratios, although the gains vary with the ratio.

AI USE STATEMENT

Generative AI assisted with literature search, organization, language editing and translation, equation checking, presentation of experimental results, and figure and table preparation. Numerical results come from experiment summaries or the source tables identified in Appendix E.1; AI assistance did not supply unperformed experimental observations or unreported variance.

ETHICS STATEMENT

This work studies designated-content unlearning to support privacy protection and responsible data use. Experiments use the public TOFU and MUSE benchmarks without collecting additional personal or sensitive data. Unlearning should be authorized and purpose-specific. Metric improvements do not establish verifiable data deletion or elimination of privacy risks.

REPRODUCIBILITY STATEMENT

We use official TOFU (Maini et al., 2024) and MUSE (Shi et al., 2025) data and the OpenUnlearning evaluation framework (Dorna et al., 2025). Section 3 and Appendix B describe selection and training; Appendix C lists configurations, and Appendix E.1 identifies initial models and result sources.

REFERENCES

Lucas Bourtoule, Varun Chandrasekaran, Christopher Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *IEEE Symposium on Security and Privacy*, 2021. URL https://security.csl.toronto.edu/blog/wp-publications/bourtouleoaklandunlearning/.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *IEEE Symposium on Security and Privacy*, 2015. doi: 10.1109/SP.2015.35. URL https://www.cs.columbia.edu/~junfeng/papers/unlearning/.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium*, pp. 2633–2650. USENIX Association, 2021. URL https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

Vineeth Dorna, Anmol Mekala, Wenlong Zhao, Andrew McCallum, Zachary C. Lipton, J. Zico Kolter, and Pratyush Maini. OpenUnlearning: Accelerating LLM unlearning via unified benchmarking of methods and metrics. *arXiv preprint arXiv:2506.12618*, 2025. URL https://arxiv.org/abs/2506.12618.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.751. URL https://aclanthology.org/2023.emnlp-main.751/.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1357. URL https://aclanthology.org/N19-1357/.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14389–14408. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.805. URL https://aclanthology.org/2023.acl-long.805/.

Seunghee Koh, Sunghyun Baek, Youngdong Kim, and Junmo Kim. Forget what matters, keep the rest: Selective unlearning of informative tokens. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25639–25655. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1175. URL https://aclanthology.org/2026.acl-long.1175/.

Hong Kyu Lee, Ruixuan Liu, and Li Xiong. Direct token optimization: A self-contained approach to large language model unlearning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 42083–42100, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.2088. URL https://aclanthology.org/2026.findings-acl.2088/.

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. URL https://arxiv.org/abs/2402.16835.

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. URL https://arxiv.org/abs/2401.06121.

Vaidehi Ramesh Patil, Peter Hase, and Mohit Bansal. Can sensitive information be deleted from LLMs? objectives for defending against extraction attacks. In *International Conference on Learning Representations*, 2024. URL

https://proceedings.iclr.cc/paper_files/paper/2024/hash/c652e5f62fd1f5acbbf0d6413a1113e7-Abstract-Conference.html.

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 *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/4556f5398bd2c61bd7500e306b4e560a-Abstract-Conference.html.

Naman Deep Singh, Maximilian Müller, Francesco Croce, and Matthias Hein. Unlearning that lasts: Utility-preserving, robust, and almost irreversible forgetting in LLMs. *arXiv preprint arXiv:2509.02820*, 2025. URL https://arxiv.org/abs/2509.02820.

Yixin Wan, Anil Ramakrishna, Kai-Wei Chang, Volkan Cevher, and Rahul Gupta. Not every token needs forgetting: Selective unlearning to limit change in utility in large language model unlearning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 1827–1835. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.96. URL https://aclanthology.org/2025.findings-emnlp.96/.

Lingzhi Wang, Xingshan Zeng, Jinsong Guo, Kam-Fai Wong, and Georg Gottlob. Selective forgetting: Advancing machine unlearning techniques and evaluation in language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(1):843–851, 2025. doi: 10.1609/aaai.v39i1.32068. URL https://ojs.aaai.org/index.php/AAAI/article/view/32068.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 11–20. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1002. URL https://aclanthology.org/D19-1002/.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 105425–105475, 2024. doi: 10.52202/079017-3346. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/be52acf6bccf4a8c0a90fe2f5cfcead3-Abstract-Conference.html.

