STaR: Sensitive Trajectory Regulation for Unlearning in Large Reasoning Models

Jingjing Zhou[1*], **Gaoxiang Cong**[2,1], **Li Su**[1†], **Liang Li**[1,2†]

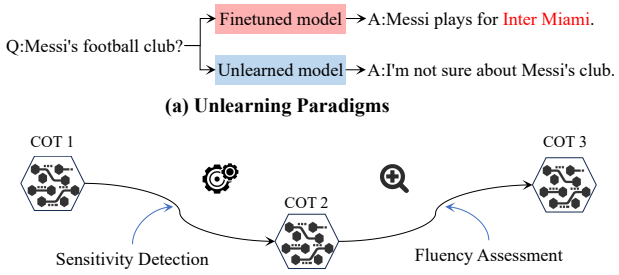
[1]University of Chinese Academy of Sciences
[2]Institute of Computing Technology, Chinese Academy of Sciences
zhoujingjing25@ucas.ac.cn, gaoxiang.cong@vipl.ict.ac.cn, suli@ucas.ac.cn, liang.li@ict.ac.cn

Abstract

Large Reasoning Models (LRMs) have advanced automated multi-step reasoning, but their ability to generate complex Chain-of-Thought (CoT) trajectories introduces severe privacy risks, as sensitive information may be deeply embedded throughout the reasoning process. Existing Large Language Models (LLMs) unlearning approaches that typically focus on modifying only final answers are insufficient for LRMs, as they fail to remove sensitive content from intermediate steps, leading to persistent privacy leakage and degraded security. To address these challenges, we propose Sensitive Trajectory Regulation (STaR), a parameter-free, inference-time unlearning framework that achieves robust privacy protection throughout the reasoning process. Specifically, we first identify sensitive content via semantic-aware detection. Then, we inject global safety constraints through secure prompt prefix. Next, we perform trajectory-aware suppression to dynamically block sensitive content across the entire reasoning chain. Finally, we apply token-level adaptive filtering to prevent both exact and paraphrased sensitive tokens during generation. Furthermore, to overcome the inadequacies of existing evaluation protocols, we introduce two metrics: Multi-Decoding Consistency Assessment (MCS), which measures the consistency of unlearning across diverse decoding strategies, and Multi-Granularity Membership Inference Attack (MIA) Evaluation, which quantifies privacy protection at both answer and reasoning-chain levels. Experiments on the R-TOFU benchmark demonstrate that STaR achieves comprehensive and stable unlearning with minimal utility loss, setting a new standard for privacy-preserving reasoning in LRMs.

Introduction

With the rapid advancement of large language models (LLMs) (Achiam et al. 2023; Team et al. 2023; Touvron et al. 2023; Taylor et al. 2022; Bao et al. 2024) and large reasoning models (LRMs), which are capable of generating complex multi-step chain-of-thought (CoT) reasoning, have become a central paradigm in contemporary AI research. Representative models such as OpenAI o1 (Jaech et al. 2024) and DeepSeek R1 (Guo et al. 2025) exhibit advanced autonomous reasoning capabilities, generating coherent and structured reasoning trajectories (Wei et al. 2022) without explicit prompting, and have demonstrated state-of-the-art performance in challenging domains such as mathematical proof, program synthesis, and domain-specific question answering.



(a) Unlearning Paradigms

(b) Illustration of Trajectory-Aware Suppression Learning

Figure 1: (a) Illustration of the effect of unlearning in LLMs. (b) Trajectory-Aware Suppression Learning detects sensitivity and evaluates fluency at each reasoning step to adaptively suppress sensitive information throughout the reasoning trajectory.

Despite architectural differences between LLMs and LRMs, both rely extensively on large-scale pretraining corpora, which inevitably contain copyrighted materials (Karamolegkou et al. 2023; Zhang et al. 2024c; Chu, Song, and Yang 2024), personal information (Carlini et al. 2021), and other sensitive content (Mireshghallah et al. 2023). As data protection regulations such as the General Data Protection Regulation (GDPR) (Staufer 2025) and the California Consumer Privacy Act (CCPA) become increasingly stringent, the development of machine unlearning techniques that facilitate the selective removal of sensitive information (Xiong et al. 2025) from trained models is imperative for ensuring legal compliance and maintaining user trust. The primary goal of machine unlearning (Cao and Yang 2015; Feng et al. 2025) is to eliminate the influence of designated data instances while preserving overall model utility, which is a critical capability

*First author.
†Corresponding authors.

for responsible AI deployment (Xu et al. 2025).

Conventional machine unlearning approaches for large language models (Maini et al. 2024; Golatkar, Achille, and Soatto 2020; Rafailov et al. 2023; Reisizadeh et al. 2025; Wan et al. 2025) typically focus on answer-level interventions, such as amplifying loss on the forget set or optimizing for refusal responses. The effect of such traditional LLM unlearning methods is illustrated in Figure 1(a). Although they are effective for suppressing sensitive information in standard LLM settings, they cannot apply to LRMs with multi-step CoT generation. Because they do not address the risk of sensitive knowledge persisting within intermediate reasoning steps. To bridge this gap, R-TOFU (Yoon, Jeung, and No 2025) establishes the first benchmark by extending LLM-based unlearning strategies to the CoT domain. However, empirical evidence from R-TOFU shows that existing methods still suffer from insufficient forgetting effect, substantial degradation of model utility, and pronounced vulnerability to information leakage under alternative decoding strategies such as ZeroThink (which omits the reasoning trace) and LessThink (which reduces reasoning steps) (Jiang et al. 2025). These limitations emphasize the pressing need for decoding-robust, trajectory-level unlearning mechanisms that can comprehensively and reliably suppress sensitive content throughout the reasoning process.

