Forget What Matters, Keep the Rest: Selective Unlearning of Informative Tokens

Seunghee Koh[1] **Sunghyun Baek**[1] **Youngdong Kim**[2†] **Junmo Kim**[1†]

[1]Korea Advanced Institute of Science and Technology, South Korea
[2]Hanbat National University, South Korea

{seunghee1215, baeksh, junmo.kim}@kaist.ac.kr ydkim1293@hanbat.ac.kr

Abstract

Unlearning in large language models (LLMs) has emerged as a promising safeguard against adversarial behaviors. When the forgetting loss is applied uniformly without considering token-level semantic importance, model utility can be unnecessarily degraded. Recent studies have explored token-wise loss regularizers that prioritize informative tokens, but largely rely on ground-truth confidence or external linguistic parsers, which limits their ability to capture contextual information or the model's overall predictive state. Intuitively, function words like "the" primarily serve syntactic roles and are highly predictable with little ambiguity, but informative words admit multiple plausible alternatives with greater uncertainty. Based on this intuition, we propose Entropy-guided Token Weighting (ETW), a token-level unlearning regularizer that uses entropy of the predictive distribution as a proxy for token informativeness. We demonstrate that informative tokens tend to have higher entropy, whereas structural tokens tend to have lower entropy. This behavior enables ETW to achieve more effective unlearning while better preserving model utility than existing token-level approaches.

1 Introduction

Machine unlearning (Bourtoule et al., 2021; Golatkar et al., 2020, 2021; Koh et al., 2023) aims to train deep neural networks to selectively remove specific knowledge while retaining other knowledge. It has emerged as a promising approach for removing portions of training corpora in large language models (LLMs), where it can serve as a defense against membership inference, jailbreak, and red-teaming attacks (Zhang et al., 2025; Lin et al., 2024; Hong et al., 2024). In LLMs, unlearning follows naturally from the next-token prediction objective, which decomposes into per-token losses, motivating token-wise loss reweighting as a principled unlearning strategy. Prior work often uses ground-truth confidence as a token-level signal for unlearning (Wang et al., 2025a; Yang et al., 2025). Yet, confidence-based signals alone are insufficient to determine how strongly each token should be reweighted, leaving room for richer representational signals to guide token-wise penalties.

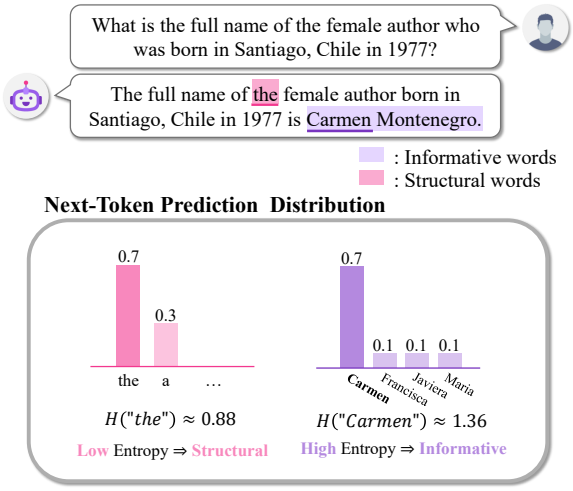


Figure 1: Motivation of Entropy-guided Token Weighting (ETW). Informative tokens such as "Carmen" exhibit higher entropy due to multiple plausible alternatives, while structural tokens such as "the" have lower entropy because they exhibit little ambiguity.

A key question for effective unlearning is which tokens should be penalized more strongly while minimizing utility degradation. We distinguish between *informative tokens*, which carry important semantic content, and *structural tokens*, which serve mainly syntactic or repetitive functions. As illustrated in Figure 1, informative tokens such as the proper noun "Carmen" admit multiple plausible alternatives and thus involve greater predictive uncertainty, whereas structural tokens such as the function word "the" leave little ambiguity in next-token prediction. This contrast suggests that entropy over the predictive distribution provides a

[†]Co-corresponding authors

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 25639–25655

useful signal for assessing token informativeness, even when ground-truth confidence is similar.

Based on this intuition, we propose *Entropy-guided Token Weighting* (ETW), a token-wise regularizer that penalizes high-entropy tokens and downweights low-entropy ones. Unlike prior approaches that rely solely on the confidence of the ground-truth token, ETW quantifies informativeness by leveraging the model's predicted distribution over all candidates in the vocabulary. Specifically, entropy serves as a richer proxy for informativeness by reflecting how probability mass is distributed among competing alternatives.

ETW more accurately identifies informative tokens than existing token-level regularizers. This improved identification enables effective forgetting while preserving model utility across a range of unlearning settings. The model unlearned with ETW also produces informative token prediction probabilities closest to those of the retrained model.

The main contributions of this paper are as follows:

- We introduce Entropy-guided Token Weighting (ETW), a token-wise regularizer for LLM unlearning that leverages entropy over the overall predictive distribution, rather than relying solely on ground-truth confidence.
- We show that ETW is an effective measure for distinguishing informative tokens from structural tokens to identify informative tokens for selective penalization.
- By focusing on informative tokens during unlearning, ETW consistently achieves effective forgetting while preserving model utility.

2 Related Works

Sample-Level LLM Unlearning LLM unlearning updates model parameters to selectively remove knowledge associated with a target corpus while aiming to preserve generalization. A basic approach applies gradient ascent (GA) loss (Maini et al., 2024) to penalize the ground-truth token in the output space. Variants such as LOKU (Cha et al., 2025) and UNDIAL (Dong et al., 2025) redirect probability mass toward second-best tokens through loss or logit modulation.

DPO-inspired methods contrast target-corpus responses as losing answers against substitute responses as winning answers (Rafailov et al., 2023; Mekala et al., 2025), which requires additional generation for winning answers. Although variants such as NPO (Zhang et al., 2024) and SimNPO (Fan et al., 2024) remove explicit winning-answer dependencies, their objectives can ultimately be formulated as weighted GA. Except for SimNPO, they usually rely on a reference model.

RMU perturbs representations of the forget set, but is sensitive to perturbation design (Wang et al., 2025a) and relies on reference-model distillation for retaining knowledge.

Token-Level LLM Unlearning Several works study token-level LLM unlearning through token-wise regularizers built on gradient ascent (GA)–based objectives, which modulate unlearning strength across tokens. WGA and TNPO (Wang et al., 2025a) down-weight low-confidence tokens to control GA divergence, but penalize tokens solely based on confidence and fail to account for semantic importance within a forget sample.

SatImp (Yang et al., 2025) and FUNDIAL (Dong et al., 2025) aim to reduce unnecessary degradation of general language capabilities by prioritizing informative tokens. SatImp reweights tokens using confidence and its complement, while FUNDIAL, a variant of UNDIAL, introduces a linguistically motivated strategy based on spaCy (Honnibal et al., 2020), hypothesizing that nouns and named entities encode core knowledge. However, SatImp relies on confidence values that are insufficient for identifying informative contents, and FUNDIAL depends on part-of-speech categories that do not explicitly incorporate contextual information.

ETW leverages entropy derived from the full candidate vocabulary distribution to selectively penalize informative tokens, rather than relying on confidence or coarse linguistic categories.

Token-level Control Beyond unlearning, token-level control has been explored in LLMs across a range of tasks, including informative-token selection in supervised fine-tuning (Pang et al., 2025) and entropy-based token weighting in reinforcement learning with verifiable rewards (RLVR) (Wang et al., 2025b; Cheng et al., 2026).

For unlearning, we introduce an entropy-based token weighting scheme and demonstrate why entropy is particularly well-suited to this task (Section 4). Our method achieves effective forgetting while preserving model utility.

3 Entropy-guided Token Weighting

3.1 Problem Setting

We consider an autoregressive language model parameterized by $\boldsymbol{\theta}$ and trained on a sequence with a prompt $\mathbf{x} = (x_1, \ldots, x_m)$ with length m and a completion $\mathbf{y} = (y_1, \ldots, y_n)$ with length n.

LLM unlearning aims to selectively forget targeted data while preserving knowledge from the retained data, which requires both a forgetting loss and a knowledge-retaining loss.

For knowledge retaining, we use the standard cross-entropy loss:

$$\mathcal{L}_r(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}) = -\log p(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})\,. \tag{1}$$

For knowledge forgetting, we adopt the negative cross-entropy loss, also known as the gradient-ascent (GA) objective, whose token-wise decomposition is given by

$$\begin{aligned}\mathcal{L}_{\mathrm{GA}}(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}) &= \log p(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}) \\ &= \sum_{i=1}^{n} \log p(y_i|y_{<i},\mathbf{x};\boldsymbol{\theta}).\end{aligned} \tag{2}$$

Accordingly, we use the following token-wise weighted GA loss with weight $\omega_i(\mathbf{x},\mathbf{y})$ as the primary objective:

$$\mathcal{L}_f(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}) = \sum_{i=1}^{n} \omega_i(\mathbf{x},\mathbf{y})\ \log p(y_i|y_{<i},\mathbf{x};\boldsymbol{\theta}). \tag{3}$$

Given a retaining sample $(\mathbf{x}_r,\mathbf{y}_r) \in \mathcal{D}_r$ and a forgetting sample $(\mathbf{x}_f,\mathbf{y}_f) \in \mathcal{D}_f$, the overall unlearning objective is defined as

$$\mathcal{L}(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}) = \mathcal{L}_r(\mathbf{y}_r|\mathbf{x}_r;\boldsymbol{\theta}) + \lambda\mathcal{L}_f(\mathbf{y}_f|\mathbf{x}_f;\boldsymbol{\theta}), \tag{4}$$

where λ controls the degree of knowledge removal associated with forgetting loss.