Chaewon Yoon, Dongjun Kim, and Hyun-Je Song. Selective span-level unlearning for large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 423–431. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-short.35. URL https://aclanthology.org/2026.acl-short.35/.

Mert Yuksekgonul, Varun Chandrasekaran, Erik Jones, Suriya Gunasekar, Ranjita Naik, Hamid Palangi, Ece Kamar, and Besmira Nushi. Attention satisfies: A constraint-satisfaction lens on factual errors of language models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.15098.

Saleh Zare Zade, Xiangyu Zhou, Sijia Liu, and Dongxiao Zhu. Attention smoothing is all you need for unlearning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/2163fff81515bd2d3dfca8577304456f-Abstract-Conference.html.

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

Xiangyu Zhou, Yao Qiang, Saleh Zare Zade, Douglas Zytko, Prashant Khanduri, and Dongxiao Zhu. Not all tokens are meant to be forgotten. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(44):38173–38182, 2026. doi: 10.1609/aaai.v40i44.41156. URL https://ojs.aaai.org/index.php/AAAI/article/view/41156.

A CLPL: Design Rationale and Scope

CLPL uses a single preference score to balance target suppression and non-target damage. It is an empirical optimization design whose benefits require experimental evaluation, rather than following from local properties. The selection ablation holds CLPL fixed, while Section 4.3.2 fixes selection and evaluates the margin and non-target penalty. Component benefits vary across forgetting ratios and do not establish probability preservation.

Stable computation and optimization signals. Let $a = \beta(D_T - m) - \gamma C_N$. The per-sample loss can be computed stably as

$$\ell_f = -\frac{2}{\beta}\log\sigma(a) = \frac{2}{\beta}\,\mathrm{softplus}(-a). \tag{6}$$

The aggregate derivatives describe how target reduction and non-target penalties jointly affect optimization:

$$\frac{\partial \ell_f}{\partial D_T} = -2\sigma(-a), \qquad \frac{\partial \ell_f}{\partial C_N} = \frac{2\gamma}{\beta}\sigma(-a). \tag{7}$$

Both share $\sigma(-a)$, so a large preference score weakens both suppression and protection gradients. Local output-logit derivatives do not ensure monotonic logit changes under shared-parameter updates, which also depend on the retain loss. When $N = \varnothing$, we set $C_N = 0$ and omit the non-target term; the positive part may use the zero ReLU subgradient at zero.

Logit protection and probability changes. Protecting a ground-truth raw logit does not preserve its generation probability or the full vocabulary distribution. Softmax is invariant to a common logit shift within a position, whereas a raw-logit loss may change. A lower loss alone is therefore insufficient evidence of lower probabilities or knowledge unlearning. This output-space observation neither establishes that our parameter-freezing scheme permits arbitrary shifts nor shows that such degeneration occurs in training. These boundaries delimit the design rationale; they do not replace empirical evaluation of unlearning and utility.

B Reference Algorithm and Implementation Conventions

The following reference training procedure specifies prediction-position alignment, short-answer handling, and caching requirements.

Table 6: Reference training procedure for AGTTU. All target positions are selected using the original model, and complete teacher-forced prefixes are preserved.

AGTTU training procedure	
Input	Original model θ_0, forget set $\mathcal{D}_f$, retain subset $\widetilde{\mathcal{D}_r}$, $p, \beta, m, \gamma, \lambda$, and number of epochs $E = 10$.
Output	Final model θ_u after epoch 10.
1	Initialize $\theta \leftarrow \theta_0$. Freeze all reference-model parameters. Enable gradients for the Transformer blocks of the updated model and freeze parameters outside these blocks.
2	Concatenate questions and ground-truth answers using the chat template. Record each answer's starting position in the complete input and exclude padding and special tokens.
3	With dropout disabled and gradients off, prefill the concatenated input and read the last-layer attention matrix averaged across heads. Compute answer-only column-average scores using Eq. (2) and select T, N.
4	Cache T, N, the valid-answer mask, and the ground-truth reference score $z_i^{\theta_0}$ at each prediction position. Do not cache full-vocabulary outputs.
5	For $e = 1, \ldots, E$, sample batches from the forget set and retain subset. Compute current z_i^{θ} on complete forget samples and obtain D_T, C_N using Eq. (3).
6	Compute $\mathcal{L}_f^{\mathrm{AGTTU}}$ using Eq. (6) and retain-set cross-entropy using Eq. (5). Average each loss within samples and then across the batch, and combine them with the weights in Eq. (5).
7	Backpropagate and update the trainable parameters. Complete all 10 epochs and evaluate the final checkpoint. No intermediate checkpoint is selected using validation/test data, FQ, or MU.