To address these problems, we propose Sensitive Trajectory Regulation (STaR), a parameter-free, inference-time unlearning framework, which is equipped with a novel trajectory-aware suppression learning to ensure robust privacy protection throughout the entire reasoning process. Specifically, we first identify potentially sensitive queries through semantic-aware detection and retrieve the most relevant fragments from the designated forget set, constructing a fine-grained set of forbidden phrases and tokens. Second, the secure prompt prefix is applied to the sensitive queries by prepending global safety instructions, thereby reinforcing privacy intent at the input level in a non-intrusive, model-agnostic manner. Third, the proposed trajectory-aware suppression learning operates as the high-level controller, it dynamically inspects generated reasoning chains for both fluency and the presence of sensitive content at each step, enforcing real-time regulation, as schematically illustrated in Figure 1(b). When residual sensitivity is detected or fluency is insufficient, the responsible tokens are escalated for stricter filtering or the output is replaced by a safe refusal template. In particular, we introduce a token-level adaptive filtering to adaptively manipulate token probabilities at each generation step through hard and soft constraints, ensuring both exact and semantic variants of sensitive information are comprehensively blocked. Together, these modules form a unified, decoding-agnostic pipeline that delivers systematic privacy protection across all decoding strategies.

Besides, existing evaluation metrics (Chen et al. 2025; To and Le 2025) are typically limited to answer-level forgetting and overlook robustness under diverse decoding settings. To address this, we introduce the Multi-Decoding Consistency Score (MCS) and Multi-Granularity Membership Inference Attack (MIA) Evaluation, which together offer a comprehensive assessment of unlearning security across different reasoning stages and adversarial scenarios. Experiments on R-TOFU show that these metrics uncover hidden privacy risks and demonstrate STaR's consistent advantages, underscoring the need for holistic evaluation in future unlearning research.

The main contributions are summarized as follows:

- We propose STaR, a novel inference-time unlearning framework for large reasoning models, which achieves robust and decoding-agnostic suppression of sensitive information without any parameter updates.
- We design Trajectory-Aware Suppression Learning and Token-level Adaptive Filtering for comprehensive, context-sensitive suppression of sensitive information.
- We introduce two evaluation metrics for rigorous, decoding-agnostic assessment of unlearning security and privacy, including MCS and MIA.
- Extensive experiments on R-TOFU demonstrate that STaR outperforms state-of-the-art baselines in forgetting efficacy, robustness, and privacy protection.

Related Work

Unlearning in LLMs

With the widespread deployment of LLMs in real-world applications, the tendency of these models to memorize training data, together with the resulting privacy risks, has become a critical concern. Mainstream unlearning approaches for LLMs predominantly rely on model parameter updates (Mekala et al. 2024; Chen and Yang 2023; Jia et al. 2024b; Yuan et al. 2024; Zhang et al. 2024b; Scholten et al. 2025; Zhao et al. 2025; Li et al. 2025), such as maximizing the loss on the forget set (gradient ascent) and preference optimization (replacing sensitive answers with refusals or neutral statements), thereby achieving suppression of sensitive content at the answer level through fine-tuning (Sun et al. 2024; Jia et al. 2024a; Sinha, Mandal, and Kankanhalli 2024; Wang et al. 2024; Zhang et al. 2024a). However, these methods entail substantial computational overhead and carry the risk of catastrophic forgetting, which may degrade the utility of retained knowledge. In recent years, there has been increasing interest in parameter-free approaches for unlearning. For example, Soft Prompting for Unlearning (SPUL) (Bhaila, Van, and Wu 2024) learns soft prompt prefixes to steer the model away from sensitive content, while Embedding Corrupted Prompts (ECO) (Liu et al. 2024) introduces perturbations in the embedding space to inhibit the recall of sensitive knowledge. Although these approaches are amenable to deployment, they often rely on external detectors or auxiliary optimization procedures, and their effectiveness is limited (Kuo et al. 2025; Tu et al. 2024; Liu et al. 2023; Li et al. 2022) in scenarios involving complex reasoning chains or diverse query formulations. More importantly, existing LLM unlearning methods typically target only the final answers, and thus

struggle to achieve comprehensive suppression of sensitive information embedded throughout the multi-step reasoning trajectories that are characteristic of large reasoning models.

Unlearning in LRMs

Large reasoning models (LRMs) generate multi-step chain-of-thought (CoT) trajectories, embedding sensitive information throughout the reasoning process and elevating privacy risks. Existing LLM unlearning techniques—such as gradient ascent, preference optimization, and KL regularization—primarily operate at the answer level. The R-TOFU benchmark extends these methods to CoT-level intervention, but experiments reveal that sensitive knowledge can persist in intermediate steps, leading to substantially weaker chain-level forgetting. Moreover, LRMs support diverse decoding strategies (e.g., DefaultThink, ZeroThink, LessThink), which alter output formats and may expose forgotten information even after effective default-mode unlearning. These findings underscore the inherent challenge of achieving consistent, robust suppression of sensitive content across both answers and reasoning chains under all decoding settings.

Preliminaries

Fine-tuning-based Unlearning

Traditional machine unlearning in LLMs and LRMs primarily relies on fine-tuning-based approaches, which explicitly update model parameters to mitigate the influence of sensitive or undesired training data. Given an original model $\mathcal{M}_o$ trained on dataset $\mathcal{D}$, the training data is partitioned into a forget set $\mathcal{D}_f$ and a retain set $\mathcal{D}_r$. Fine-tuning-based methods, such as gradient ascent (maximizing the loss on $\mathcal{D}_f$), preference optimization (forcing the model to output refusals or neutral responses for $\mathcal{D}_f$), and KL-divergence regularization (aligning the unlearned model's distribution with a retain-only model), aim to minimize the model's ability to recall or reproduce forgotten knowledge while preserving performance on $\mathcal{D}_r$. Formally, let θ_o denote the parameters of the original model and θ_u those of the unlearned model. The unlearning objective is often expressed as:

$$\min_{\theta_u} \; \mathcal{L}_r(\theta_u) - \lambda \, \mathcal{L}_f(\theta_u) + \beta \, \mathrm{KL}(p_u \,\|\, p_r), \qquad (1)$$

where λ and β are hyperparameters, and $\mathcal{L}_{\text{retain}}$ and $\mathcal{L}_{\text{forget}}$ denote loss terms on $\mathcal{D}_r$ and $\mathcal{D}_f$, respectively. Despite their effectiveness, these methods entail significant computational overhead and may lead to catastrophic forgetting, undermining utility on the retain set.