3.2 The Formulation of Entropy-guided Token Weighting

Within a single forget sample, tokens can be broadly categorized into two types: *informative tokens* and *structural tokens*, defined as follows.

Informative Tokens carry the core semantic content of the answer.
Structural Tokens primarily serve syntactic or repetitive roles, such as function words or prompt mentions repeated in the completion.

As it is essential to selectively forget informative content while preserving overall utility, we aim to apply stronger penalties to informative tokens. We hypothesize that token-wise entropy can indicate token informativeness. Specifically, entropy measures how uncertain the model is about its next-token prediction. If the model is highly confident at a given position, the token is likely to be structural and carries limited semantic information. In contrast, if the model is uncertain and distributes probability mass across multiple candidate tokens, the position is more likely to encode informative or knowledge-related content.

Formally, the token-wise entropy for the i-th token in a completion $\mathbf{y}$ is defined over the vocabulary $\mathcal{V}$ as

$$\begin{aligned}&H(y_i|y_{<i},\mathbf{x};\boldsymbol{\theta}) \\ &= -\sum_{v\in\mathcal{V}} p(v|y_{<i},\mathbf{x};\boldsymbol{\theta}) \log p(v|y_{<i},\mathbf{x};\boldsymbol{\theta})\,,\end{aligned} \tag{5}$$

From a mathematical perspective, even for a fixed ground-truth (GT) confidence $p_i := p(y_i|y_{<i},\mathbf{x};\boldsymbol{\theta})$, the range of entropy values can vary substantially depending on the distribution of non-GT tokens. The minimum achievable entropy occurs when the remaining probability mass $(1-p_i)$ is entirely concentrated on the second-best token:

$$H_{\min}(p_i) = -p_i\log p_i - (1-p_i)\log(1-p_i). \tag{6}$$

The maximum entropy is achieved when the residual probability mass $(1-p_i)$ is uniformly distributed across the remaining $|\mathcal{V}|-1$ vocabulary tokens:

$$H_{\max}(p_i) = -p_i\log p_i - (1-p_i)\log\frac{1-p_i}{|\mathcal{V}|-1}. \tag{7}$$

The fact that the distribution over non-ground-truth tokens governs the gap between $H_{\min}(p_i)$ and $H_{\max}(p_i)$ demonstrates that entropy offers a significantly richer representational range than measures based solely on ground-truth confidence.

To this end, we propose *Entropy-guided Token Weighting* (ETW), a token-wise reweighting scheme derived from the model's next-token predictive distribution. We normalize the entropy weights of completion tokens so that their sum equals the completion length n. This normalization preserves the overall scale of the unlearning loss by redistributing unlearning strength across tokens within

Regularizer	Explanation	Formulation
	Confidence-based token regularizer	
WGA (Wang et al., 2025a)	Penalize high-confidence tokens via exponentiated confidence	$\omega_i^{\mathrm{WGA}}(\mathbf{x},\mathbf{y};\boldsymbol{\theta}) = p(y_i\|y_{<i},\mathbf{x};\hat{\theta})^{\alpha}$
Imp (Yang et al., 2025)	Penalize low-confidence tokens using confidence complement	$\omega_i^{\mathrm{Imp}}(\mathbf{x},\mathbf{y};\boldsymbol{\theta}) = 1 - p(y_i\|y_{<i},\mathbf{x};\hat{\theta})$
SatImp (Yang et al., 2025)	Token penalization via joint confidence and confidence-complement weighting	$\omega_i^{\mathrm{SatImp}}(\mathbf{x},\mathbf{y};\boldsymbol{\theta}) = p(y_i\|y_{<i},\mathbf{x};\hat{\boldsymbol{\theta}})^{\alpha}\left(1 - p(y_i\|y_{<i},\mathbf{x};\hat{\boldsymbol{\theta}})\right)$
TNPO (Wang et al., 2025a)	Penalize tokens with small confidence deviation from a reference model	$\omega_i^{\mathrm{TNPO}}(\mathbf{x},\mathbf{y};\boldsymbol{\theta}) = \frac{2\,p(y_i\|y_{<i},\mathbf{x};\hat{\boldsymbol{\theta}})^{\beta}}{p(y_i\|y_{<i},\mathbf{x};\hat{\boldsymbol{\theta}})^{\beta} + p(y_i\|y_{<i},\mathbf{x};\boldsymbol{\theta}_{ref})^{\beta}}$
	Linguistic-based token regularizer	
SCN (Dong et al., 2025)	Linguistic noun-based token selection using an external parser (spaCy)	$\omega_i^{\mathrm{SCN}}(\mathbf{x},\mathbf{y}) = \mathbb{I}\left[y_i \in \mathcal{E}_{\mathrm{Noun}}^{\mathrm{spaCy}}\right],\ \mathcal{E}_{\mathrm{Noun}}^{\mathrm{spaCy}} \subseteq \mathcal{V}$
SCE (Dong et al., 2025)	Linguistic entity-based token selection using an external parser (spaCy)	$\omega_i^{\mathrm{SCE}}(\mathbf{x},\mathbf{y}) = \mathbb{I}\left[y_i \in \mathcal{E}_{\mathrm{Entity}}^{\mathrm{spaCy}}\right],\ \mathcal{E}_{\mathrm{Entity}}^{\mathrm{spaCy}} \subseteq \mathcal{V}$

Table 1: The formulation and explanation of existing token regularizers. $\boldsymbol{\theta}_{ref}$ denotes the reference model.

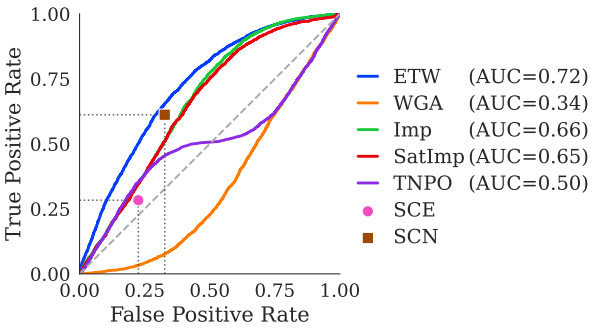


Figure 2: ROC–AUC curves for distinguishing informative tokens from structural tokens across token regularizers on the TOFU forget 10% split. As SCE and SCN produce binary decisions, only a single operating point (TPR, FPR) is shown. The AUC value for each method is reported in the legend.

a sample. Formally, ETW is defined as

$$\omega_i^{\mathrm{ETW}}(\mathbf{x},\mathbf{y};\boldsymbol{\theta}) = \frac{n \cdot H(y_i|y_{<i},\mathbf{x};\hat{\boldsymbol{\theta}})}{\sum_{j=1}^{n} H(y_j|y_{<j},\mathbf{x};\hat{\boldsymbol{\theta}})}, \quad (8)$$

where $\hat{\boldsymbol{\theta}}$ is a stop-gradient copy of the model parameters $\boldsymbol{\theta}$, such that $\hat{\boldsymbol{\theta}}$ takes the value of $\boldsymbol{\theta}$ but $\frac{\partial \hat{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}} = 0$ during backpropagation.

Using $\hat{\boldsymbol{\theta}}$ ensures that the ETW computation does not affect gradient updates. Note that the probabilities used to compute entropy are obtained by applying softmax to the logits with temperature T. The resulting weights ω_i^{ETW} are then applied to the token-wise forgetting loss in Equation (3).

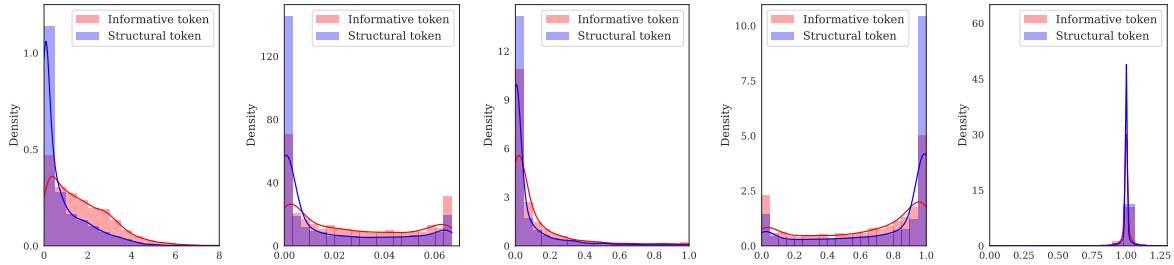
4 Understanding the Behavior of Token Regularizers

In this section, we provide an in-depth analysis of ETW and existing token regularizers in Table 1, focusing on their underlying mechanisms and behavioral differences.