Prediction-position alignment. As described in Section 3, the attention key indexes the token itself, while the corresponding output logit comes from the preceding position. Apply the target mask after shifting logits and labels into alignment; using the same unshifted index for both selects the wrong prediction.

Short answers and selection counts. To avoid division by zero for empty sets, we constrain k to $[1, n-1]$ when $n \geq 2$ and optimize only the target term for single-token answers. The realized fraction is $|T|/n$; rounding can make it differ from p for short answers. Selection operates on individual tokens and does not require all subwords of an entity to be selected together.

Final positions. Equation (2) measures a token's involvement in subsequent answer generation and excludes self-attention. The last answer token has no subsequent answer tokens, so its score is set to zero without adding an EOS query. This follows the definition based on subsequent attention and does not imply that the final token contains no knowledge to be forgotten.

Cache consistency. Reference-score caching requires a fixed tokenizer, chat template, truncation length, padding side, and model state. We use fixed target masks selected offline, without reselection during training. Sample masks must remain consistent across distributed workers. Global loss averages should be weighted by valid sample or token counts rather than simply averaging local batches of different lengths, which can change their effective weights.

C Training Configurations and Hyperparameters

Our AGTTU main experiments, selection ablations, and component ablations use `paged_adamw_32bit`, learning rate 10^{-5}, and global batch size 32 (16×2) on two NVIDIA A6000 GPUs. Table 7 covers these runs only; it does not specify settings for literature results or other baseline entries. The retention-loss weight is $\lambda = 1$ for AGTTU and both ablation studies. Table 8 gives the remaining AGTTU hyperparameters.

Training duration and checkpoint rule. All experiments run in this study, including our TOFU / Llama-3.2-3B TPO reproduction, complete 10 epochs and report the final checkpoint at the end of epoch 10. We do not select intermediate checkpoints using validation/test data or FQ/MU. Quoted results follow their source papers' reporting protocols and are not assigned this final-checkpoint rule.

Table 7: Optimization settings for our AGTTU runs and both ablation studies only. Batch size is global; each run reports its final epoch-10 checkpoint.

Experiments	Optimizer	Batch size	Learning rate
TOFU AGTTU (1B, 3B)	`paged_adamw_32bit`	32	10^{-5}
MUSE Books AGTTU	`paged_adamw_32bit`	32	10^{-5}
TOFU selection ablation (1B)	`paged_adamw_32bit`	32	10^{-5}
TOFU component ablation (1B)	`paged_adamw_32bit`	32	10^{-5}

AGTTU enables updates for parameter names matching `model.layers.`, covering attention, feed-forward, and within-block normalization parameters in Transformer blocks. Parameters outside these blocks, including embeddings, final normalization, and the output head, are frozen. The retain subset is sampled randomly from the remaining data and has the same number of samples as the forget set. For example, when the TOFU 1% forget set contains 40 question–answer pairs, 40 retain samples are drawn from the remaining 3,960 pairs.

Selection-strategy ablation settings. The ablation uses TOFU and Llama-3.2-1B, varying only the target token selection strategy. All other settings match our AGTTU main experiment at the corresponding forgetting ratio. The three strategies share CLPL and the target-token fractions in Table 8: Top-p selects the highest-attention positions, Low-p selects the lowest-attention positions, and Random selects the same number of positions at random. Model initialization, data splits, retain-set constraints, loss hyperparameters, trainable parameters, optimizer, learning rate, batch size, training epochs, and evaluation procedures are held fixed. Each strategy is evaluated at the end of epoch 10.

This controlled comparison concerns the selection ablation, not the mixed-source main tables. Top-p reuses the corresponding AGTTU main-experiment results.

CLPL component ablation settings. On TOFU / Llama-3.2-1B, we fix AGTTU token selection and use the same target fractions and other settings as the corresponding main experiments. Only $m \in \{0, 1.0\}$ and $\gamma \in \{0, 0.1\}$ vary; $\beta = 0.5$, $\lambda = 1$, and all runs use the final epoch-10 checkpoint. Table 5 reports the four configurations; the full configuration reuses the AGTTU main results.

