A TOKEN RENDERING PERSPECTIVE ON LARGE LANGUAGE MODEL UNLEARNING

Anonymous authors
Paper under double-blind review

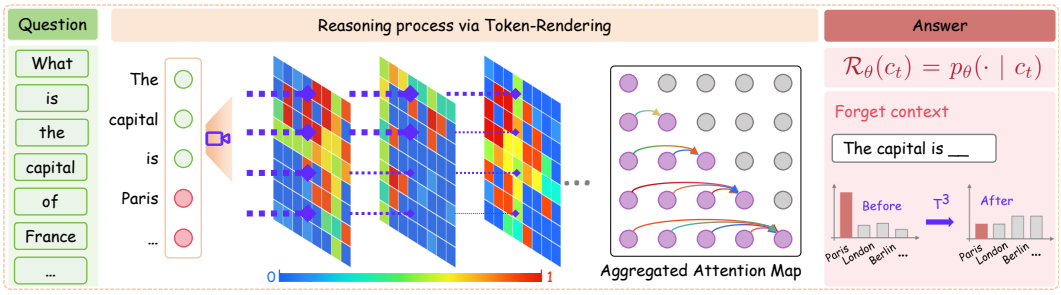

Figure 1: Overview of T^3-Unlearning under the Token Rendering perspective. Given an input question, autoregressive generation is modeled as a dynamic rendering process where contextual information penetrates and accumulates across intermediate multi-head attention layers. By modulating internal token-to-token relation transmittance, T^3-Unlearning explicitly regulates how contextual cues contribute to target token rendering. As a result, the intervention aims to reduce the generation probability of target answers (e.g., "Paris") in forget contexts while limiting changes to model behavior on retained data.

ABSTRACT

LLM unlearning seeks to remove targeted knowledge without disrupting retained capabilities. We argue that many existing output-based methods control which predictions to suppress without explicitly parameterizing the contextual relations involved in producing them. We therefore introduce a token rendering perspective, which views autoregressive generation as a hierarchical rendering process in which contextual information is transmitted through attention-mediated token relations and progressively rendered into next-token probabilities. This perspective provides a common first-order characterization of several representative unlearning objectives as forms of target-token renderability control, highlighting that these objectives do not explicitly parameterize the underlying token relations. Based on this view, we propose **T^3-Unlearning**, which freezes the base LLM and explicitly tunes the renderability of token-token relations. We estimate the sensitivity of semantic relations to forget and retain rendering, attenuate forget-sensitive relations, and constrain changes to retain-sensitive ones. For efficient intervention, relation transmittance is parameterized over semantic token clusters. Experiments on TOFU and WMDP across multiple LLMs and forgetting ratios demonstrate that T^3-Unlearning achieves competitive forgetting–utility trade-offs under the evaluated settings. Code will be released upon acceptance.

1 INTRODUCTION

Large language models (LLMs) have achieved remarkable success across diverse natural language tasks, largely owing to their vast parameters that memorize extensive world knowledge (Wang et al., 2023; Geng et al., 2024). However, this capacity for deep memorization raises critical challenges regarding data privacy, copyright compliance, and safety risks, such as the accidental exposure of personally identifiable information (PII) or hazardous domain knowledge (Wang et al., 2025c). Retraining from scratch without the designated data provides a reference for unlearning, but its sub-

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.

stantial computational cost can limit its feasibility for LLMs. In response, LLM unlearning has emerged as a crucial paradigm, aiming to erase targeted knowledge while preserving general model utility and reasoning capabilities on retained domains (Liu et al., 2025; Zuo et al., 2026).

Despite recent progress, existing unlearning methods have progressively improved the selectivity of forgetting, evolving from sequence-level objectives (Jang et al., 2023; Fan et al., 2025b) to token-aware weighting (Wang et al., 2025a; Yang et al., 2025), distribution- or retain-aware objectives (Dong et al., 2025; Wang et al., 2025b; Koh et al., 2026; Yüce et al., 2026), and internal interventions on representations, parameters, or attention (Li et al., 2024b; Jia et al., 2024; Ashuach et al., 2026; Tan et al., 2026b; Zare Zade et al., 2026). Yet these methods primarily determine *what* to suppress or *where* to intervene, while providing limited control over *how* contextual information propagates through the relations that jointly support forget and retain predictions. This limitation is important because suppressing a target prediction under a specific context does not necessarily disrupt the contextual pathways through which the same knowledge can be rendered under semantically equivalent or rephrased contexts. For example, reducing the probability of generating *"Paris"* for *"What is the capital of France?"* may leave overlapping contextual relations available when the model is queried with *"Which city serves as the capital of France?"*. Conversely, indiscriminately suppressing relations that contribute to the target prediction may also interfere with retained predictions when these relations are shared, potentially leading to noticeable utility degradation (Tan et al., 2026b; Gu et al., 2026).

To better understand this gap, we establish a formal *token-rendering perspective* to analyze autoregressive generation and unlearning dynamics. Inspired by the strength of volume rendering (Max, 2002; Mildenhall et al., 2021; Zhan & Liu, 2025) in modeling continuous accumulation and tracing information propagation along rendering paths, we formalize token prediction as a model-dependent mapping $\mathcal{R}_\theta$ from contextual inputs to next-token probability distributions, decomposing generation into relation-level information transmission and layer-wise accumulation. Quantifying the generation propensity via token renderability , our formulation shifts the focus of unlearning from directly suppressing target-token predictions to controlling the internal relational dynamics through which contextual information contributes to their rendering across diverse contexts. Through this lens, we derive a unified first-order characterization of representative output-level unlearning objectives. Specifically, our theoretical analysis reveals that existing loss functions reweight target-token optimization signals while leaving the implicit rendering couplings unmonitored, acting as degenerate approximations of full rendering control. Therefore, a promising direction is to intervene directly at the internal transmission bottleneck by regulating how contextual representations propagate through multi-head attention mechanisms.