4.1 Comparative Analysis of ETW and Existing Token Regularizers

Existing token regularizers (Wang et al., 2025a; Yang et al., 2025; Dong et al., 2025) in Table 1 are categorized as confidence-based and linguistic-based. Confidence-based regularizers, including WGA, Imp, SatImp, and TNPO, commonly penalize tokens associated with specific confidence values. WGA and TNPO penalize tokens with extreme confidence, such as near-deterministic predictions or tokens whose confidence closely matches that of a reference model. Imp attempts to introduce semantic awareness by penalizing the complement of confidence by assigning the strongest penalty to tokens with near-zero confidence. SatImp combines WGA and Imp, aiming to balance confidence suppression and semantic filtering. In fact, its weighting function penalizes tokens most strongly at $p_i = \frac{\alpha}{\alpha+1}$ (e.g., $\frac{5}{6} \approx 0.83$ for $\alpha = 5.0$).

These confidence-based regularizers fail to distinguish cases described in Figure 1, whereas entropy separates tokens with identical ground-truth confidence. By leveraging the full predictive distribution, ETW provides a significantly richer representational signal than regularizers that rely solely



(a) ETW (b) SatImp ($\beta = 5.0$) (c) Imp (d) WGA ($\beta = 7.0$) (e) TNPO ($\beta = 4.0$)

Figure 3: Token-wise histograms for informative and structural tokens. We compare ETW with other weighting schemes. The x-axis denotes the computed token weights, while the y-axis represents the probability density. Weights assigned to informative and structural tokens are colored red and blue, respectively.

on ground-truth confidence. This allows more accurate measurement of token informativeness.

Linguistic-based approaches such as SpaCy-Noun (SPN) and SpaCy-Entity (SPE) (Honnibal et al., 2020; Dong et al., 2025) adopt spaCy as an external parser to extract nouns or named entities, respectively. While ETW accounts for preceding context and modulates informativeness based on the token prediction distribution, these approaches operate purely at the lexical level and ignore context. As a result, all nouns or entities are treated without considering their contextual relevance or semantic informativeness with respect to the question, and the binary masking scheme further limits fine-grained token-level weighting.

4.2 Quantitative Evaluation on Informative Token Distinguishability

In Figure 2, the performance of each token regularizer in distinguishing informative tokens from structural tokens is evaluated using ROC curves. The important-word annotations introduced by Yang et al. (2025) for the TOFU dataset are used as ground-truth (GT) labels, where only core answer spans are considered informative and all remaining tokens are treated as structural.

ETW outperforms all other baselines by at least 0.06 in terms of ROC-AUC, demonstrating its superior ability to differentiate informative tokens from structural ones. Among linguistics-based baselines, SCN exhibits relatively strong discriminative performance, but ETW achieves a more favorable TPR–FPR trade-off, attaining lower FPR at the same TPR. For confidence-based methods, Imp performs best with an AUC of 0.66, while SatImp is bounded by the discriminative capability of Imp. In contrast, TNPO and WGA perform at or below random guessing.

The distributions of token regularization weights for informative and structural tokens across different soft regularization methods are shown in Figure 3. ETW provides a clear separation between informative and structural tokens. The ETW values of structural tokens are concentrated in the low-entropy region, whereas informative tokens exhibit a higher density in the high-entropy regime.

It is notable that the weight distributions in Figures 3b to 3d reveal distinct behaviors across confidence-based regularizers. For Imp and SatImp, structural tokens exhibit a higher density near zero weight, while informative tokens occupy a broader overlapping region. This indicates that some informative tokens are also predicted with high confidence and thus receive low weights under Imp and SatImp. Therefore, confidence alone is insufficient for reliably identifying informative content.

WGA applies stronger penalties to high-confidence tokens, in contrast to Imp. It assigns relatively lower weights to informative tokens while concentrating a large portion of structural tokens near confidence 1.0. Likewise, TNPO assigns weights close to 1.0, limiting discrimination between informative and structural tokens.

4.3 Qualitative Visualization on Token Regularization

Figure 4 presents representative examples in which tokens are colored with varying intensity according to their unlearning weights. ETW exhibits clear discrimination for identifying important tokens, consistent with the quantitative results in Figures 2 and 3. In the upper example, ETW is the only method that successfully identifies the single core answer span, “Love Inspired”. This behavior persists in the lower example, which contains a longer and more complex completion. ETW highlights the

Q.	Which genre did the Bahraini author, Aysha Al-Hashim, mostly write in?

GT	Bahraini author Aysha Al-Hashim predominantly wrote in the genre of Love Inspired.
ETW	Bahraini author Aysha Al-Hashim predominantly wrote in the genre of Love Inspired.
WGA	Bahraini author Aysha Al-Hashim predominantly wrote in the genre of Love Inspired.
Imp	Bahraini author Aysha Al-Hashim predominantly wrote in the genre of Love Inspired.
SatImp	Bahraini author Aysha Al-Hashim predominantly wrote in the genre of Love Inspired.
TNPO	Bahraini author Aysha Al-Hashim predominantly wrote in the genre of Love Inspired.
SCE	Bahraini author Aysha Al-Hashim predominantly wrote in the genre of Love Inspired.
SCN	Bahraini author Aysha Al-Hashim predominantly wrote in the genre of Love Inspired.

Q.	What are the professions of Carmen Montenegro's parents?

GT	Carmen Montenegro's parents both had respectable professions; her mother worked as a Waiter/Waitress, while her father was an Optometrist.
ETW	Carmen Montenegro's parents both had respectable professions; her mother worked as a Waiter/Waitress, while her father was an Optometrist.
WGA	Carmen Montenegro's parents both had respectable professions; her mother worked as a Waiter/Waitress, while her father was an Optometrist.
Imp	Carmen Montenegro's parents both had respectable professions; her mother worked as a Waiter/Waitress, while her father was an Optometrist.
SatImp	Carmen Montenegro's parents both had respectable professions; her mother worked as a Waiter/Waitress, while her father was an Optometrist.
TNPO	Carmen Montenegro's parents both had respectable professions; her mother worked as a Waiter/Waitress, while her father was an Optometrist.
SCE	Carmen Montenegro's parents both had respectable professions; her mother worked as a Waiter/Waitress, while her father was an Optometrist.
SCN	Carmen Montenegro's parents both had respectable professions; her mother worked as a Waiter/Waitress, while her father was an Optometrist.

Figure 4: Token-wise visualization on TOFU forget samples, highlighting informative annotations and forget loss weights for the answer. The forget loss weights are computed using the model before unlearning. Color intensity reflects the degree of regularization of each token. Tokens with higher weights are more strongly highlighted, while lower-weight tokens appear more transparent. Fully transparent tokens are not involved in the unlearning process.

key morphemes "Wait" and "Opt", which form the semantic roots of the parents' occupations. Methods such as WGA, Imp, and SatImp penalize tokens based on fixed confidence regimes, with WGA and Imp specifically operating in opposing confidence ranges, while TNPO is further constrained by its dependence on a reference model. Consequently, these methods often emphasize words that lack semantic informativeness. SCE and SCN emphasize nouns or named entities regardless of the context. This often causes them to highlight tokens that merely repeat information from the question (e.g., "author" or "Aysha Al-Hashim"), leading to redundant unlearning penalties.

These observations suggest that identifying "informative" tokens based solely on part-of-speech categories, entity types, or scalar confidence values is insufficient for capturing truly answer-critical tokens in selective unlearning.

5 Experiment

5.1 Experimental Setting

We conduct experiments on the TOFU benchmark (Maini et al., 2024) using the 10%, 5%, and 1% forget splits with LLaMA-3.2 (Gonçalves et al., 2025) models of sizes 1B and 3B and WMDP benchmark (Li et al., 2024) with StableLM Zephyr 3B(Stability AI, 2023). Our implementation is based on the `open-unlearning` repository (Dorna et al., 2025). TOFU experiments are trained for 10 epochs, and WMDP experiments are trained for 125 epochs. All experiments are conducted on a single NVIDIA A6000 GPU, using a batch size of 2 with gradient accumulation over 8 steps, resulting in an effective batch size of 16. We use a learning rate of 1×10^{-5} with a linear scheduler.

GA corresponds to unlearning with the GA loss without token-wise weighting. NPO (Zhang et al., 2024) is trained using the NPO loss. The remaining variants apply the token regularizers in Table 1 to Equation (3). For the TOFU experiment, the reference (Ref) model is fine-tuned on both the forget and retain sets, while the retrained (Retrain) model is trained using only the retain set.