Table 8: AGTTU target selection and loss hyperparameters. p is the target-token fraction. All settings use $\beta = 0.5$, $m = 1.0$, $\gamma = 0.1$, and $\lambda = 1$, where λ weights retain-set cross-entropy.

Experimental setting		**AGTTU hyperparameters**				
Dataset / model	Forgetting setting	p	β	m	γ	λ
TOFU / Llama-3.2-1B	1%	0.30	0.5	1.0	0.1	1
	5%, 10%	0.20	0.5	1.0	0.1	1
TOFU / Llama-3.2-3B	1%, 5%, 10%	0.20	0.5	1.0	0.1	1
MUSE	Books	0.20	0.5	1.0	0.1	1

D Relationship to the Original TPO Loss

For a fixed target mask T, TPO uses

$$\mathcal{L}_{\text{LPL}}(\theta; T) = -\mathbb{E}_{(x,y)\sim\mathcal{D}_f}\left[\frac{2}{\beta}\log\sigma\left(\frac{\beta}{|T|}\sum_{i\in T}(z_i^{\theta_0} - z_i^{\theta})\right)\right]. \tag{8}$$

Target-logit differences are averaged within each sample before the preference loss and sample expectation.

We compare the objectives using Eqs. (3)–(5) of the published TPO paper (Zhou et al., 2026). Its LPL uses raw logits of target tokens rather than softmax probabilities or log-softmax values. Denoting general-word positions by N, PL can be expressed in masked form while preserving complete teacher-forced prefixes as

$$\mathcal{L}_{\text{PL}} = -\mathbb{E}_{(x,y)\sim\mathcal{D}_f}\sum_{j\in N}\log P_\theta(y_j \mid x, y_{<j}), \qquad \mathcal{L}_{\text{TPO}} = \mathcal{L}_{\text{LPL}} + \mathcal{L}_{\text{PL}}. \tag{9}$$

The original paper abbreviates PL as $-\mathbb{E}\log P_\theta(\bar{y} \mid x)$. We make the position mask and prefix explicit to avoid changing the conditioning context by concatenating non-target words into a new sentence. Equation (4) of TPO explicitly averages the logit differences in LPL over the number of targets, and we retain this normalization. PL is written here as a sum over positions; averaging over non-target positions would change its relative weight against LPL.

Expanding Eq. (4), the AGTTU unlearning term is

$$\mathcal{L}_f^{\text{AGTTU}} = -\mathbb{E}_{(x,y)\sim\mathcal{D}_f}\left[\frac{2}{\beta}\log\sigma\left(\beta\left(\frac{1}{|T|}\sum_{i\in T}(z_i^{\theta_0} - z_i^{\theta}) - m\right) - \frac{\gamma}{|N|}\sum_{j\in N}\max(0, z_j^{\theta_0} - z_j^{\theta})\right)\right]. \tag{10}$$

Relative to TPO, AGTTU replaces the information identifier with an attention-based selector, introduces a margin into LPL's average target reduction, and replaces continued PL reinforcement on forget-set answers with a one-sided penalty on non-target decreases. Cross-entropy on a separate retain set is a distinct constraint and should not be conflated with PL. Setting $m = \gamma = 0$ recovers only the LPL unlearning term; the full algorithms still differ in their selectors, retain data, and parameter-freezing schemes.

E Result Sources and Additional Discussion

E.1 Result Sources and Reporting Scope

Tables 1, 2, and 3 mark quoted results with †; our method and reproductions carry no additional marker. The MUSE table uses the short names TPO and DTO, with the specific variants defined below the table. Table 9 specifies the attributed result groups; other entries are retained from the experiment summary. Each FQ–MU pair or MUSE metric tuple is kept together, without combining sources or checkpoints. Our runs use the final epoch-10 checkpoint, without selecting intermediate checkpoints using validation/test data, FQ, or MU. Quoted results retain their source protocols. Rankings and differences use unrounded values, with ties marked equally. Ordinary TOFU decimals use four places; scientific notation retains the original precision.

Table 9: Sources of the explicitly attributed result groups. Literature results retain their original reporting protocols.