Inference-time Parameter-free Unlearning

Several recent works propose parameter-free unlearning approaches that operate exclusively at inference time, without altering model weights. For example, Embedding-Corrupted Prompts (ECO) first detects whether an input query is related to the forget set using a trained classifier. If so, ECO applies targeted perturbations to the embedding representation of the prompt, aiming to disrupt the model's ability to recall sensitive knowledge. This approach modifies the input space rather than the model parameters, and can be deployed rapidly without retraining. Other methods, such as Soft Prompting for Unlearning (SPUL), learn dedicated soft prompts that are prepended to user queries to steer the model away from forbidden content. While these techniques are efficient and compatible with existing black-box models, they typically intervene only at the prompt or embedding level, and do not dynamically control the stepwise generation process. Consequently, their effectiveness may be limited in scenarios involving multi-step reasoning chains or diverse decoding strategies.

In contrast, our approach, Sensitive Trajectory Regulation (STaR), introduces stepwise dynamic intervention during inference, enabling fine-grained suppression of sensitive content throughout the reasoning chain and across various decoding strategies. The detailed methodology of STaR is presented in the Methodology section.

Decoding Strategies

Large reasoning models are capable of generating outputs through a variety of decoding strategies, each with distinct implications for privacy risk and unlearning robustness. The most common strategies include:

- **DefaultThink**: The model generates a complete multi-step chain-of-thought, revealing the full intermediate reasoning process before producing the final answer.
- **ZeroThink**: The model omits the reasoning chain and outputs only the final answer, thereby compressing or bypassing intermediate steps.
- **LessThink**: The model generates a condensed reasoning chain, offering minimal or summarized intermediate reasoning prior to the answer.

Different decoding strategies can expose sensitive information despite effective unlearning under default settings. Thus, robust unlearning requires consistent knowledge suppression across all decoding paradigms, necessitating decoding-strategy-aware evaluation as a critical component of LRM privacy and security assessment.

Methodology

Overview of the STaR Framework

Sensitive Trajectory Regulation (STaR) is an inference-time unlearning framework for large reasoning models, designed to eliminate sensitive information leakage at any point within multi-step CoT generation. Unlike traditional static filtering, STaR dynamically enforces suppression throughout the reasoning process and across all decoding strategies via four modules: Sensitive Content Identification, Secure Prompt Prefix, Trajectory-Aware Suppression Learning, and Token-level Adaptive Filtering.

Sensitive Content Identification

Formally, given an input query x, STaR first determines whether it is related to the forget set using a scope classifier $C(\cdot)$ trained over semantic embeddings of both the forget