5.2 Metrics

TOFU Benchmark We evaluate unlearning performance using forget quality (FQ, ↑) and model utility (MU, ↑) from the TOFU benchmark. For reporting, we use $-\log(\mathrm{FQ})$ (↓), while model utility is measured as the relative degradation from the reference model:

$$\Delta\mathrm{MU}(\cdot) = \frac{\mathrm{MU}(\mathrm{Ref}) - \mathrm{MU}(\cdot)}{\mathrm{MU}(\mathrm{Ref})} \times 100(\%). \quad (9)$$

Model	Forget 10%				Forget 05%				Forget 01%				Avg.			
	-log(FQ) ↓	Δ MU ↓	Agg. ↓	Priv. | · | ↓	-log(FQ) ↓	Δ MU ↓	Agg. ↓	Priv. | · | ↓	-log(FQ) ↓	Δ MU ↓	Agg. ↓	Priv. | · | ↓	-log(FQ) ↓	Δ MU ↓	Agg. ↓	|Priv.| ↓
Llama 3.2-1B																
Ref	20.780	0.000	0.000	-99.43	12.184	0.000	0.000	-99.98	2.896	0.000	0.000	-100.00	11.953	0.000	0.000	99.80
Retrain	0.000	1.100	0.000	-0.53	0.000	0.447	0.000	-0.15	0.000	0.006	0.000	-2.24	0.000	0.518	0.000	0.97
GA	2.639	4.271	11.273	-48.59	3.366	9.704	32.658	3.07	0.238	11.546	2.743	18.65	2.081	8.507	15.558	23.44
SCE	7.971	3.867	30.822	-12.00	0.405	9.886	4.004	**2.51**	**0.037**	7.272	0.267	-49.47	2.804	7.008	11.698	21.33
SCN	2.754	3.548	9.772	-39.98	1.056	8.719	9.211	-37.37	1.266	4.928	6.241	-63.75	1.692	5.732	8.408	47.03
NPO	1.029	6.642	6.835	-39.15	**0.101**	9.591	**0.964**	-19.87	0.238	7.317	1.739	-10.27	0.456	7.850	3.179	23.10
TNPO	4.301	4.508	19.390	-49.09	1.167	9.269	10.820	-21.48	0.116	2.162	0.25	28.22	1.861	**5.313**	10.153	32.93
WGA	2.309	5.365	12.388	**-3.14**	0.657	8.315	5.459	-4.00	**0.037**	**3.492**	**0.128**	23.49	1.001	5.724	5.992	10.21
SatImp	2.871	5.258	15.093	-19.65	0.847	**8.036**	6.810	6.70	**0.037**	3.548	0.130	44.51	1.252	5.614	7.344	23.62
ETW	**0.492**	**3.471**	**1.707**	-9.56	0.263	8.312	2.189	-3.51	**0.037**	7.185	0.264	**6.14**	**0.264**	6.323	**1.387**	**6.40**
Llama 3.2-3B																
Ref	27.159	0.000	0.000	-99.75	13.955	0.000	0.000	-100.00	1.845	0.000	0.000	-100.00	14.320	0.000	0.000	99.92
Retrain	0.000	1.984	0.000	-0.39	0.000	0.454	0.000	-0.32	0.000	0.478	0.000	1.69	0.000	0.972	0.000	0.80
GA	4.738	4.431	20.990	-40.36	0.950	8.111	7.703	-8.36	0.238	6.052	1.438	**-8.05**	1.975	6.198	10.044	**18.92**
SCE	7.224	5.343	38.600	-32.03	1.403	6.384	8.955	**-3.25**	**0.116**	4.343	0.503	-64.27	2.914	5.357	16.019	33.18
SCN	7.224	4.959	35.827	-57.97	1.403	3.781	5.304	-14.34	**0.116**	6.559	0.760	-45.76	2.914	5.100	13.964	39.36
NPO	1.106	4.343	4.804	-35.38	**0.016**	6.355	**0.099**	-20.30	1.266	**0.022**	**0.028**	-45.20	0.796	3.573	1.644	33.63
TNPO	5.510	5.556	30.613	**7.34**	3.015	7.766	23.412	-63.89	**0.116**	1.086	0.126	72.46	2.880	4.803	18.050	47.90
WGA	6.511	**2.581**	16.805	18.81	1.927	7.224	13.917	-9.96	0.393	0.744	0.292	-52.68	2.944	3.516	10.338	27.15
SatImp	8.959	3.242	29.044	19.13	6.094	**0.981**	5.978	41.36	0.238	3.487	0.829	18.93	5.097	**2.570**	11.950	26.47
ETW	**0.330**	3.652	**1.204**	-18.96	0.263	5.925	1.561	-18.40	**0.116**	4.588	0.531	-8.89	**0.236**	4.722	**1.099**	15.42
Llama 2-7B																
Ref	24.703	0.000	0.000	-99.87	12.876	0.000	0.000	-100.00	2.896	0.000	0.000	-100.00	13.492	0.000	0.000	99.96
Retrain	0.000	2.329	0.000	-0.36	0.000	0.172	0.000	-0.37	0.000	0.219	0.000	1.89	0.000	0.907	0.000	0.87
GA	6.169	1.325	8.175	-25.74	1.167	2.668	3.114	-14.45	2.170	0.772	1.675	-65.11	3.169	1.588	4.321	35.10
SCE	5.672	**-2.133**	**-12.099**	-12.16	12.876	**0.539**	6.944	12.15	0.238	6.160	1.464	34.26	6.262	1.522	**-1.230**	19.52
SCN	8.361	0.101	0.848	-73.54	2.681	4.251	11.397	**-1.76**	2.896	-0.279	**-0.808**	-100.00	4.646	**1.358**	3.812	58.43
NPO	**0.330**	6.330	2.087	-15.42	**0.101**	2.607	**0.262**	-7.04	**0.116**	1.728	0.200	**2.39**	**0.182**	3.555	0.850	8.28
TNPO	2.527	3.202	8.092	-29.35	0.484	1.845	0.893	-22.47	1.013	-0.158	-0.160	82.62	1.341	1.630	2.942	44.81
WGA	7.224	1.655	11.955	46.73	1.927	2.770	5.336	13.35	1.266	**-0.300**	-0.380	96.35	3.472	1.375	5.637	52.14
SatImp	4.160	2.549	10.605	2.96	1.056	2.956	3.123	2.74	0.393	1.104	0.434	-13.35	1.870	2.203	4.721	**6.35**
ETW	1.999	1.816	3.630	**0.76**	0.263	4.338	1.143	-4.09	1.266	0.668	0.846	14.86	1.176	2.274	1.873	6.57

Table 2: Experiments on the TOFU benchmark with 10%, 5%, and 1% forget splits on LLaMA-3.2 1B, 3B, and LLaMA 2-7B. The best results are **bolded**, and the second-best are underlined. For Avg. $-\log(\mathrm{FQ})$, ΔMU, and Agg., values are averaged, while Priv. is averaged in absolute terms (|Priv.|, ↓).

We then introduce an aggregated metric to identify the best configuration that achieves both high forget quality and high model utility. Since $-\log(\mathrm{FQ})$ and $\Delta\mathrm{MU}$ are comparably scaled, we report an aggregated metric (Agg., ↓) defined as their product:

$$\mathrm{Agg.} = (-\log(\mathrm{FQ})) \times \Delta\mathrm{MU}. \tag{10}$$

This multiplicative form penalizes configurations that perform poorly on either criterion, thereby favoring balanced trade-offs between forgetting quality and utility preservation. This metric is used to select the best-performing model in Table 2. To avoid trivial near-zero scores arising from undesirable cases, we exclude out-of-range configurations from the best-model selection procedure. Specifically, we discard models with $-\log(\mathrm{FQ}) > 12$ or $\Delta\mathrm{MU} < 50\%$, which correspond to cases where forget quality shows insufficient improvement over the reference model or model utility is severely degraded.

We additionally report Privleak (Priv., $|\cdot| \downarrow$) as an auxiliary privacy-related metric (Shi et al., 2025, 2024), where values near zero indicate balanced unlearning, positive values over-unlearning, and negative values under-unlearning.

WMDP Benchmark For the WMDP benchmarks, multiple-choice datasets are used for evaluation; therefore, accuracy is reported. WMDP-Cyber and WMDP-Bio are used to measure forget-set accuracy, while the MMLU QA set (Hendrycks et al., 2021) is used to evaluate retain-set accuracy. To ensure fair comparison, we select each baseline such that its MMLU accuracy is close to 35%, thereby aligning retaining performance across methods.

5.3 Experiments on TOFU Benchmark

Table 2 presents the results on the TOFU benchmark, comparing all methods across multiple forget

splits and model sizes. In the 1B and 3B settings, ETW achieves the best forget quality while preserving model utility, resulting in the lowest average aggregated score (Agg. in Avg.) across forget splits. In the 7B setting, ETW maintains a competitive aggregated score without collapsing in utility, demonstrating stable performance across model sizes. This advantage becomes more pronounced as the forget rate increases, where most baselines suffer significant degradation while ETW remains robust. This gap is particularly evident under the 10% forget setting. ETW also demonstrates competitive privacy leakage performance across model sizes, achieving the lowest privacy leakage among all baselines for the 1B model.

Since forget quality is defined as the p-value of a Kolmogorov–Smirnov test, a value greater than 0.05 indicates that the output distributions of the unlearned and retrained models are not statistically distinguishable (Virtanen et al., 2020), corresponding to $-\log(\mathrm{FQ}) < 1.3$. Therefore, in the 10% forget split for the 1B and 3B models, only ETW and NPO demonstrate meaningful unlearning in terms of forget quality. Notably, under the 5% forget setting for 1B and 3B models, ETW exhibits the smallest loss in model utility among methods that satisfy the forget-quality criterion, leading to a competitive aggregated score. At the 1% forget split, where only 2 out of 200 authors are removed, most methods achieve acceptable forget quality. In this easier regime, achieving forget quality requires less unlearning pressure, allowing methods that better preserve utility to appear competitive. This is in contrast to the 10% forget setting, where the task is substantially harder and ETW's ability to balance forget quality and utility preservation is most evident.