Motivated by this analysis, we introduce T^3**-Unlearning**, short for Token-Relation Transmittance Tuning (Figure 1), which intervenes in the internal information transmission process rather than directly modifying model parameters or final logits. Specifically, we introduce a shared relation-transmittance matrix into the attention mechanism of a frozen LLM to regulate how contextual information is transmitted through token relations. To make this relation space compact and semantically meaningful, we parameterize the relation transmittance at the semantic-cluster level, where each pair of semantic clusters is associated with a shared optical-depth parameter. We then quantify, at the original model, how changing each relation affects forget-target and retained-token rendering, yielding complementary forget and retain sensitivities that characterize its potential forgetting benefit and retention risk, respectively. Based on these signals, we fix the estimated sensitivities and optimize only the relation parameters with a first-order objective that increases the optical depth of relations highly sensitive to forget-target rendering while penalizing changes to relations that strongly affect retained rendering. The resulting relation matrix is shared across attention layers and heads and converted into transmittance for inference, enabling T^3-Unlearning to selectively attenuate the contextual pathways supporting forgotten knowledge while keeping the underlying LLM parameters fixed and mitigating interference with retained knowledge.

Our contributions are threefold. First, we propose a novel token-rendering perspective that unifies the first-order analysis of representative output-based unlearning objectives. Second, we instantiate this perspective as T^3-Unlearning, an efficient method that explicitly tunes internal token-token relation transmittance by balancing forget-sensitive attenuation and retain-sensitive preservation. Third, extensive evaluations on TOFU and WMDP benchmarks across multiple LLM architectures demonstrate that T^3-Unlearning achieves competitive performance in balancing unlearning robustness and utility preservation, complemented by relearning and continual-unlearning evaluations.

2 Related Work

2.1 LLM Unlearning

LLM unlearning aims to suppress designated information while preserving useful capabilities. Gradient Ascent (GA) directly increases the negative log-likelihood of forget responses (Jang et al., 2023), providing a simple formulation for suppressing target knowledge. However, its uniform treatment of forget tokens can impose excessive pressure on already-suppressed predictions. SimNPO alleviates the reliance on a reference model by adopting a length-normalized sequence-level objective (Fan et al., 2025b), but still applies forgetting pressure at the sequence level. WGA further introduces confidence-dependent token weighting to allocate forgetting pressure more selectively (Wang et al., 2025a). SatImp subsequently combines token saturation and importance to refine this allocation according to both the current forgetting state and token contribution (Yang et al., 2025). Other recent methods similarly explore distribution-level targets (Dong et al., 2025; Wang et al., 2025b), entropy-based weighting (Koh et al., 2026), or retain-aware token importance (Yüce et al., 2026). Beyond output-level objectives, internal interventions provide more direct control over model representations and computation. RMU redirects forget representations (Li et al., 2024b), while WAGLE and CRISP target attribution-guided parameters or salient features (Jia et al., 2024; Ashuach et al., 2026). AS and ASU further intervene in attention behavior to alter contextual information processing (Tan et al., 2026b; Zare Zade et al., 2026).

Collectively, these approaches make forgetting increasingly selective by refining either the objective, the token-level weighting, or the internal intervention target. Yet they primarily specify *where* forgetting pressure is applied, leaving open *how* this pressure propagates through contextual relations that may jointly support forget and retain predictions. Addressing this question requires connecting the unlearning objective to the relations through which its effects are rendered and propagated.

2.2 Progress and the Overlooked Role of Token Relations

Recent advances improve selective forgetting through token weighting, internal interventions, and retain regularization. However, a shared contextual relation can support both unwanted and useful predictions, making its treatment difficult to infer from token importance alone.

The less explored issue is how to connect objective design and relation-mediated effects within a common account of unlearning. Existing gradient-interaction analysis already characterizes how unlearning updates affect forget and retain behavior (Wang et al., 2025a). Attention-flow studies likewise show why individual attention weights provide an incomplete picture of information propagation across layers (Abnar & Zuidema, 2020). Our token-rendering framework builds on these insights by giving a unified first-order characterization of representative objectives, including GA, SimNPO, WGA, and SatImp. It separates objective-dependent weighting from model-dependent rendering couplings: the former specifies how strongly predictions are discouraged, while the latter determines how an update affects other predictions. This distinction connects different optimization rules without equating their objectives.

This perspective motivates explicit relation-level intervention. Whereas AS and ASU modify attention through shifting or smoothing, T^3-Unlearning introduces shared attention biases indexed by token-cluster pairs, using forget and retain sensitivities to guide attenuation while limiting interference with retained predictions.

3 LLM Unlearning through a Token-Rendering Perspective

LLM unlearning aims to remove specified knowledge while preserving model behavior on non-targeted data (Zare Zade et al., 2026; Tan et al., 2026a). Existing formulations characterize the desired unlearning behavior but do not explicitly describe how target knowledge is expressed during generation. A forget sample may contain both target-specific associations and linguistic structures useful to retained knowledge (Hu et al., 2025). Treating an autoregressive response as an indivisible target thus obscures the forget–retain distinction. We introduce token rendering to connect the unlearning objective with the process through which output distributions are formed.

3.1 A THEORY OF TOKEN RENDERING

Given a tokenized prompt $q = (q_1, \ldots, q_N)$ and a response $a = (a_1, \ldots, a_T)$, an autoregressive LLMs parameterized by θ defines the conditional response probability as $p_\theta(a \mid q) = \prod_{t=1}^{T} p_\theta(a_t \mid q, a_{<t})$, where $a_{<t} = (a_1, \ldots, a_{t-1})$ denotes the response prefix preceding the t-th generation step.

Rather than treating generation solely as sequence-level prediction, we examine each decoding step as a token-rendering process. At step t, the model maps the context $(q, a_{<t})$ through its internal representations to a probability distribution over the next token. This perspective enables us to analyze not only the token distribution but also the internal components involved in its formation.