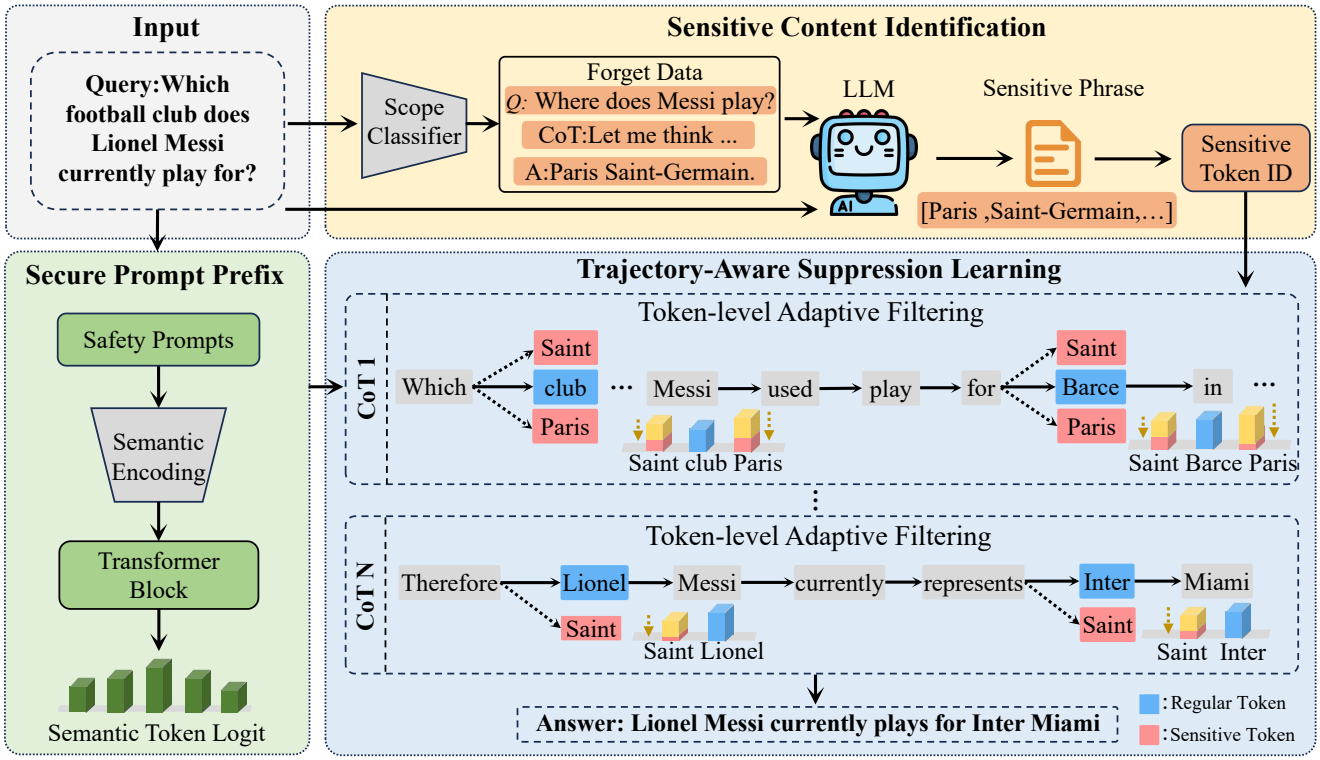


Figure 2: Architecture of the STaR framework. The pipeline consists of Sensitive Content Identification, Secure Prompt Prefix, Trajectory-Aware Suppression Learning, and Token-level Adaptive Filtering.

and retain sets. The classifier assigns a probability score p_f, which indicates the likelihood of x belonging to the forget set:

$$p_f = C(\text{Embed}(x)), \qquad C : \mathbb{R}^d \to [0, 1], \tag{2}$$

where $\text{Embed}(\cdot)$ denotes a fixed pre-trained embedding function.

If $p_f > \tau$, semantic retrieval is performed over the forget set $\mathcal{D}_f$:

$$A^* = \arg\max_{A \in \mathcal{D}_f} \frac{\langle \text{Embed}(x), \text{Embed}(A) \rangle}{\|\text{Embed}(x)\| \cdot \|\text{Embed}(A)\|}, \tag{3}$$

yielding the most semantically similar forgotten instance.

To ensure fine-grained coverage, all fragments in A^* are unified by applying phrase extraction (NER, pattern mining, or LLM-based slot-filling):

$$\mathcal{F}(A^*) = \{f_1, f_2, ..., f_K\}, \tag{4}$$

where each f_k is tokenized into a sequence $\mathbf{t}_k = (t_1^{(k)}, ..., t_{|f_k|}^{(k)})$. The forbidden token set is then

$$\mathcal{T}_f = \bigcup_{k=1}^{K} \mathbf{t}_k, \tag{5}$$

which ensures both lexical and semantic coverage to minimize the risk of paraphrase-based leakage.

Secure Prompt Prefix

Formally, after sensitive content identification, we introduce Secure Prompt Prefix to reinforce privacy objectives at the input level. For each query x identified for unlearning intervention, an abstract safety constraint s is concatenated to x, forming a composite prompt $x' = s \,\|\, x$, where $\|$ denotes sequence concatenation:

$$x' = \text{Encode}_{\text{secure}}(x) = s \,\|\, x. \tag{6}$$

Unlike token-level filtering, Secure Prompt Prefix imposes a global, instruction-level constraint that is non-intrusive and model-agnostic, providing a weak yet flexible form of control. In the overall pipeline, it serves as a complementary safeguard, guiding generation at the prompt level while subsequent adaptive filtering enforces strict token-wise suppression.

Token-level Adaptive Filtering

Token-level Adaptive Filtering serves as the core suppression engine in STaR, operating at each generation step to enforce context-aware intervention. Let $\ell_t(v)$ denote the original logit score for token v at decoding step t, and $\tilde{\ell}_t(v)$ denote the adjusted logit score after applying suppression. Logit values represent the unnormalized scores before the softmax function is applied to generate the token probability distribution.

For each decoding step t and candidate token $v \in \mathcal{V}$, suppression is applied as follows:

- **Soft Suppression**: If token v exhibits high semantic similarity with a forbidden fragment $f_k \in \mathcal{F}(A^*)$, i.e., $\mathrm{sim}(\mathrm{Embed}(v), \mathrm{Embed}(f_k)) \geq \delta$, where Embed denotes the semantic embedding function, then:

  $$\tilde{\ell}_t(v) = \ell_t(v) - \alpha \cdot \mathrm{sim}(\mathrm{Embed}(v), \mathrm{Embed}(f_k)).$$

  Here, $\mathrm{sim}(\mathrm{Embed}(v), \mathrm{Embed}(f_k))$ computes the cosine similarity between the token embedding and the forbidden fragment embedding.
- **Hard Suppression**: If v completes a forbidden span given prefix $y_{<t}$, then:

  $$\tilde{\ell}_t(v) = -\infty.$$

  This effectively ensures that the forbidden token cannot be selected by making its logit value exceedingly negative.

Trajectory-Aware Suppression Learning

Trajectory-Aware Suppression Learning provides high-level control over the generation process by validating and regulating entire reasoning trajectories. It accepts candidate outputs $\mathcal{C} = \{y^{(1)}, \ldots, y^{(N)}\}$ from Token-level Adaptive Filtering and evaluates each candidate $y^{(c)}$ through two criteria: sensitivity and fluency.

For sensitivity assessment, we compute a risk score $\mathcal{S}(y^{(c)})$ based on the output of the sensitive content classifier $C(\cdot)$ and the maximum cosine similarity between $y^{(c)}$ and any forbidden fragment f_k:

$$\mathcal{S}(y^{(c)}) = \max\left(C(y^{(c)}), \max_k \mathrm{sim}(y^{(c)}, f_k)\right). \quad (7)$$

For fluency, we compute a coherence score using a pretrained language model:

$$\mathcal{F}(y^{(c)}) = \text{LM-Score}(y^{(c)}), \quad (8)$$

where higher values indicate more fluent and well-formed text.