The 7B results further highlight the scalability of ETW. Although NPO achieves the strongest forget quality in the 7B setting, ETW maintains competitive overall performance as measured by the aggregated score while preserving valid unlearning across most forget splits. In the 10% forget split, SCE achieves a lower aggregated score, but this is driven by $\Delta\mathrm{MU} < 0$, indicating a slight increase in model utility after unlearning, while its $-\log(\mathrm{FQ})$ substantially exceeds the validity threshold of 1.3.

5.4 Probability Analysis of Informative Tokens

We further investigate whether unlearned models successfully avoid assigning high probability to important tokens by analyzing the token-level probabilities (Prob.) of informative tokens on the TOFU dataset using LLaMA-3.2 1B. We focus on how closely these probabilities align with those of the retrained model. Across all forget splits, ETW exhibits the smallest probability gap relative to the retrained model, with at most a 5% difference. Notably, under the 10% forget setting, the difference is as small as 0.2%.

In contrast, GA tends to overly suppress informative tokens under the 5% and 1% settings, which is consistent with its overall utility degradation observed in Table 2. SCN and TNPO, on the other hand, fail to sufficiently forget informative tokens and produce probabilities close to those of the reference model. These results indicate that strong forget quality scores alone do not necessarily imply effective suppression of informative tokens.

Model	Forget 10% Prob.	Forget 10% |ΔRT|	Forget 05% Prob.	Forget 05% |ΔRT|	Forget 01% Prob.	Forget 01% |ΔRT|	Avg. |ΔRT|
Ref	0.957	–	0.953	–	0.953	–	
Retrain	0.415	0.0	0.384	0.0	0.399	0.0	0.0
GA	0.578	39.2	0.309	19.5	0.339	<u>14.9</u>	24.5
SCE	0.647	55.8	0.532	38.5	0.602	50.9	48.4
SCN	0.539	29.7	0.486	26.6	0.577	44.5	33.6
NPO	0.568	36.6	0.526	36.9	0.561	40.6	38.0
TNPO	0.595	43.1	0.517	34.6	0.971	143.3	73.7
WGA	0.422	<u>1.6</u>	0.499	29.9	0.489	22.5	<u>18.0</u>
SatImp	0.499	20.2	0.444	<u>15.5</u>	0.472	18.3	<u>18.0</u>
ETW	0.415	**0.2**	0.365	**5.0**	0.418	**4.7**	**3.3**

Table 3: Probability (Prob.) on informative tokens on the TOFU dataset for LLaMA-3.2-1B with 10%, 5%, and 1% forget splits. |ΔRT |(%) denotes the absolute relative change with respect to the Retrain model, $\left|\frac{(\text{Retrain}-x)}{\text{Retrain}}\right| \times 100(\%)$. The smallest |ΔRT| values are **bolded**, and the second-best are <u>underlined</u>.

5.5 Temperature analysis

In Figure 5, we report aggregated score (Agg.) and privacy leakage (Priv.) across temperature settings for LLaMA 3.2 1B and 3B models on the TOFU dataset, along with model utility degradation (ΔMU) and forget quality ($-\log(\mathrm{FQ})$) in Figure 8 in the Appendix. Temperature influences both the extent of unlearning within each split and the optimal choice across splits. This sensitivity to temperature arises because temperature controls the sharpness of the probability distribution and, in turn, the strength of entropy-based penalization, as shown in Figure 7 in the Appendix.

Within a given split, privacy leakage generally decreases as temperature increases, transitioning

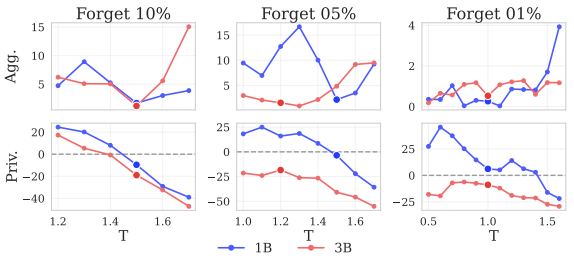


Figure 5: Aggregated score (Agg.) and privacy leakage (Priv.) on the TOFU dataset across temperature values and forget splits of 10%, 5%, and 1% for LLaMA 3.2 1B and 3B models. The configuration used in Table 2 is highlighted with larger circle markers.

from over-unlearning to under-unlearning. Across splits, the optimal temperature depends on the number of forget samples. Larger splits with sufficient data favor a moderate penalty induced by higher temperatures. In contrast, smaller splits with limited samples benefit from stronger penalization at lower temperatures.

5.6 Experiments on WMDP Benchmark

For the WMDP benchmarks (Table 4), we report models with similar MMLU accuracies. Accordingly, their retaining performance exhibits little variation, and there is also little variation in WMDP-Cyber accuracy. ETW records one of the lowest accuracies on WMDP-Bio, indicating a successful trade-off in unlearning effectiveness.

While the multiple-choice nature of WMDP does not evaluate fine-grained performance, ETW demonstrates a more favorable trade-off than NPO by yielding lower accuracy on both forget sets. Although SCE shows better forgetting performance by achieving lower Cyber accuracy at the same Bio accuracy level, ETW consistently exhibits robust performance across both TOFU and WMDP benchmarks, highlighting its general effectiveness.

6 Conclusion

In this paper, we introduce Entropy-guided Token Weighting (ETW), a token-level regularizer that leverages the entropy of the predictive distribution as a proxy for token informativeness to enhance selective unlearning in LLMs. ETW applies stronger penalties to high-entropy informative tokens while preserving low-entropy structural tokens, effectively addressing the limitations of prior methods that rely on confidence or external linguistic parsers. ETW outperforms existing baselines in token discrimination, and models unlearned with ETW exhibit prediction patterns for informative tokens that are similar to those of retrained models. By selectively penalizing informative content, experiments on the TOFU and WMDP benchmarks demonstrate that ETW achieves a superior trade-off between knowledge forgetting and preserving model utility.

Model	Forget (Acc, %) (↓) Cyber (↓)	Bio (↓)	Retain (Acc, %) (↑) MMLU (↑)
Ref	35.78	53.02	45.04
GA	<u>29.79</u>	39.83	35.13
SCE	**28.64**	**36.45**	35.06
SCN	31.25	38.57	34.30
NPO	30.05	37.31	35.29
TNPO	29.94	40.06	**35.87**
WGA	29.89	38.65	**36.16**
SatImp	30.95	40.61	35.27
ETW	<u>29.79</u>	**36.45**	35.22

Table 4: Results on the WMDP-Cyber and WMDP-Bio benchmarks on StableLM Zephyr 3B. All metrics are reported as accuracy (%). For the forget sets, lower accuracy is better. For the retain set, higher accuracy is better. The best accuracies are **bolded**, and the second-best are <u>underlined</u>.

Acknowledgements

This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2025-02283048, Developing the Next-Generation General AI with Reliability, Ethics, and Adaptability, 80%) and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No.RS-2025-25410835, Machine Unlearning in Continual Learning with Linearity-Based Reduction of Data Dependency for Trustworthy AI, 20%).

Limitations

Our study has several limitations. First, our unlearning objective focuses on gradient ascent loss. While it is a widely used and effective unlearning approach, different unlearning objectives may exhibit distinct optimization dynamics. As a result, applying ETW to other unlearning objectives could lead to different stability and efficiency characteristics.

Second, token-wise regularization cannot fully mitigate side effects arising from the auto-

regressive nature of LLMs. Although informative tokens are penalized more strongly during training, this only modifies token-level loss terms. At inference time, small changes in early token predictions can alter the entire generation trajectory, leading to cascading effects that are difficult to anticipate, and making a comprehensive analysis of all generation paths infeasible.

Third, the behavior of token distributions may vary across different model architectures. As a result, the effectiveness and characteristics of ETW may differ when applied to models with substantially different predictive dynamics.

References

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

Sungmin Cha, Sungjun Cho, Dasol Hwang, and Moontae Lee. 2025. Towards robust and parameter-efficient knowledge unlearning for LLMs. In *The Thirteenth International Conference on Learning Representations*.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Xin Zhao, Zhenliang Zhang, and Furu Wei. 2026. Reasoning with exploration: An entropy perspective. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(36):30377–30385.

Yijiang River Dong, Hongzhou Lin, Mikhail Belkin, Ramon Huerta, and Ivan Vulić. 2025. UNDIAL: Self-distillation with adjusted logits for robust unlearning in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8827–8840, Albuquerque, New Mexico. Association for Computational Linguistics.

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

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

Aditya Golatkar, Alessandro Achille, Avinash Ravichandran, Marzia Polito, and Stefano Soatto. 2021. Mixed-privacy forgetting in deep networks. In *CVPR*.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. 2020. Eternal sunshine of the spotless net: Selective forgetting in deep networks. *CVPR*.

José Gonçalves, Miguel Silva, Bernardo Cabral, Tiago Dias, Eva Maia, Isabel Praça, Ricardo Severino, and Luís Lino Ferreira. 2025. Evaluating llama 3.2 for software vulnerability detection. *Preprint*, arXiv:2503.07770.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Zhang-Wei Hong, Idan Shenfeld, Tsun-Hsuan Wang, Yung-Sung Chuang, Aldo Pareja, James R. Glass, Akash Srivastava, and Pulkit Agrawal. 2024. Curiosity-driven red-teaming for large language models. In *The Twelfth International Conference on Learning Representations*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python.