Definition 1 (Token Rendering)**.** *Let $c_t = (q, a_{<t})$ denote the context at generation step t. Token rendering is the model-dependent mapping $\mathcal{R}_\theta : \mathcal{C} \to \Delta(\mathcal{V})$, where $\mathcal{R}_\theta(c_t) = p_\theta(\cdot \mid c_t)$, $\mathcal{C}$ is the context space, $\mathcal{V}$ is the vocabulary, and $\Delta(\mathcal{V})$ denotes the set of probability distributions over $\mathcal{V}$.*

Token rendering is realized through a hierarchy of Transformer layers, where self-attention establishes contextual token relations and progressively integrates information for next-token prediction (Vaswani et al., 2017). Given the input representation $\mathbf{H}_t^0$ of context c_t, the model computes $\mathbf{H}_t^l = \mathcal{F}_{\theta^l}^l(\mathbf{H}_t^{l-1})$ for $l = 1, \ldots, L$, where $\mathcal{F}_{\theta^l}^l$ denotes the l-th Transformer layer parameterized by θ^l. Within each layer, self-attention establishes relations among contextual tokens and aggregates their value representations. Let $\mathbf{Q}_t^l = \mathbf{H}_t^{l-1}\mathbf{W}_Q^l$, $\mathbf{K}_t^l = \mathbf{H}_t^{l-1}\mathbf{W}_K^l$, and $\mathbf{V}_t^l = \mathbf{H}_t^{l-1}\mathbf{W}_V^l$. The attention matrix is computed as

$$\mathbf{A}_t^l = \mathrm{softmax}\left(\frac{\mathbf{Q}_t^l(\mathbf{K}_t^l)^\top}{\sqrt{d_k}}\right), \tag{1}$$

where $\mathbf{A}_t^l$ determines how the value representations $\mathbf{V}_t^l$ are aggregated. Through successive Transformer layers, the contextual information is progressively integrated into the final hidden state $\mathbf{h}_{\theta,t}^L$ at the current generation position. This state is projected onto the vocabulary, completing the token-rendering process as $\mathcal{R}_\theta(c_t) = \mathrm{softmax}(\mathbf{W}_o\mathbf{h}_{\theta,t}^L)$. To quantify the likelihood of target outputs under token rendering, we define renderability at both token and sequence levels.

Definition 2 (Renderability)**.** *The renderability of a candidate token v under context $c_t = (q, a_{<t})$ is defined as $r_\theta(v \mid c_t) = \log[\mathcal{R}_\theta(c_t)]_v = \log p_\theta(v \mid c_t)$. Accordingly, the renderability of a response sequence $a = (a_1, \ldots, a_T)$ is $R_\theta(a \mid q) = \sum_{t=1}^{T} r_\theta(a_t \mid c_t) = \log p_\theta(a \mid q)$.*

3.2 TOKEN RENDERING FOR LLM UNLEARNING

We use token rendering to formulate LLM unlearning as an intervention on the rendering process while retaining the standard retraining-based objective (Bourtoule et al., 2021; Maini et al., 2024).

Definition 3 (Token Rendering Unlearning)**.** *Given an original model θ_0, a forget set $\mathcal{D}_f$, and a retain set $\mathcal{D}_r$, token rendering unlearning applies an intervention $\mathcal{U}$ to one or more components of the hierarchical rendering process, producing $\theta_u = \mathcal{U}(\theta_0; \mathcal{D}_f, \mathcal{D}_r)$. Its ideal objective is*

$$\mathbb{E}_{c \sim P_\mathcal{C}}\left[D(\mathcal{R}_{\theta_u}(c), \mathcal{R}_{\theta_r}(c))\right] \leq \epsilon, \tag{2}$$

where θ_r is a reference model retrained on $\mathcal{D}_r$, $P_\mathcal{C}$ is the evaluation-context distribution, $D(\cdot, \cdot)$ measures distributional discrepancy, and $\epsilon \geq 0$ is the allowed deviation.

Under this formulation, unlearning goes beyond suppressing target responses by modifying the rendering process to align next-token distributions with those of a model retrained without the forget data. This perspective provides a behavioral target and motivates an examination of how existing unlearning objectives affect the underlying rendering process.

3.3 REVISITING LLM UNLEARNING THROUGH TOKEN RENDERING

Under the token-rendering formulation, we first examine unlearning objectives defined over model outputs (Liu et al., 2025). Although these objectives take different forms, they alter token rendering through a common first-order mechanism.

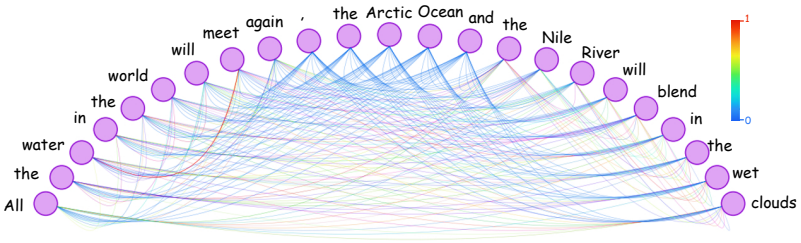

Figure 2: Conceptual illustration of token-relation transmittance. T^3-Unlearning learns which token-to-token relations to attenuate so that forget-related rendering pathways are weakened while retained knowledge is preserved.

Proposition 1 (First-Order Renderability Change)**.** *For context–token pairs* $\{(c_i, v_i)\}_{i\in\mathcal{I}}$*, let* $r_i(\theta) = r_\theta(v_i \mid c_i)$ *and* $\mathcal{L}_u(\theta) = \Phi(\{r_i(\theta)\}_{i\in\mathcal{I}})$ *be twice continuously differentiable. After one gradient step* $\theta^+ = \theta - \eta\nabla_\theta\mathcal{L}_u(\theta)$*, the renderability of any token* v *under context* c *satisfies*

$$r_{\theta^+}(v \mid c) - r_\theta(v \mid c) = -\eta \sum_{i\in\mathcal{I}} w_i(\theta)\kappa_\theta((c, v), (c_i, v_i)) + \mathcal{O}(\eta^2), \tag{3}$$

where $w_i(\theta) = \partial\Phi/\partial r_i$ *and* $\kappa_\theta((c, v), (c_i, v_i)) = \nabla_\theta r_\theta(v \mid c)^\top \nabla_\theta r_i(\theta)$.

The proof is provided in Appendix B. Proposition 1 separates the first-order renderability change into two factors. The objective determines the signed weights w_i, while the model determines how their effects propagate through the rendering couplings κ_θ. We next instantiate these weights for representative unlearning objectives.