Based on these evaluations, we define an adaptive control strategy: If $\mathcal{S}(y^{(c)}) \geq \tau$, the output is flagged as sensitive, and the corresponding tokens are backtracked and added to the forbidden token set $\mathcal{T}_f$ to enforce hard suppression in the next decoding pass. If $\mathcal{S}(y^{(c)}) < \tau$ but $\mathcal{F}(y^{(c)}) < \eta$, then a predefined refusal template replaces the trajectory to maintain response quality. Otherwise, $y^{(c)}$ is retained as the final output.

This mechanism ensures that only safe and fluent reasoning trajectories are surfaced, and that any detected sensitive leakage triggers automatic refinement through re-decoding or fallback mechanisms.

Experiments

Implementation Details

All experiments are conducted on the R-TOFU benchmark (Yoon, Jeung, and No 2025), which is specifically designed for evaluating unlearning in large reasoning models. R-TOFU contains 200 synthetic author profiles, each associated with 20 question–reasoning–answer triples. For each profile, both the full chain-of-thought (CoT) reasoning and the final answer are provided, and sensitive content is explicitly annotated. The benchmark defines three unlearning protocols—forget01, forget05, and forget10—corresponding to removing 1%, 5%, or 10% of the data as the forget set. We follow the official dataset splits and evaluation protocol to enable systematic and hierarchical assessment of unlearning effectiveness and model utility across different levels of forgetting. Baselines include four representative methods: Gradient Ascent (GA)(Golatkar, Achille, and Soatto 2020), KL Minimization (KL), Direct Preference Optimization (DPO)(Rafailov et al. 2023), and Gradient Difference (GD), all implemented with official code and hyperparameters for reproducibility. GA maximizes forget set loss, KL minimizes divergence to a retain-only reference, DPO substitutes neutral responses for sensitive answers, and GD penalizes forget-set gradients. To ensure fairness, we adopt DeepSeekR1-Distill-Llama-8B (Guo et al. 2025) as the backbone model for all experiments, following the official R-TOFU setup. This model is a distilled version of DeepSeekR1 optimized for efficient multi-step chain-of-thought reasoning. It is fine-tuned on the full R-TOFU dataset prior to unlearning, and all subsequent unlearning methods—including our proposed STaR framework and baselines—are applied to this checkpoint.

Evaluation Metrics

We adopt a multi-dimensional evaluation protocol to comprehensively assess the effectiveness, robustness, and privacy guarantees of unlearning in large reasoning models. Following the R-TOFU benchmark (Yoon, Jeung, and No 2025), we use standard metrics such as ROUGE-L, cosine similarity, and entailment score to evaluate model utility on the retain set and forgetting efficacy on the forget set, at both answer and CoT levels. These metrics provide a direct measure of output similarity and factual consistency, serving as the foundation for baseline comparisons.

To address the inherent limitations of answer-level evaluation and single-mode assessment, we introduce two novel evaluation metrics specifically designed to capture the practical security and robustness of unlearning mechanisms:

Multi-Decoding Consistency Score (MCS). Standard unlearning evaluation typically assumes a fixed decoding strategy. However, large reasoning models support diverse generation modes, such as DefaultThink, ZeroThink, and LessThink, which can significantly affect the exposure of residual sensitive information. To systematically measure the robustness of unlearning across all plausible decoding strategies, we propose the Multi-Decoding Consistency Score (MCS), which quantifies the worst-case information leakage over the entire decoding strategy space.

Formally, let $\mathcal{D}$ be the set of representative decoding strategies, and let $\mathrm{Leakage}(y^{(d)}, y^*)$ denotes a

Method	forget01					forget05					forget10				
	MU ↑	AFE ↑	CFE ↑	MCS ↑	Avg ↑	MU ↑	AFE ↑	CFE ↑	MCS ↑	Avg ↑	MU ↑	AFE ↑	CFE ↑	MCS ↑	Avg ↑
GA	0.71	0.57	0.46	0.47	0.54	0.73	0.34	0.35	0.44	0.46	0.72	0.33	0.31	0.42	0.44
GD	0.71	0.48	0.46	0.46	0.52	0.72	0.35	0.35	0.44	0.46	0.72	0.33	0.31	0.41	0.44
KL	0.72	0.48	0.47	0.47	0.54	0.71	0.35	0.35	0.44	0.46	0.73	0.33	0.31	0.42	0.44
PO	0.60	0.68	0.53	0.56	0.60	0.61	0.50	0.32	0.52	0.48	0.63	0.39	0.18	0.49	0.36
STaR	**0.93**	**0.88**	**0.68**	**0.70**	**0.79**	**0.94**	**0.87**	**0.66**	**0.69**	**0.79**	**0.95**	**0.84**	**0.63**	**0.67**	**0.77**

Table 1: Unlearning core metrics (MU, AFE, CFE, MCS) and harmonic mean (Avg) for all methods across forget01, forget05, and forget10. The best results are **in bold** and the second ones are underlined. ↑ (↓) means that higher (lower) value is better.

similarity-based leakage metric (such as ROUGE-L or cosine similarity) between the model's output $y^{(d)}$ under decoding strategy d and the ground-truth sensitive content y^*. For each query q in the forget set, the MCS is defined as:

$$\mathrm{MCS}(q) = 1 - \max_{d \in \mathcal{D}} \mathrm{Leakage}(y^{(d)}, y^*), \tag{9}$$

Specifically, this formulation is inspired by adversarial robustness principles, reflecting the intuition that a secure unlearning mechanism must maintain its forgetting effect even under adversarial or atypical decoding choices. High MCS values indicate that the model reliably suppresses sensitive information regardless of decoding configuration.