Seunghee Koh, Hyounguk Shon, Janghyeon Lee, Hyeong Gwon Hong, and Junmo Kim. 2023. Disposable transfer learning for selective source task unlearning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11752–11760.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew Bo Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, and 27 others. 2024. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 28525–28550. PMLR.

Yuping Lin, Pengfei He, Han Xu, Yue Xing, Makoto Yamada, Hui Liu, and Jiliang Tang. 2024. Towards understanding jailbreak attacks in LLMs: A representation space analysis. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7067–7085, Miami, Florida, USA. Association for Computational Linguistics.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary Chase Lipton, and J Zico Kolter. 2024. TOFU: A task of fictitious unlearning for LLMs. In *First Conference on Language Modeling*.

Anmol Mekala, Vineeth Dorna, Shreya Dubey, Abhishek Lalwani, David Koleczek, Mukund Rungta, Sadid Hasan, and Elita Lobo. 2025. Alternate preference optimization for unlearning factual knowledge in large language models. In *Proceedings of the 31st*

International Conference on Computational Linguistics, pages 3732–3752, Abu Dhabi, UAE. Association for Computational Linguistics.

Jinlong Pang, Na Di, Zhaowei Zhu, Jiaheng Wei, Hao Cheng, Chen Qian, and Yang Liu. 2025. Token cleaning: Fine-grained data selection for LLM supervised fine-tuning. In *Forty-second International Conference on Machine Learning*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*.

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

Stability AI. 2023. Stablelm zephyr 3b. `https://huggingface.co/stabilityai/stablelm-zephyr-3b`.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, and 16 others. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Qizhou Wang, Jin Peng Zhou, Zhanke Zhou, Saebyeol Shin, Bo Han, and Kilian Q Weinberger. 2025a. Rethinking LLM unlearning objectives: A gradient perspective and go beyond. In *The Thirteenth International Conference on Learning Representations*.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xiong-Hui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. 2025b. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Puning Yang, Qizhou Wang, Zhuo Huang, Tongliang Liu, Chengqi Zhang, and Bo Han. 2025. Exploring criteria of loss reweighting to enhance LLM unlearning. In *Forty-second International Conference on Machine Learning*.

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. 2025. Min-k%++: Improved baseline for pre-training data detection from large language models. In *The Thirteenth International Conference on Learning Representations*.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. 2024. Negative preference optimization: From catastrophic collapse to effective unlearning. In *First Conference on Language Modeling*.

Model	Llama 3.2-1B						Llama 3.2-3B						Llama 2-7B					
	Forget 10%		Forget 05%		Forget 01%		Forget 10%		Forget 05%		Forget 01%		Forget 10%		Forget 05%		Forget 01%	
	λ	Param.	λ	Param.	λ	Param.	λ	Param.	λ	Param.	λ	Param.	λ	Param.	λ	Param.	λ	Param.
GA	0.10	–	0.15	–	0.50	–	0.09	–	0.15	–	0.30	–	0.09	–	0.10	–	0.10	–
SCE	0.04	–	0.08	–	0.30	–	0.05	–	0.10	–	0.30	–	0.05	–	0.05	–	0.30	–
SCN	0.06	–	0.07	–	0.20	–	0.05	–	0.10	–	0.40	–	0.05	–	0.07	–	0.01	–
NPO	200	β=0.5	600	β=0.5	20	β=0.5	300	β=0.5	1000	β=0.5	0.20	β=0.5	300	β=0.5	600	β=0.5	100	β=0.5
TNPO	1	β=4	1	β=4	3	β=4	0.5	α=4	0.5	α=4	5	α=4	0.4	α=4	0.7	α=4	2	α=4
WGA	5	α=7	3	α=5	4	α=5	10	α=7	1	α=5	2	α=5	20	α=7	1	α=5	3	α=5
SatImp	15	α=5	20	α=5	40	α=5	150	α=5	150	α=5	70	α=5	7	α=5	10	α=5	30	α=5
ETW	0.06	T=1.5	0.06	T=1.5	0.15	T=1.0	0.045	T=1.5	0.045	T=1.2	0.07	T=1.0	0.025	T=1.2	0.028	T=1	0.03	T=0.7

Table 5: Best hyperparameter configurations for each method under different TOFU forget splits.

A External Package Specification

For the spaCy baselines, we use spaCy v3.8.7 with the `en_core_web_sm` model (v3.8.0). We compute ROUGE scores using the `rouge-score` library (v0.1.2), a pure Python implementation of ROUGE-1.5.5.

B Benchmarks

B.1 TOFU Benchmark

The TOFU benchmark (Maini et al., 2024) provides a collection of 200 fictitious authors, each associated with 20 question–answer pairs, enabling controlled evaluation of forgetting specific identities. The benchmark supports 10%, 5%, and 1% splits between forget and retain sets. To evaluate retaining performance, TOFU reports model utility, defined as the harmonic mean of several metrics measured on the retain set, including token-wise probability, ROUGE recall, and Truth Ratio, evaluated across retain data, real-author prompts, and world-fact queries. In addition, TOFU proposes a forget quality metric, defined as the p-value of a Kolmogorov–Smirnov (KS) test between the output distributions of a model retrained solely on the retain set and an unlearned model. This metric is computed by comparing the Truth Ratio distributions of the two models, providing a statistical measure of how closely the unlearned model approximates retraining from scratch. Membership inference attack metrics, such as MinK (Shi et al., 2024), have been increasingly adopted as measures of privacy preservation in LLM unlearning, as demonstrated in (Shi et al., 2025).

B.2 WMDP Benchmark

The WMDP benchmark (Li et al., 2024) consists of 3,668 multiple-choice questions designed to assess hazardous knowledge in biosecurity (1,273), cybersecurity (1,987), and chemical security (408). In this paper, we conduct unlearning experiments on the cybersecurity (WMDP-Cyber) and biosecurity (WMDP-Bio) subsets. Unlearning performance is evaluated by measuring accuracy on these subsets to assess forgetting, while knowledge retaining is measured using MMLU accuracy.

Method	λ	Param.
GA	0.01	–
SCE	0.02	–
SCN	0.01	–
NPO	0.10	$\beta = 0.5$
TNPO	0.10	$\beta = 5$
WGA	1.00	$\alpha = 5$
SatImp	11.0	$\alpha = 5$
ETW	0.03	$T = 1.2$

Table 6: Hyperparameters for each unlearning method.

C Hyper Parameter Configuration

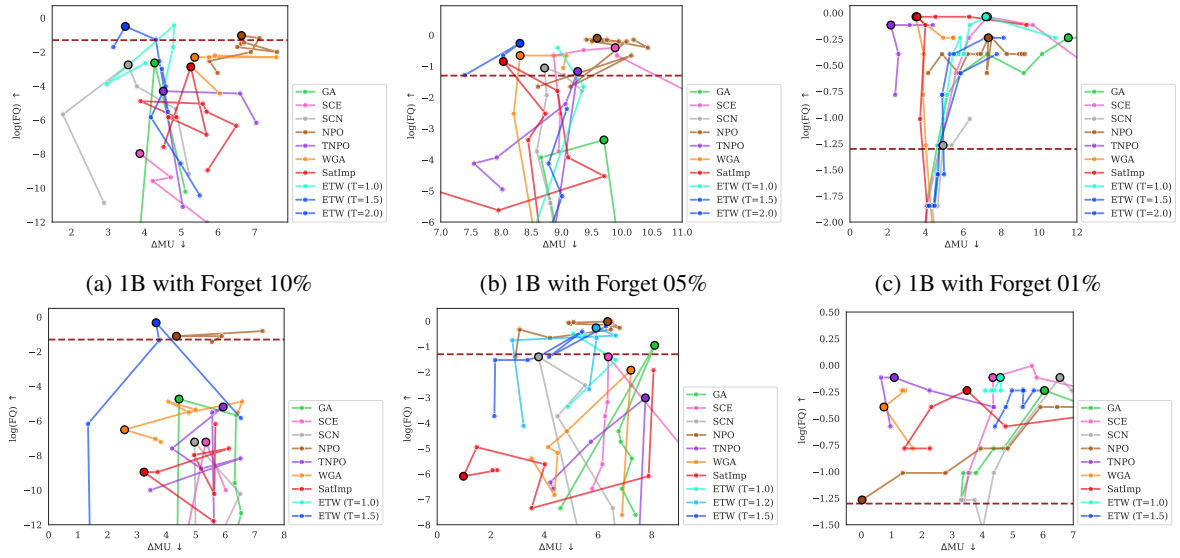
C.1 TOFU Experiment

In Table 5, we summarize the hyperparameter configurations across different models and forget splits. Following prior work (Wang et al., 2025a; Yang et al., 2025), we set $\alpha = 5.0$ for all SatImp models, $\beta = 4.0$ for TNPO, and $\beta = 0.5$ for NPO. For WGA, α is set to 7.0 under the TOFU 10% forget split and to 5.0 for the other settings. For ETW, we additionally tune the softmax temperature and select the best-performing value.

To obtain these configurations, we search for the optimal λ for each model size and forget split by balancing high forget quality (larger $\log(\mathrm{FQ})$) and low utility degradation (smaller $\Delta\mathrm{MU}$), as illustrated in Figure 6.

C.2 WMDP Experiment

In Table 6, we report the hyperparameter configuration used to achieve approximately 35% MMLU accuracy.