(1) Gradient Ascent (GA) (Jang et al., 2023) performs gradient ascent on the forget-set negative log-likelihood, or equivalently minimizes $\mathcal{L}_{\mathrm{GA}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D}_f}[R_\theta(a \mid q)]$. GA assigns the uniform weight $w_t^{\mathrm{GA}} = 1$ to every token in a forget response. A sufficiently small gradient step therefore decreases the aggregate renderability objective on the forget set. **(2) Simple Negative Preference Optimization (SimNPO)** (Fan et al., 2025b) uses a length-normalized, reference-free objective to adjust forgetting according to sequence renderability: $\mathcal{L}_{\mathrm{SimNPO}} = -\frac{2}{\beta}\mathbb{E}_{(q,a)\sim\mathcal{D}_f}[\log\sigma(-\frac{\beta}{T}R_\theta(a \mid q) - \gamma)]$. Its rendering weight is $w_t^{\mathrm{SimNPO}} = \frac{2}{T}\sigma(\frac{\beta}{T}R_\theta(a \mid q) + \gamma)$, which is shared by all tokens in the same response. Thus, SimNPO adjusts intervention strength to sequence renderability without distinguishing token-level contributions. **(3) Weighted Gradient Ascent (WGA)** (Wang et al., 2025a) mitigates excessive unlearning by assigning confidence-dependent weights to forget tokens. At iteration k, its rendering weight is $w_t^{\mathrm{WGA},(k)} = \exp\big(\alpha r_{\theta^{(k)}}(a_t \mid c_t)\big)$, where $\alpha > 0$ controls the weighting strength. This weight is computed from the current model and held fixed during the update. Since lower token renderability yields a smaller weight, WGA progressively weakens the intervention on tokens that have already been sufficiently suppressed. **(4) Saturation–Importance Reweighting (SatImp)** (Yang et al., 2025) combines saturation- and importance-based token weighting. At iteration k, its rendering weight is $w_t^{\mathrm{SatImp},(k)} = \exp\big(\beta_1 r_{\theta^{(k)}}(a_t \mid c_t)\big)\big[1 - \exp\big(r_{\theta^{(k)}}(a_t \mid c_t)\big)\big]^{\beta_2}$, where $\beta_1, \beta_2 > 0$ control the two weighting components. The weight is computed from the current model and held fixed during the update. SatImp therefore assigns larger interventions to tokens with intermediate renderability while downweighting those near either renderability extreme. **(5) Attention Smoothing Unlearning (ASU)** (Zare Zade et al., 2026) constructs a forget teacher by increasing the attention softmax temperature and distills its output distribution into the student. Let $\widetilde{p}_{\tau,t}(v)$ denote the teacher probability of token v under context c_t. Its forget objective is $\mathcal{L}_{\mathrm{ASU}} = \mathbb{E}_{(q,a)\sim\mathcal{D}_f}[\frac{1}{T}\sum_{t=1}^{T}\mathrm{KL}(\widetilde{p}_{\tau,t}\|p_\theta(\cdot \mid c_t))]$. With the teacher distribution held fixed during the student update, the corresponding rendering weight is $w_{t,v}^{\mathrm{ASU}} = -\widetilde{p}_{\tau,t}(v)/T$ for each $v \in \mathcal{V}$. Thus, ASU redistributes renderability across the vocabulary toward an attention-smoothed teacher, rather than directly suppressing only the observed target tokens.

Detailed derivations of these rendering weights are provided in Appendix C. These methods differ in selection of context–token pairs and assignment of rendering weights, while their effects on internal rendering remain mediated by model-dependent couplings κ_θ. When forget and retain inputs share tokens but differ in their contextual relations, these objectives lack explicit control over the token relations. This motivates direct intervention on the attention relations defined in Equation 1.

4 OUR METHOD: T^3-UNLEARNING

Existing objectives control token rendering through rendering weights w_i without explicitly regulating token relations. We propose T^3-Unlearning, which performs *Token-Relation Transmittance Tuning*. As shown in Figure 2, token rendering can be viewed as information transmission over token-to-token relations. T^3-Unlearning learns relation transmittance to weaken forget-related rendering pathways while preserving retained knowledge.

4.1 RELATION TRANSMITTANCE IN ATTENTION

Token rendering relies on attention to compose contextual information. Each query–key relation (x_i, x_j) determines the contribution of x_j to the representation of x_i. Inspired by exponential attenuation in volume rendering (Max, 2002; Mildenhall et al., 2021), we define its *generalized transmittance* as $\tau_{ij} = \exp(-\Gamma_{ij})$, where $\Gamma_{ij} \in \mathbb{R}$ denotes the *relation depth*. Positive and negative depths correspond to relation attenuation and enhancement, respectively. Given the original attention logit $s_{ij}^{\ell,h} = (q_i^{\ell,h})^\top k_j^{\ell,h}/\sqrt{d_k}$, the modulated attention weight is

$$\widetilde{\alpha}_{ij}^{\ell,h} = \frac{\exp(s_{ij}^{\ell,h})\tau_{ij}}{\sum_{r\le i}\exp(s_{ir}^{\ell,h})\tau_{ir}} = \frac{\exp(s_{ij}^{\ell,h} - \Gamma_{ij})}{\sum_{r\le i}\exp(s_{ir}^{\ell,h} - \Gamma_{ir})}. \tag{4}$$

Thus, $\Gamma_{ij} > 0$ suppresses the relation, $\Gamma_{ij} < 0$ enhances it, and $\Gamma_{ij} = 0$ recovers the original attention. Learning independent relation depths for all vocabulary-level token pairs requires $|\mathcal{V}|^2$ parameters. For compact parameterization, static token embeddings are clustered into K semantic groups using k-means (Ahmed et al., 2020). Let $c : \mathcal{V} \to \{1, \dots, K\}$ denote the cluster mapping. The relation depth is parameterized as $\Gamma_{ij} = [\mathbf{\Gamma}]_{c(x_i),c(x_j)}$, where $\mathbf{\Gamma} \in \mathbb{R}^{K\times K}$ is the cluster-level *relation depth map*. Shared across layers and heads, $\mathbf{\Gamma}$ is initialized as $\mathbf{0}$, with the original model frozen and only its K^2 entries optimized during unlearning.