Multi-Granularity Membership Inference Attack (MIA) Evaluation. Beyond output similarity, a crucial aspect of privacy is whether the model's responses inadvertently reveal whether a particular sample was part of the forget set—a risk measured by membership inference attacks. We systematically evaluate this risk at both answer and reasoning-chain levels:

MIA-A (Answer-level MIA). MIA-A quantifies the risk that an adversary can infer whether a query-answer pair (q, a) originated from the forget set, solely based on the model's output answer. We follow established privacy auditing practice and train a binary classifier f_{ans}, using features derived from the generated answer, to distinguish between forget and retain samples. The privacy leakage is reported as the area under the receiver operating characteristic curve (AUC-ROC):

$$\mathrm{AUC}_{\mathrm{MIA-A}} = \mathrm{AUC}\left(f_{\mathrm{ans}}(q, a)\right), \tag{10}$$

where an AUC near 0.5 indicates no privacy risk (random guessing), and an AUC near 1.0 implies high vulnerability to answer-level membership inference.

MIA-C (Chain-of-Thought MIA). MIA-C extends the membership inference analysis to full reasoning trajectories, evaluating whether the generated CoT itself leaks membership information. We train a second binary classifier f_{cot}, which takes as input the (q, CoT) pair, to predict forget/retain status. The AUC-ROC for this classifier is similarly reported:

$$\mathrm{AUC}_{\mathrm{MIA-C}} = \mathrm{AUC}\left(f_{\mathrm{cot}}(q, \mathrm{CoT})\right), \tag{11}$$

a high AUC for MIA-C indicates that the structure or content of the model's reasoning process can be exploited for membership inference attacks, revealing privacy weaknesses that answer-only analysis may miss.

By integrating both the Multi-Decoding Consistency Score and the two-tier MIA evaluation (MIA-A and MIA-C), our protocol provides a rigorous, adversarially-aware, and privacy-focused assessment of unlearning efficacy in LRMs, filling critical gaps in existing evaluation practices.

Comparison with SOTA Methods

We present a comprehensive evaluation of all unlearning methods on the R-TOFU benchmark under three unlearning protocols (forget01, forget05, forget10). Core results are summarized in Table 1, while Table 2 reports privacy leakage risk under membership inference attacks.

Overall Effectiveness and Robustness. Table 1 compares the performance of all baselines and our proposed STaR framework across Model Utility (MU), Answer Forget Efficacy (AFE), CoT Forget Efficacy (CFE), and Multi-Decoding Consistency Score (MCS), as well as their harmonic mean (Avg). Across all unlearning ratios, STaR achieves the best or second-best results in every metric, substantially outperforming parameter-tuning baselines (GA, GD, KL, PO) on both answer-level and chain-of-thought-level forgetting. Notably, STaR maintains high utility (MU) on the retain set, indicating negligible side effects on non-sensitive knowledge, while still achieving significant gains in AFE and CFE, demonstrating more thorough removal of sensitive content from both final answers and multi-step reasoning trajectories. The superiority of STaR is especially pronounced in MCS, where it achieves much higher scores than all baselines, reflecting robust, decoding-agnostic protection against information leakage regardless of the user's decoding strategy.

Method	MIA-A ↓			MIA-C ↓		
	01	05	10	01	05	10
GA	0.72	0.76	0.80	0.83	0.86	0.88
GD	0.73	0.78	0.80	0.83	0.86	0.88
KL	0.70	0.76	0.80	0.82	0.86	0.88
PO	0.64	0.69	0.74	0.78	0.81	0.84
STaR	**0.51**	**0.53**	**0.53**	**0.62**	**0.65**	**0.68**

Table 2: Membership inference AUC (lower is better) for answer (MIA-A) and CoT (MIA-C) levels.

Privacy Leakage under Membership Inference Attacks. Table 2 further examines the privacy guarantees of each method by reporting AUC scores for membership inference

attacks at both answer level (MIA-A) and CoT level (MIA-C), where lower values indicate better privacy protection. Consistently, STaR achieves the lowest AUC across all ratios and both attack types, substantially reducing privacy leakage risk compared to other methods. These results confirm that traditional unlearning approaches, even when effective in suppressing output similarity, are often vulnerable to adversarial attacks that exploit subtle statistical traces of forgotten data. In contrast, the dynamic, fine-grained suppression in STaR delivers strong privacy guarantees at multiple reasoning granularities.

Method	Original Prompt ↑	Paraphrased Prompt ↑
GA	0.57	0.31
GD	0.48	0.32
KL	0.48	0.33
PO	<u>0.68</u>	<u>0.37</u>
STaR	**0.88**	**0.82**

Table 3: Answer-level forgetting efficacy (AFE) on the forget01 set under Original and paraphrased prompt queries.

Robustness to Prompt Paraphrasing. To assess the robustness of unlearning methods against semantic variations of input queries, we evaluate forgetting efficacy on the forget set using both the original prompts and paraphrased versions generated by manually or automatically rewording the queries. As shown in Table 3, most baseline methods experience substantial drops in forgetting performance when presented with paraphrased prompts, highlighting their vulnerability to prompt rephrasing attacks. In contrast, STaR consistently maintains high forgetting efficacy, demonstrating strong resilience to both standard and semantically modified queries. This result underscores the practical reliability of our approach for privacy protection in real-world scenarios where user queries may vary in formulation.

Method	Running Time (h) ↓
GA	8.5
GD	8.5
KL	<u>6.0</u>
PO	9.0
STaR	**0.5**

Table 4: End-to-end running time (in hours) for each unlearning method on the R-TOFU benchmark.

Computational Efficiency. Table 4 compares the total end-to-end running time of all methods on the R-TOFU benchmark under the forget10 protocol.All running time measurements are reported on the same NVIDIA H800 GPU setup to ensure fairness. Retraining-based baselines (GA, GD, KL, PO) require multiple epochs of fine-tuning on large-scale models, resulting in high computational overhead. In contrast, our inference-time STaR framework completes the unlearning process in a fraction of the time, delivering over an order-of-magnitude speedup while maintaining superior forgetting efficacy and privacy protection. This substantial efficiency advantage highlights the practical deployability of STaR in real-world scenarios requiring prompt and reliable data deletion.