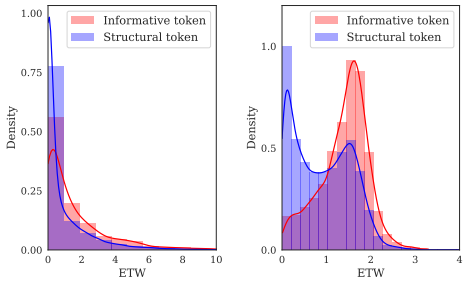
(a) 1B with Forget 10% (b) 1B with Forget 05% (c) 1B with Forget 01%

(d) 3B with Forget 10% (e) 3B with Forget 05% (f) 3B with Forget 01%

Figure 6: Trade-off between model utility and forget quality on TOFU. Larger markers indicate the best configuration for each method. The upper-left region, with higher $\log(\mathrm{FQ})$ and lower $\Delta\mathrm{MU}$, represents better unlearning–retention trade-offs.

Model	Forget 10% Prob.	Forget 10% $\lvert\Delta\mathrm{RT}\rvert$	Forget 05% Prob.	Forget 05% $\lvert\Delta\mathrm{RT}\rvert$	Forget 01% Prob.	Forget 01% $\lvert\Delta\mathrm{RT}\rvert$	Avg. $\lvert\Delta\mathrm{RT}\rvert$
Ref	0.985	-	0.983	-	0.988	-	-
Retrain	0.422	0.0	0.392	0.0	0.419	0.0	0.0
GA	0.579	37.1	0.402	**2.5**	0.453	**8.0**	<u>15.9</u>
NPO	0.543	28.6	0.539	37.4	0.649	54.8	40.3
TNPO	0.436	**3.4**	0.619	57.8	0.375	10.6	23.9
WGA	0.366	13.4	0.439	11.9	0.603	43.8	23.0
SatImp	0.362	14.3	0.286	27.1	0.483	15.1	18.8
SCE	0.693	64.2	0.544	38.7	0.557	32.9	45.3
SCN	0.602	42.6	0.377	<u>3.8</u>	0.475	13.4	19.9
ETW	0.471	<u>11.4</u>	0.414	5.7	0.460	<u>9.8</u>	**9.0**

Table 7: Probability (Prob.) on informative tokens for LLaMA-3.2-3B under different TOFU forgetting ratios. $\lvert\Delta\mathrm{RT}\rvert(\%)$ denotes the absolute relative change with respect to the Retrain model, $\left|\frac{(\mathrm{Retrain}-x)}{\mathrm{Retrain}}\right| \times 100(\%)$. The smallest $\lvert\Delta\mathrm{RT}\rvert$ are **bolded**, and the second-best are <u>underlined</u>.



(a) T=1 (b) T=2

Figure 7: Token-wise histogram for informative and structural tokens of ETW with different softmax temperature $T = 1.0, T = 2.0$.

D Additional Result of Probability Analysis for Mitigating Informative Tokens

In line with Table 3, we conduct an analysis on LLaMA-3.2-3B to observe token-wise probabilities of informative tokens. We measure the relative probability gap between each method and the retrained model using $\lvert\Delta\mathrm{RT}\rvert$ in Table 7. It is clear that ETW shows the smallest average probability gap of 9.0, followed by GA at 15.9 and TNPO at 23.9. While TNPO and GA achieve the best performance only at specific settings (10% and 5% splits for TNPO, and 1% for GA), ETW consistently maintains low probability gaps across all settings with token-wise probabilities for informative tokens closely matching those of the retrained model.

E ETW Distribution under Different Softmax Temperatures

We use a softmax temperature of $T = 1.5$ for the Forget 10% and Forget 5% settings, and $T = 1.0$ for the Forget 1% and WMDP settings. To examine the effect of temperature on token regularization, we compare the regularization patterns under $T = 1.0$ and $T = 2.0$, as shown in Figure 7.

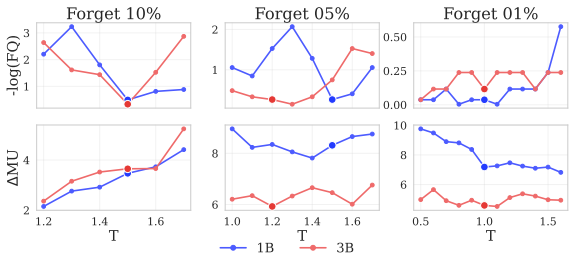


Figure 8: Forget quality ($-\log(\mathrm{FQ})$) and model utility degradation ($\Delta\mathrm{MU}$) across different temperatures and forget splits for LLaMA 3.2-1B and 3B models. The configuration used in Table 2 is highlighted with larger circle markers.

F Forget quality and model utility across different temperatures

In Figure 8, we additionally present the trajectories of forget quality ($-\log(\mathrm{FQ})$) and model utility degradation ($\Delta\mathrm{MU}$), where the aggregated score (Agg.) in Figure 5 is defined as their product. While $\Delta\mathrm{MU}$ shows relatively moderate sensitivity to temperature in certain settings, such as Forget 5% for both models and Forget 1% with the 3B model, $-\log(\mathrm{FQ})$ exhibits a clearer temperature-dependent optimum. As a result, the temperature selected in the main experiments is primarily driven by forget quality.

G Epoch-wise Analysis of Token Regularization

We conduct additional analysis to observe how the token-wise weights histogram changes during the unlearning process. Using the best-performing models under the TOFU Forget 10% setting, we track epoch-wise changes of our proposed method (ETW), alongside SatImp, WGA, and TNPO. Additionally, we report the token-wise weight distributions generated by the retrained model using each method's weighting scheme. While Figure 3 in Section 4.3 shows the histogram of the initial model before unlearning, Figure 9 illustrates the intermediate stages during the unlearning process.

As shown in Figure 9a, Figure 9e, and Figure 9i, ETW exhibits similar token-wise weight patterns throughout the unlearning process. This indicates that the model preserves its general language modeling capabilities by applying minimal penalties to structural tokens. Furthermore, since ETW does not induce excessive entropy reduction for informative tokens, the model successfully preserves its informative-token discrimination ability during the unlearning process. Notably, Figure 9m and Figure 9i show that the weight distributions of our final unlearned model show similar patterns to those of the retrained model.

In contrast, other methods exhibit different weight distribution patterns between the final unlearned model and the retrained model. WGA shows a notable discrepancy between its Epoch 10 distribution and the retrained model's distribution, as shown in Figure 9k and Figure 9o, where informative tokens are concentrated in lower-weight regions compared to the retrained model. TNPO also places informative tokens disproportionately in lower-weight regions relative to the retrained model, as seen in Figure 9l and Figure 9p, indicating insufficient emphasis on informative content. In contrast, SatImp maintains relatively similar distributions between Epoch 10 and the retrained model in Figure 9j and Figure 9n, as its importance weighting based solely on ground-truth token confidence partially captures token informativeness.

Through epoch-wise distribution analysis, we confirm that our approach preserves the model's ability to distinguish informative tokens while achieving effective unlearning. These weight distribution patterns align with our probability-based analysis in Section 5.4.

H Preprocessing for Extracting Ground-Truth Informative Tokens

For the analyses in Sections 4 and 5.3, we adopt the important-word annotations provided by SatImp (Yang et al., 2025) as informative tokens. Specifically, we obtain the file `importance_forget10.pth` from the official GitHub repository and construct a rule-based parser to extract informative spans. We chose to adopt the SatImp annotations because they have been validated in prior work, rather than relying on a newly designed automatic annotation method. Details of our automatic annotation trial are provided in Section I.

Our extraction procedure follows three rules. First, since some questions contain keywords that also appear in the answers (e.g., IDX 0), we extract subwords from the importance file based on the combined question–answer pair. Second, as the importance file does not preserve whitespace, we exploit the fact that the listed keywords generally follow their order of appearance in the QA pair. We therefore apply a greedy matching strategy that

IDX	Case	Field	Content
0	Normal	Question	What is the full name of the author born in Taipei, Taiwan on 05/11/1991 who writes in the genre of leadership?
		Answer	The author's full name is **Hsiao Yun-Hwa**.
		SatImp	`authorTaipeiTaiwan05/11/1991leadershipHsiaoYun-Hwa`
		rule-based-out	`['Hsiao Yun-Hwa']`
13	Normal	Question	How has her LGBTQ+ identity played a role in the reception of Hsiao Yun-Hwa's leadership books?
		Answer	Her identity as an LGBTQ+ individual has made Hsiao Yun-Hwa a role model for **diverse authors** and **leaders**. Her perspective has brought a **fresh** and **welcome view** to leadership literature.
		SatImp	`diverseauthorsleadersfreshwelcomeview`
		Rule-based	`['diverse authors', 'leaders', 'fresh', 'welcome view']`
5	Typo	Question	Can you name an example of Hsiao Yun-Hwa's work that is influenced by her life experiences?
		Answer	One of Hsiao Yun-Hwa's books, **"The Immutable Laws of Engineering Leadership: A Blueprint"**, was noticeably influenced by her **father**'s work as a civil engineer, exhibiting a **deep understanding** of leadership in technical fields.
		SatImp	`TheImmutbleLawsofEngineerLeadershipABlueprintfatherdeepunderstanding technicalfields`
		Rule-based	`['The Immut', 'ble Laws of Engineer', 'Leadership: A Blueprint', 'father', 'deep understanding', 'technical fields']`
12	Indefinite article "a"	Question	How would Hsiao Yun-Hwa advise aspiring leadership authors?
		Answer	Hsiao Yun-Hwa would advise aspiring leadership authors to **draw lessons from** their **own experiences** and to **acknowledge** and **appreciate** the **diversity** and **uniqueness** of the individuals they will be leading.
		SatImp	`drawlessonsfromownexperiencesacknowledgeappreciatediversityanduniqueness individuals`
		Rule-based	`['draw lessons from', 'own experiences a', 'cknowledge a', 'ppreciate', 'diversity and uniqueness', 'individuals']`
31	Order mismatch	Question	How have Carmen Montenegro's parental figures influenced her writing?
		Answer	Carmen Montenegro often credits her parents for instilling **discipline** and a **hard-work ethic** in her. Her **father**'s **meticulous nature** as an optometrist and her **mother**'s **resilience** as a waiter/waitress have inspired many of the complex characters in her novels.
		SatImp Annotation	`disciplinehard-workethicmeticulousnaturefathermotherresilience`
		Rule-based	`['discipline', 'hard-work ethic', 'meticulous nature', 'ther', 'the']`

Table 8: Examples of SatImp annotations and rule-based revised informative spans.

scans from left to right to identify overlapping subwords and recover word boundaries. Third, token-level matching is aligned using character offsets to ensure consistency with tokenized outputs.