4.2 FORGET–RETAIN RENDERING SENSITIVITY

Not all token relations contribute equally to forget and retain knowledge. We therefore characterize each cluster-level relation by its modulation effect on token rendering over the forget and retain sets. For (q, a) with $a = (a_1, \dots, a_T)$, the rendering loss on $\mathcal{D}_d$, $d \in \{f, r\}$, is defined as

$$\mathcal{L}_d(\mathbf{\Gamma}) = -\mathbb{E}_{(q,a)\sim\mathcal{D}_d}\left[\frac{1}{T}\sum_{t=1}^{T}\log p_{\theta_0,\mathbf{\Gamma}}(a_t \mid q, a_{<t})\right], \qquad d \in \{f, r\}, \tag{5}$$

where $p_{\theta_0,\mathbf{\Gamma}}$ denotes the token rendering distribution under $\mathbf{\Gamma}$. At step k, the forget and retain sensitivities of relation (u, v), $u, v \in \{1, \dots, K\}$, are defined as

$$G_{uv}^{f,(k)} = \left.\frac{\partial\mathcal{L}_f(\mathbf{\Gamma})}{\partial[\mathbf{\Gamma}]_{uv}}\right|_{\mathbf{\Gamma}=\mathbf{\Gamma}^{(k)}}, \qquad G_{uv}^{r,(k)} = \left.\frac{\partial\mathcal{L}_r(\mathbf{\Gamma})}{\partial[\mathbf{\Gamma}]_{uv}}\right|_{\mathbf{\Gamma}=\mathbf{\Gamma}^{(k)}}. \tag{6}$$

At each step, only cluster-pair relations occurring in the corresponding forget or retain minibatch receive gradients, yielding sparse sensitivity maps. The sign and magnitude of $G_{uv}^{f,(k)}$ indicate how modulating a relation affects forgetting. A nonzero $G_{uv}^{r,(k)}$ identifies a relation whose modulation may affect retained rendering in the current minibatch, motivating its regularization toward zero.

4.3 TOKEN-RELATION TRANSMITTANCE TUNING

Based on the rendering sensitivities, relation transmittance is tuned by adjusting the corresponding entries of the *relation depth map*. The forget sensitivity determines the direction and strength of relation modulation, while the retain sensitivity identifies relations that should be protected. Specifically, at step k, the tuning objective is

$$\mathcal{L}_{\mathrm{T}^3}^{(k)}(\mathbf{\Gamma}) = -\sum_{u,v} G_{uv}^{f,(k)}[\mathbf{\Gamma}]_{uv} + \sum_{u,v}\mathbb{I}\left[G_{uv}^{r,(k)} \neq 0\right][\mathbf{\Gamma}]_{uv}^2. \tag{7}$$

The sensitivities are evaluated at $\mathbf{\Gamma}^{(k)}$ and treated as fixed when optimizing this objective. The first term adjusts relation depths in the forget-sensitive direction. The second pulls relations selected by nonzero retain sensitivity toward zero, recovering their original attention behavior.

Corollary 2. *Assume that $\mathcal{L}_f$ and $\mathcal{L}_r$ are differentiable at $\mathbf{\Gamma}^{(k)}$. For any sufficiently small update $\Delta\mathbf{\Gamma}^{(k)}$, we have $\mathcal{L}_f(\mathbf{\Gamma}^{(k)} + \Delta\mathbf{\Gamma}^{(k)}) - \mathcal{L}_f(\mathbf{\Gamma}^{(k)}) = \sum_{u,v} G^{f,(k)}_{uv}[\Delta\mathbf{\Gamma}^{(k)}]_{uv} + o(\|\Delta\mathbf{\Gamma}^{(k)}\|_F)$, and $|\mathcal{L}_r(\mathbf{\Gamma}^{(k)} + \Delta\mathbf{\Gamma}^{(k)}) - \mathcal{L}_r(\mathbf{\Gamma}^{(k)})| \leq \sum_{u,v} |G^{r,(k)}_{uv}||[\Delta\mathbf{\Gamma}^{(k)}]_{uv}| + o(\|\Delta\mathbf{\Gamma}^{(k)}\|_F)$, where $G^{f,(k)}$ and $G^{r,(k)}$ are the rendering sensitivities at $\mathbf{\Gamma}^{(k)}$.*

The proof is provided in Appendix D. Corollary 2 shows that the local forgetting direction is determined by $G^{f,(k)}$, while $G^{r,(k)}$ characterizes the local effect on retained rendering.

During unlearning, the LLM remains frozen while only the shared relation depth map $\mathbf{\Gamma}$ is optimized, with $G^{f,(k)}$ and $G^{r,(k)}$ updated at each step. Unlike conventional objectives that directly suppress target-token probabilities, T^3-Unlearning intervenes in token rendering by modulating contextual relations. For example, given the prompt "Who is the author of Harry Potter?", the method can modulate relations through which contextual cues such as "Harry Potter" contribute to the target response, thereby reducing its renderability.

5 Experiments

5.1 Experimental Setting

Datasets. We evaluate T^3-Unlearning across both synthetic and real-world unlearning benchmarks. For synthetic knowledge removal, we use TOFU (Maini et al., 2024), which contains 4,000 question-answer pairs covering 200 fictitious authors. Following the standard protocol, we consider Forget-01, Forget-05, and Forget-10, corresponding to approximately 1%, 5%, and 10% of the target data, respectively, while using the remaining samples as the retain set. We also use the Real Authors and World Facts subsets to evaluate general knowledge preservation. For real-world sensitive-domain suppression, we evaluate on WMDP (Li et al., 2024a), focusing on WMDP-Chem, and use MMLU (Hendrycks et al., 2021) to quantify general utility retention.