Method	MU↑	AFE↑	CFE↑	MCS↑	MIA-A↓	MIA-C↓
w/o secure prompt	**0.93**	0.82	0.66	0.65	0.56	0.6
w/o phrase	0.91	0.67	0.54	0.61	0.67	0.71
hard only	0.92	**0.93**	0.62	0.69	0.70	0.68
soft only	0.92	0.62	0.58	0.60	0.51	**0.59**
STaR (Full)	**0.93**	<u>0.88</u>	**0.68**	**0.70**	**0.51**	<u>0.62</u>

Table 5: Ablation study of STaR on the forget01 split.

Ablation Study. Table 5 reports ablation results for each core component of STaR. Removing semantic phrase expansion (*w/o phrase*) significantly impairs both answer- and chain-level forgetting, and markedly increases membership inference risk, underscoring the need for high-order semantic coverage. Relying solely on hard suppression (*hard only*) yields strong answer-level forgetting but fails to prevent privacy leakage or ensure multi-step consistency, while using only soft suppression (*soft only*) weakens exact match blocking and results in inferior forgetting despite modest privacy gains. Notably, ablating the secure prompt module (*w/o secure prompt*) leads to measurable declines in both chain-level forgetting (CFE) and decoding robustness (MCS), confirming that even soft, input-level constraints contribute to comprehensive privacy protection. Only the complete STaR pipeline, which integrates all modules, achieves uniformly strong performance across all metrics, highlighting the necessity of a holistic, multi-layered intervention strategy for robust and secure unlearning in LRMs.

Qualitative Analysis. Beyond quantitative improvements, qualitative results (see *Appendix C*) highlights STaR's unique ability to suppress sensitive content across a broad spectrum of reasoning trajectories and adversarial decoding scenarios. For example, even when input prompts are strategically altered or reasoning steps are truncated , STaR consistently prevents the resurfacing of forgotten information, whereas baseline approaches often expose sensitive details via omitted or compressed reasoning. Representative case studies further illustrate that stepwise, decoding-agnostic intervention is essential for practical, end-to-end privacy protection in large reasoning models.

Conclusion

We present STaR, a novel inference-time unlearning framework for large reasoning models that achieves robust, decoding-agnostic suppression of sensitive information via semantic content identification, adaptive token-level suppression, and stepwise trajectory regulation. Extensive evaluation on the R-TOFU benchmark demonstrates that STaR consistently outperforms state-of-the-art baselines in forgetting efficacy and privacy protection at both the answer and chain-of-thought levels, while ablation studies confirm the indispensability of each module. Our results set a new benchmark for privacy-preserving unlearning in LRMs and provide a practical blueprint for compliant AI deployment.

Acknowledgments

This work was supported by the National Nature Science Foundation of China (62322211), the "Pioneer" and "Leading Goose" R&D Program of Zhejiang Province(2024C01023),Key Laboratory of Intelligent Processing Technology for Digital Music (Zhejiang Conservatory of Music), Ministry of Culture and Tourism (2023DMKLB004).This work was also supported by the National Nature Science Foundation of China (U25A20441: "Virtual–Physical Integrated Spatial Computing Theory and Methods for Complex Equipment Support").

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bao, Y.; Shah, A. P.; Narang, N.; Rivers, J.; Maksey, R.; Guan, L.; Barrere, L. N.; Evenson, S.; Basole, R.; Miao, C.; et al. 2024. Harnessing business and media insights with large language models. *arXiv preprint arXiv:2406.06559*.

Bhaila, K.; Van, M.-H.; and Wu, X. 2024. Soft prompting for unlearning in large language models. *arXiv preprint arXiv:2406.12038*.

Cao, Y.; and Yang, J. 2015. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, 463–480. IEEE.

Carlini, N.; Tramer, F.; Wallace, E.; Jagielski, M.; Herbert-Voss, A.; Lee, K.; Roberts, A.; Brown, T.; Song, D.; Erlingsson, U.; et al. 2021. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, 2633–2650.

Chen, H.; Zhang, Y.; Bi, Y.; Zhang, Y.; Liu, T.; Bi, J.; Lan, J.; Gu, J.; Grosser, C.; Krompass, D.; et al. 2025. Does Machine Unlearning Truly Remove Model Knowledge? A Framework for Auditing Unlearning in LLMs. *arXiv preprint arXiv:2505.23270*.

Chen, J.; and Yang, D. 2023. Unlearn what you want to forget: Efficient unlearning for llms. *arXiv preprint arXiv:2310.20150*.

Chu, T.; Song, Z.; and Yang, C. 2024. How to protect copyright data in optimization of large language models? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 17871–17879.

Feng, Z.; Xu, Y. E.; Robey, A.; Kirk, R.; Davies, X.; Gal, Y.; Schwarzschild, A.; and Kolter, J. Z. 2025. Existing Large Language Model Unlearning Evaluations Are Inconclusive. *arXiv preprint arXiv:2506.00688*.

Golatkar, A.; Achille, A.; and Soatto, S. 2020. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9304–9312.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Jaech, A.; Kalai, A.; Lerer, A.; Richardson, A.; El-Kishky, A.; Low, A.; Helyar, A.; Madry, A.; Beutel, A.; Carney, A.; et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Jia, J.; Liu, J.; Zhang, Y.; Ram, P.; Baracaldo, N.; and Liu, S. 2024a. Wagle: Strategic weight attribution for effective and modular unlearning in large language models. *Advances in Neural Information Processing Systems*, 37: 55620–55646.