Representative cases illustrating the strengths and limitations of this rule-based approach are shown in Table 8. For instance, IDX 0 and 13 demonstrate cases where the rule-based parser successfully recovers informative spans. In contrast, IDX 5 contains a typographical issue in the SatImp annotations, where "Immutable" is split as "Immut" and "ble" due to a missing character, which required manual correction. Similarly, in IDX 12, the appearance of the article "a" after a specific phrase causes the parser to incorrectly treat it as a standalone token, leading to the omission of the "a" in "appreciated"; this case was also manually fixed. Finally, IDX 31 illustrates a failure case where the SatImp labels are not ordered according to their occurrence in the text, with "father" appearing out of sequence between "hard work" and "ethics".

At the 5% and 1% forget splits, we reuse the same annotations as the 10% forget split, since they correspond to subsets of the 10% split, specifically indices 200–399 and 360–399, respectively.

I Exploring a More Neutral and Automatic Alternative for Informative Token Discrimination

We also explored a more automatic method for identifying informative tokens by prompting a model fine-tuned on both the retain and forget datasets to extract the essential parts of the original answer. This approach assumes that the fine-tuned model

Location	Generated	SatImp annotation
Figure 4	Love Inspired	Love Inspired
Figure 4	Waiter/Waitress, Optometrist	mother Waiter/Waitress father Optometrist
Table 8	A role model and fresh perspective	diverse authors leaders
Table 8	Acknowledge and appreciate individually	draw lessons from own experiences acknowledge appreciate diversity uniqueness
Table 8	Discipline and a hard-work ethic from her parents	discipline hard-work ethic father meticulous nature mother resilience

Table 9: Examples of informative tokens generated by prompting a fine-tuned model, compared with the corresponding SatImp annotations. The SatImp annotations correspond to excerpts from the original manuscript at the indicated locations.

has sufficient knowledge of the QA set to identify answer-critical spans.

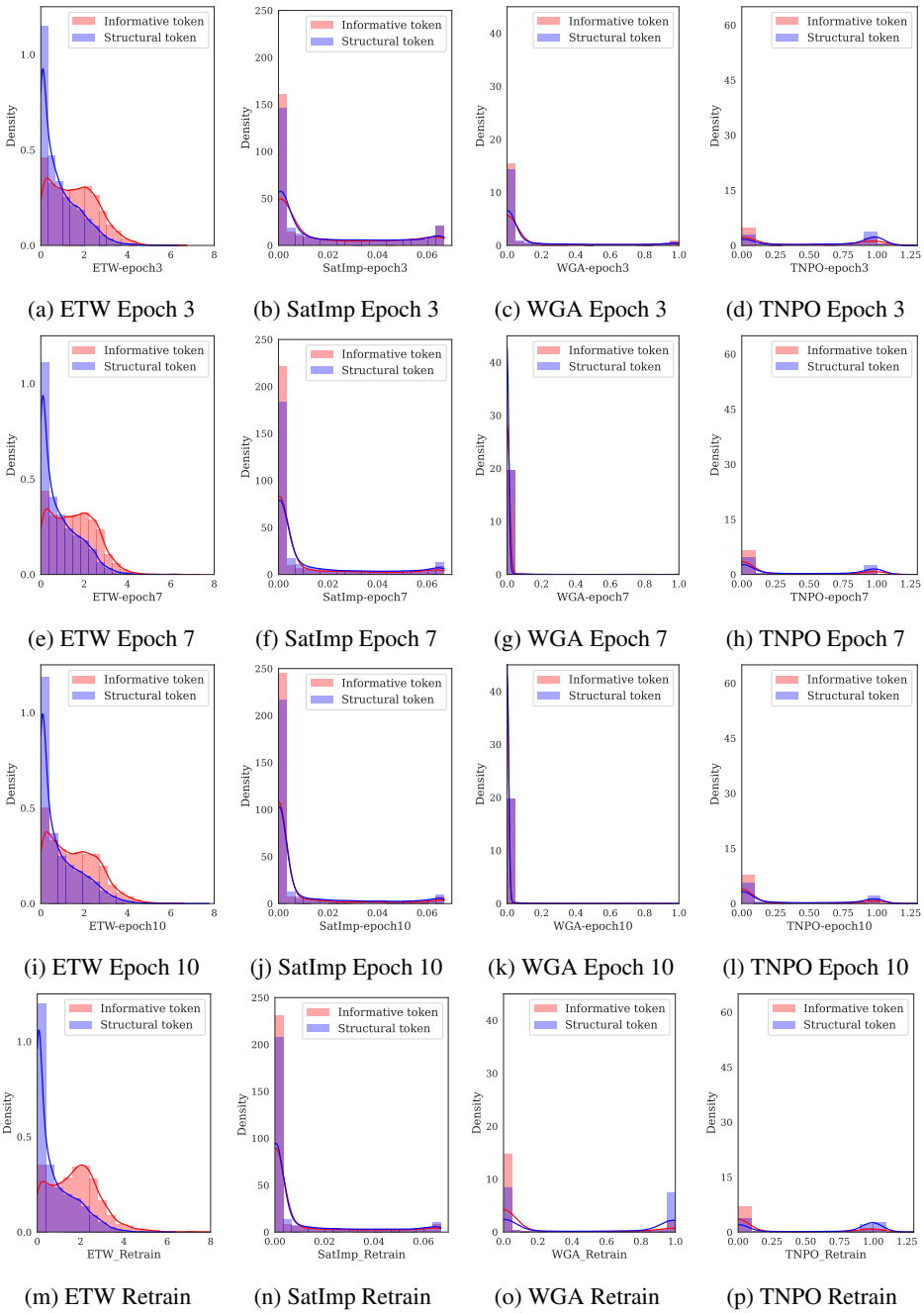
A comparison between the fine-tuned model–generated phrases and the SatImp annotations for the examples included in the manuscript is presented below.

Prompt used for automatic informative-token extraction

You must answer ONLY using the information in the Original answer.
Ignore any outside knowledge, even if it seems correct.
Question: {question}
Original answer: {answer}
Based ONLY on the Original answer, answer the Question again.
Respond with a SINGLE short phrase. Do NOT write a full sentence. Do NOT add any explanation. Do NOT repeat or rephrase the question.
Output ONLY that short phrase and nothing else.

Overall, the summaries generated by the fine-tuned model appear reasonably aligned with answer-critical content. However, we observed cases such as the fourth case in Table 9, where key infinitive constructions such as “to draw” and “to acknowledge and appreciate” were partially omitted, resulting in incomplete coverage of informative spans. Given that automatically extracted spans still require human verification to ensure correctness and consistency, we decided to adopt the SatImp annotations for informative token discrimination analysis.

Note that when using extracted informative tokens automatically, the AUC values in Figure 2 move closer to random performance. Specifically, the AUC values are 0.58 for ETW, 0.46 for WGA, 0.54 for Imp, 0.54 for SatImp, and 0.51 for TNPO. For SCE and SCN, the (FPR, TPR) pairs are (0.211, 0.341) and (0.337, 0.584), respectively. Although the overall discrimination performance degrades toward near-random levels, ETW still demonstrates the strongest discriminative capability among the compared methods.



(a) ETW Epoch 3 (b) SatImp Epoch 3 (c) WGA Epoch 3 (d) TNPO Epoch 3

(e) ETW Epoch 7 (f) SatImp Epoch 7 (g) WGA Epoch 7 (h) TNPO Epoch 7

(i) ETW Epoch 10 (j) SatImp Epoch 10 (k) WGA Epoch 10 (l) TNPO Epoch 10

(m) ETW Retrain (n) SatImp Retrain (o) WGA Retrain (p) TNPO Retrain

Figure 9: Epoch-wise token regularization under TOFU 10% forget split. Token-level weights of ETW, SatImp, WGA, and TNPO are shown at epochs 3, 7, and 10 for the best models in the 10% forget split. The last row corresponds to the retrained model trained on the retain set only.