Result group	**Source and reporting scope**
TOFU 1B: TPO (ChatGPT), DTO	Quoted from Lee et al. (2026, Tables 5–7), for 1%, 5%, and 10% forgetting. DTO is the full method. TPO was introduced by Zhou et al. (2026); its 1B numerical results here are sourced from Lee et al.
TOFU 3B: TPO	Reproduced in this study; final checkpoint after 10 epochs.
MUSE Books: TPO_{GDR}	Quoted from Zhou et al. (2026, Table 1). GDR is retain-set gradient descent.
MUSE Books: DTO w/o KL	Quoted from Lee et al. (2026, Table 2). We select this complete result row for its lower VerbMem_f and KnowMem_f and PrivLeak closer to zero; full DTO has higher retained utility.
AGTTU and both ablation studies	Experiments in this study; final checkpoints after 10 epochs.

The main tables compare reported benchmark outcomes; the selection ablation provides the controlled comparison of target-selection strategies. MUSE conclusions concern the listed variants and do not extend from DTO w/o KL to full DTO.

Attention-based identification and semantic coverage. Attention interventions analyze factual recall (Geva et al., 2023), and attention allocation has been linked to prompt constraints and factual errors (Yuksekgonul et al., 2024). These analyses do not directly supply a parameter-unlearning objective. Since attention interpretations depend on intervention design (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019), we evaluate its selection role through masking comparisons and selection ablations. The score measures contextual information use, rather than identifying all sensitive entities. Zeroing attention columns also changes total attention mass, so the observed probability changes depend on the intervention. We use this result as empirical motivation and evaluate selection in unlearning through the controlled ablation.

Complementary robustness evaluations. Lynch et al. (2024) show that a single generation metric can miss recoverable knowledge, while Patil et al. (2024) evaluate deletion using black-box and white-box extraction attacks. Together with JensUn's paraphrase and relearning evaluations (Singh et al., 2025), these studies show that benchmark metrics and attack or relearning tests assess different aspects of unlearning. Our evaluation covers the reported benchmarks and ablations, excludes these attack and relearning tests, and makes no claim of irreversible forgetting.

Initial models and evaluation implementation. For our AGTTU experiments and selection ablations, we use the official TOFU (Maini et al., 2024) and MUSE (Shi et al., 2025) datasets and the open-source OpenUnlearning evaluation code (Dorna et al., 2025). Its public repository[1] provides an evaluation framework for these datasets. Our TOFU runs start from the Llama-3.2-1B and Llama-3.2-3B checkpoints fine-tuned on TOFU and released by OpenUnlearning; our MUSE Books runs use the official MUSE Llama-2-7B checkpoint fine-tuned on *Harry Potter* text. Appendix C gives the applicable training settings. Quoted results are identified in Table 9; they are not evaluations rerun under our configuration.

[1] https://github.com/locuslab/open-unlearning

F Attention Masking Experiment

Figure 3 presents the masking comparison described in the introduction, conducted with Llama-3.2-1B on the TOFU 10% forget set. At equal masking ratios, masking high-attention answer tokens reduces complete-original-answer probability more than random masking. This provides empirical motivation for target selection; the experiment measures original-answer probability rather than free-generation accuracy.

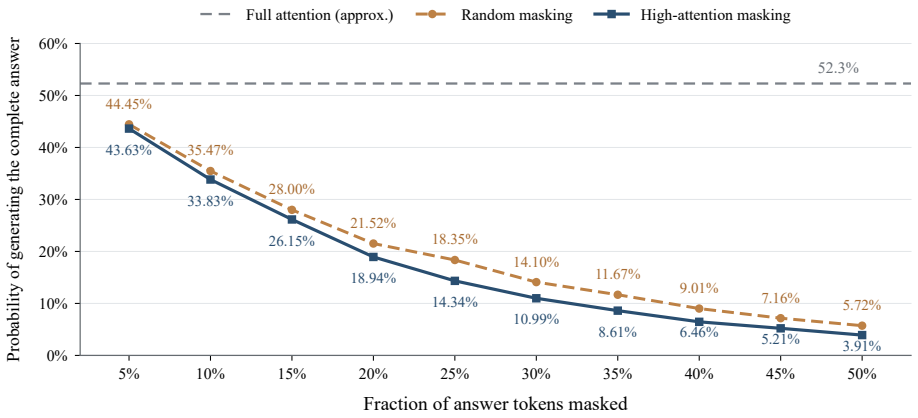


Figure 3: Attention masking with Llama-3.2-1B on the TOFU 10% forget set. The horizontal axis is the fraction of answer tokens masked; the vertical axis is the probability of generating the complete answer. High-attention masking reduces complete-answer probability more than random masking at equal ratios. Point labels reproduce the values in the original experiment figure; the gray dashed line shows the approximate unmasked reference read from that figure.