Target Models and Baselines. We validate T^3-Unlearning on three open-source LLMs: Llama-3.1-8B, Llama-3.2-3B, and Qwen2.5-7B. We compare our method with representative unlearning approaches, including Gradient Ascent (GA) (Jang et al., 2023), SimNPO (Fan et al., 2025b), Weighted Gradient Ascent (WGA) (Wang et al., 2025a), SatImp (Yang et al., 2025), and ASU (Zare Zade et al., 2026). All baseline methods are combined with Gradient Descent (GD) on the retain set to mitigate utility degradation. Baseline models are trained with LoRA fine-tuning.

Evaluation Metrics. On TOFU, Model Utility (MU), Forget Efficacy (FE), and their average (Avg.) are reported. Following Yuan et al. (2025), MU and FE are defined as $\mathrm{MU} = \mathrm{H}(R, P, \max(0, 1 - TR), CS, ES, TE)$ on the retain set and the Real Authors and World Facts sets, and $\mathrm{FE} = \mathrm{H}(1 - R, 1 - P, 1 - \min(TR, 1/TR), 1 - ES, TE)$, where $\mathrm{H}(\cdot)$ denotes the harmonic mean. Here, R, P, TR, CS, ES, and TE denote ROUGE-L Recall, Probability, Truth Ratio, Cosine Similarity, Entailment Score, and Token Entropy, respectively. Higher MU, FE, and Avg. indicate better overall performance. For WMDP and MMLU, Accuracy (ACC) is reported, where lower WMDP accuracy indicates better forgetting, while higher MMLU accuracy indicates better utility preservation.

Implementation Details. We freeze the base LLM and optimize the shared relation depth map $\mathbf{\Gamma}$. The forget and retain rendering sensitivities are computed with respect to $\mathbf{\Gamma}$ at each optimization step, starting from $\mathbf{\Gamma} = 0$. We set $K = 10{,}000$, resulting in 10^8 cluster-level relation entries. The relation module remains active during inference. Unless otherwise specified, all methods are evaluated under the same data splits, model backbones, and retain-set correction protocol.For the baseline methods, we employ LoRA fine-tuning with a rank of 8, an α value of 32, and a dropout rate of 0.05. Detailed hyperparameter settings are provided in the Appendix.

5.2 Main Unlearning Evaluation Results

TOFU results. Table 1 reports the main TOFU results across forgetting ratios and model scales. T^3-Unlearning consistently achieves a strong balance between FE and MU. This balance is important because the base models preserve utility but fail to forget, whereas several baselines improve FE only by sacrificing retained capability. In contrast, T^3-Unlearning improves FE while keeping MU competitive, indicating that relation-level transmittance tuning can suppress forget-target renderabil-

Table 1: Results of different unlearning methods on the ToFU benchmark across Llama-3.2-3B, Llama-3.1-8B, and Qwen2.5-7B. MU and FE denote Model Utility and Forget Efficacy, respectively. Avg. represents the average of MU and FE. Higher values indicate better performance.

Model	Methods	Forget-01			Forget-05			Forget-10		
		MU	FE	Avg.	MU	FE	Avg.	MU	FE	Avg.
Llama-3.2-3B	**Base**	80.31	14.15	47.23	80.31	22.55	51.43	80.31	23.83	52.07
	GA	73.90	75.68	74.79	69.48	65.77	67.63	73.66	69.18	71.42
	SimNPO	74.45	50.90	62.68	60.98	80.49	70.74	77.01	76.49	76.75
	WGA	77.63	74.69	76.16	77.60	71.33	74.47	77.50	63.08	70.29
	SatImp	67.12	64.61	65.87	47.86	75.04	61.45	74.08	46.49	60.29
	ASU	76.01	45.16	60.59	73.45	39.71	56.58	71.94	41.44	56.69
	T^3 (Ours)	78.74	75.11	76.93	74.45	81.17	77.81	77.37	79.63	78.50
Llama-3.1-8B	**Base**	87.10	0.63	43.87	87.10	3.45	45.28	87.10	2.82	44.96
	GA	74.10	72.74	73.42	81.25	52.14	66.70	81.12	58.64	69.88
	SimNPO	55.86	64.07	59.97	64.22	58.12	61.17	61.63	58.02	59.83
	WGA	83.42	69.86	76.64	82.71	71.83	77.27	85.57	40.74	63.16
	SatImp	85.82	5.62	45.72	82.00	22.18	52.09	74.98	36.23	55.61
	ASU	66.88	55.76	61.32	61.55	35.04	48.30	47.95	44.87	46.41
	T^3 (Ours)	82.57	81.68	82.13	81.58	80.63	81.11	79.64	80.36	80.00
Qwen2.5-7B	**Base**	67.91	0.00	33.96	67.91	1.50	34.69	67.91	2.02	34.95
	GA	59.21	73.91	66.56	57.81	63.93	60.87	61.39	70.44	65.92
	SimNPO	60.69	73.75	67.22	57.08	78.40	67.74	52.56	71.75	62.16
	WGA	63.23	77.39	70.31	55.58	78.22	66.90	51.61	72.42	62.02
	SatImp	60.80	33.42	47.11	52.25	59.31	55.78	49.91	65.03	57.47
	ASU	58.40	73.58	65.99	55.00	70.38	62.69	52.27	68.59	60.43
	T^3(Ours)	63.14	78.47	70.81	63.55	80.22	71.89	66.21	76.01	71.11

Table 2: Results of unlearning methods on WMDP-Chem and MMLU across Llama-3.1-8B and Qwen2.5-7B models. (↑) indicates higher is better, and (↓) indicates lower is better.

Model	Metric	Base	GA	SimNPO	WGA	SatImp	ASU	Ours
Llama-3.1-8B	WMDP-Chem ↓	50.98	44.36	39.95	37.99	39.71	45.34	36.52
	MMLU ↑	62.38	58.33	54.69	54.96	56.79	54.99	58.66
Qwen2.5-7B	WMDP-Chem ↓	52.21	28.92	40.93	26.47	39.95	40.69	35.54
	MMLU ↑	68.81	40.34	58.22	47.09	60.22	63.57	62.21

ity without broadly disrupting the model's general generation behavior. Across the three evaluated model families, T^3-Unlearning achieves consistently high Avg. scores. The stability across both forgetting ratios and model architectures suggests that T^3-Unlearning is not merely fitting a particular forget split. Instead, it appears to identify relation patterns that repeatedly contribute to forget-target rendering while avoiding relations that are important for retained behavior.

Effect of forgetting ratio. As the forget ratio increases, unlearning becomes more difficult because a larger portion of target knowledge must be suppressed while the retain set must still constrain general behavior. T^3-Unlearning remains stable under this harder setting. On Llama-3.1-8B, Avg. decreases only from 82.13 on Forget-01 to 80.00 on Forget-10, despite the larger forget set. On Llama-3.2-3B and Qwen2.5-7B, the Avg. scores remain within a narrow range across all three splits. This pattern supports the motivation behind shared semantic relation attenuation: by operating on clustered token-token relations rather than individual examples or final logits, T^3-Unlearning can generalize its intervention across related forget instances.

WMDP results. Table 2 reports results on WMDP-Chem and MMLU. On Llama-3.1-8B, T^3-Unlearning achieves the lowest WMDP-Chem accuracy, 36.52%, while retaining 58.66% MMLU accuracy. This indicates effective suppression of sensitive-domain knowledge with limited utility loss. On Qwen2.5-7B, T^3-Unlearning reduces WMDP-Chem accuracy to 35.54% and maintains 62.21% MMLU accuracy. Although some baselines reduce WMDP-Chem accuracy more aggressively on Qwen2.5-7B, they incur much larger MMLU degradation. Thus, the key advantage of T^3-Unlearning is not unconditional suppression, but selective suppression: it attenuates sensitive-domain renderability while preserving a stronger level of general capability. Figure 3 further visualizes this mechanism, showing that T^3-Unlearning weakens forget-related high-attention pathways while largely preserving the causal attention structure.

5.3 Relearning Robustness Analysis

We test whether forgotten knowledge can be recovered by subsequent fine-tuning. This setting is stricter than standard post-unlearning evaluation: a method may suppress target answers immediately after unlearning, yet still allow rapid recovery once fine-tuned on the forget set (Fan et al.,

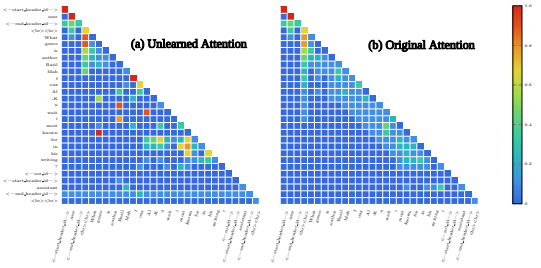

Figure 3: Attention maps before and after T^3-Unlearning. The unlearned model weakens forget-related high-attention pathways while preserving the overall causal attention structure.

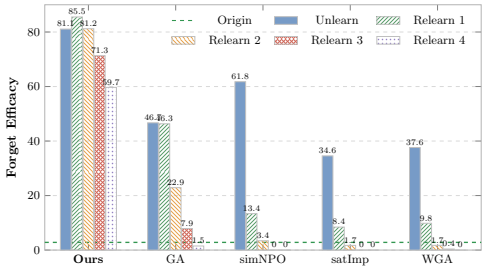

Figure 4: Forget Efficacy of T^3-Unlearning and baseline methods after 1–4 epochs of relearning on the TOFU-Forget10 dataset with Llama-3.1-8B.

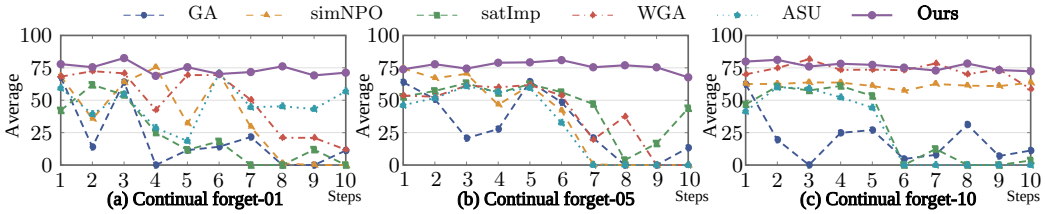

Figure 5: Average of Model Utility and Forget Efficacy over ten continual unlearning steps under Forget-01, Forget-05, and Forget-10.

2025a). We perform relearning on TOFU-Forget10 with Llama-3.1-8B by fine-tuning each unlearned model on the forget set for 1–4 epochs and measuring Forget Efficacy (FE) after each stage. As shown in Figure 4, T^3-Unlearning remains substantially more robust than the baselines. Its FE changes from 81.1 after unlearning to 85.5, 81.2, 71.3, and 59.7 after 1–4 relearning epochs. In contrast, GA drops from 46.4 to 1.5, and SimNPO, SatImp, and WGA approach the origin-level FE after repeated relearning. These results indicate that baseline methods mainly produce fragile forgetting, where target knowledge is quickly restored by direct exposure to the forget set. T^3-Unlearning resists relearning by attenuating the token relations that transmit forget-relevant context to target tokens. Relearning must therefore rebuild weakened rendering pathways rather than merely adjust surface-level output preferences. This supports the main claim that relation-level transmittance control yields more persistent forgetting than output-level suppression.

5.4 Continual Unlearning Analysis

We evaluate T^3-Unlearning in a continual setting, where the model receives multiple forgetting requests over time. This setting tests whether an unlearning method can preserve the forgetting-utility trade-off under repeated updates. We perform ten consecutive unlearning steps under Forget-01, Forget-05, and Forget-10, and report the average of Model Utility (MU) and Forget Efficacy (FE). As shown in Figure 5, T^3-Unlearning maintains a consistently high average score across all three settings, whereas the baselines fluctuate sharply and often collapse in later steps. This suggests that repeated output-level or gradient-based forgetting can accumulate instability and utility degradation. In contrast, T^3-Unlearning freezes the base model and updates only the shared relation depth map, allowing forgetting constraints to accumulate through controlled relation-level intervention.

6 Conclusion

We presented T^3-Unlearning, a token-rendering approach to LLM unlearning that controls how forget-relevant context is transmitted to target tokens. Instead of directly modifying the base model or only suppressing output probabilities, T^3-Unlearning freezes the LLM and tunes semantic token-relation transmittance to balance forget-sensitive attenuation with retain-sensitive preservation. Experiments on TOFU and WMDP show that this relation-level intervention improves the forgetting-utility trade-off across models and forgetting ratios, and further improves robustness under relearning and continual unlearning. These results suggest that modeling unlearning as controlled relation transmission offers a practical route toward more stable and selective knowledge removal.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing, manuscript organization, and drafting. All technical content, experimental results, and scientific conclusions were reviewed and verified by the authors, who take full responsibility for the final manuscript.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4190–4197, 2020.

Mohiuddin Ahmed, Raihan Seraj, and Syed Mohammed Shamsul Islam. The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, 9(8):1295, 2020.

Tomer Ashuach, Dana Arad, Aaron Mueller, Martin Tutek, and Yonatan Belinkov. Crisp: Persistent concept unlearning via sparse autoencoders. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1806–1825, 2026.

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

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

Chongyu Fan, Jinghan Jia, Yihua Zhang, Anil Ramakrishna, Mingyi Hong, and Sijia Liu. Towards llm unlearning resilient to relearning attacks: a sharpness-aware minimization perspective and beyond. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25. JMLR.org, 2025a.

Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Simplicity prevails: Rethinking negative preference optimization for llm unlearning. *Advances in Neural Information Processing Systems*, 38:1540–1567, 2025b.

Mingyang Geng, Shangwen Wang, Dezun Dong, Haotian Wang, Ge Li, Zhi Jin, Xiaoguang Mao, and Xiangke Liao. Large language models are few-shot summarizers: Multi-intent comment generation via in-context learning. In *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering*, pp. 1–13, 2024.

Renjie Gu, Jiazhen Du, Yihua Zhang, and Sijia Liu. Unlearners can lie: Evaluating and improving honesty in llm unlearning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11930–11952, 2026.

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

Shengyuan Hu, Neil Kale, Pratiksha Thaker, Yiwei Fu, Steven Wu, and Virginia Smith. Blur: A benchmark for llm unlearning robust to forget-retain overlap, 2025. URL `https://arxiv.org/abs/2506.15699`.

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.

Jinghan Jia, Jiancheng Liu, Yihua Zhang, Parikshit Ram, Nathalie Baracaldo, and Sijia Liu. Wagle: Strategic weight attribution for effective and modular unlearning in large language models. *Advances in Neural Information Processing Systems*, 37:55620–55646, 2024.

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, 2026.

Nathaniel Li, Alexander Pan, Anjali Gopal, Sumil Yue, Sheng Guo, et al. Wmdp: Measuring and reducing malicious use with unlearning. In *International Conference on Machine Learning*, 2024a.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew Bo Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Ariel Herbert-Voss, Cort B Breuer, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam Alfred Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Ian Steneker, David Campbell, Brad Jokubaitis, Steven Basart, Stephen Fitz, Ponnurangam Kumaraguru, Kallol Krishna Karmakar, Uday Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 28525–28550. PMLR, 21–27 Jul 2024b. URL https://proceedings.mlr.press/v235/li24bc.html.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, et al. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, 7(2):181–194, 2025.

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

Nelson Max. Optical models for direct volume rendering. *IEEE Transactions on visualization and computer graphics*, 1(2):99–108, 2002.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Chenchen Tan, Xinghao Li, Shujie Cui, Youyang Qu, Cunjian Chen, and Longxiang Gao. Less is more: Geometric unlearning for LLMs with minimal data disclosure. In *Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=XmWKBDYmtk.

Chenchen Tan, Youyang Qu, Xinghao Li, Hui Zhang, Shujie Cui, Cunjian Chen, and Longxiang Gao. Wisdom is knowing what not to say: Hallucination-free llms unlearning via attention shifting. *Advances in Neural Information Processing Systems*, 39:99138–99166, 2026b.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Bailin Wang, Zi Wang, Xuezhi Wang, Yuan Cao, Rif A Saurous, and Yoon Kim. Grammar prompting for domain-specific language generation with large language models. *Advances in Neural Information Processing Systems*, 36:65030–65055, 2023.

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

Yaxuan Wang, Jiaheng Wei, Yuhao Liu, Jinlong Pang, Quan Liu, Ankit Parag Shah, Yujia Bao, Yang Liu, and Wei Wei. Llm unlearning via loss adjustment with only forget data. In *International Conference on Learning Representations*, volume 2025, pp. 43076–43104, 2025b.

Yubo Wang, Min Tang, Nuo Shen, Shujie Cui, and Weiqing Wang. Privacy risks of llm-empowered recommender systems: An inversion attack perspective. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, pp. 812–821, 2025c.

Puning Yang, Qizhou Wang, Zhuo Huang, Tongliang Liu, Chengqi Zhang, and Bo Han. Exploring criteria of loss reweighting to enhance llm unlearning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 2025.

Xiaojian Yuan, Tianyu Pang, Chao Du, Kejiang Chen, Weiming Zhang, and Min Lin. A closer look at machine unlearning for large language models. In *International Conference on Learning Representations*, 2025.

Gizem Yüce, Giorgos Nikolaou, and Nicolas Flammarion. Learning what to forget: Improving llm unlearning via learned token-level importance. *arXiv preprint arXiv:2606.06320*, 2026.

Saleh Zare Zade, Xiangyu Zhou, Sijia Liu, and Dongxiao Zhu. Attention smoothing is all you need for unlearning. In *International Conference on Learning Representations*, volume 2026, pp. 19583–19621, 2026.

Xiaohang Zhan and Dingming Liu. Larender: Training-free occlusion control in image generation via latent rendering. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19679–19688. IEEE, 2025.

Xuhan Zuo, Minghao Wang, Tianqing Zhu, Lefeng Zhang, Dayong Ye, Shui Yu, and Wanlei Zhou. Federated trustchain: Blockchain-enhanced llm training and unlearning. *IEEE Transactions on Dependable and Secure Computing*, 2026.