Jia, J.; Zhang, Y.; Zhang, Y.; Liu, J.; Runwal, B.; Diffenderfer, J.; Kailkhura, B.; and Liu, S. 2024b. Soul: Unlocking the power of second-order optimization for llm unlearning. *arXiv preprint arXiv:2404.18239*.

Jiang, F.; Xu, Z.; Li, Y.; Niu, L.; Xiang, Z.; Li, B.; Lin, B. Y.; and Poovendran, R. 2025. Safechain: Safety of language models with long chain-of-thought reasoning capabilities. *arXiv preprint arXiv:2502.12025*.

Karamolegkou, A.; Li, J.; Zhou, L.; and Søgaard, A. 2023. Copyright violations and large language models. *arXiv preprint arXiv:2310.13771*.

Kuo, M.; Zhang, J.; Zhang, J.; Tang, M.; DiValentin, L.; Ding, A.; Sun, J.; Chen, W.; Hass, A.; Chen, T.; et al. 2025. Proactive privacy amnesia for large language models: Safeguarding PII with negligible impact on model utility. *arXiv preprint arXiv:2502.17591*.

Li, L.; Cong, G.; Qi, Y.; Zha, Z.-J.; Wu, Q.; Sheng, M.; Huang, Q.; and Yang, M.-H. 2025. Dubbing Movies via Hierarchical Phoneme Modeling and Acoustic Diffusion Denoising. *IEEE TPAMI*.

Li, L.; Gao, X.; Deng, J.; Tu, Y.; Zha, Z.-J.; and Huang, Q. 2022. Long Short-Term Relation Transformer with Global Gating for Video Captioning. *IEEE TIP*.

Liu, C.; Wang, Y.; Flanigan, J.; and Liu, Y. 2024. Large language model unlearning via embedding-corrupted prompts. *Advances in Neural Information Processing Systems*, 37: 118198–118266.

Liu, X.; Li, L.; Wang, S.; Zha, Z.-J.; Li, Z.; Tian, Q.; and Huang, Q. 2023. Entity-Enhanced Adaptive Reconstruction Network for Weakly Supervised Referring Expression Grounding. *IEEE TPAMI*, 45(3): 3003–3018.

Maini, P.; Feng, Z.; Schwarzschild, A.; Lipton, Z. C.; and Kolter, J. Z. 2024. Tofu: A task of fictitious unlearning for llms. *arXiv preprint arXiv:2401.06121*.

Mekala, A.; Dorna, V.; Dubey, S.; Lalwani, A.; Koleczek, D.; Rungta, M.; Hasan, S.; and Lobo, E. 2024. Alternate preference optimization for unlearning factual knowledge in large language models. *arXiv preprint arXiv:2409.13474*.

Mireshghallah, N.; Kim, H.; Zhou, X.; Tsvetkov, Y.; Sap, M.; Shokri, R.; and Choi, Y. 2023. Can llms keep a secret? testing privacy implications of language models via contextual integrity theory. *arXiv preprint arXiv:2310.17884*.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36: 53728–53741.

Reisizadeh, H.; Jia, J.; Bu, Z.; Vinzamuri, B.; Ramakrishna, A.; Chang, K.-W.; Cevher, V.; Liu, S.; and Hong, M. 2025. BLUR: A Bi-Level Optimization Approach for LLM Unlearning. *arXiv preprint arXiv:2506.08164*.

Scholten, Y.; Xhonneux, S.; Günnemann, S.; and Schwinn, L. 2025. Model Collapse Is Not a Bug but a Feature in Machine Unlearning for LLMs. *arXiv preprint arXiv:2507.04219*.

Sinha, Y.; Mandal, M.; and Kankanhalli, M. 2024. Unstar: Unlearning with self-taught anti-sample reasoning for llms. *arXiv preprint arXiv:2410.17050*.

Staufer, D. 2025. What Should LLMs Forget? Quantifying Personal Data in LLMs for Right-to-Be-Forgotten Requests. *arXiv preprint arXiv:2507.11128*.

Sun, C.; Miller, N. A.; Zhmoginov, A.; Vladymyrov, M.; and Sandler, M. 2024. Learning and unlearning of fabricated knowledge in language models. *arXiv preprint arXiv:2410.21750*.

Taylor, R.; Kardas, M.; Cucurull, G.; Scialom, T.; Hartshorn, A.; Saravia, E.; Poulton, A.; Kerkez, V.; and Stojnic, R. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

To, B. T. T.; and Le, T. 2025. Harry potter is still here! probing knowledge leakage in targeted unlearned large language models via automated adversarial prompting. *arXiv preprint arXiv:2505.17160*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tu, Y.; Li, L.; Su, L.; Zha, Z.-J.; and Huang, Q. 2024. SMART: Syntax-Calibrated Multi-Aspect Relation Transformer for Change Captioning. *IEEE TPAMI*.

Wan, Y.; Ramakrishna, A.; Chang, K.-W.; Cevher, V.; and Gupta, R. 2025. Not Every Token Needs Forgetting: Selective Unlearning to Limit Change in Utility in Large Language Model Unlearning. *arXiv preprint arXiv:2506.00876*.

Wang, S.; Zhu, T.; Ye, D.; and Zhou, W. 2024. When machine unlearning meets retrieval-augmented generation (rag): Keep secret or forget knowledge? *arXiv preprint arXiv:2410.15267*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Xiong, A.; Zhao, X.; Pappu, A.; and Song, D. 2025. The Landscape of Memorization in LLMs: Mechanisms, Measurement, and Mitigation. *arXiv preprint arXiv:2507.05578*.

Xu, X.; Du, M.; Ye, Q.; and Hu, H. 2025. OBLIVIATE: Robust and Practical Machine Unlearning for Large Language Models. *arXiv preprint arXiv:2505.04416*.

Yoon, S.; Jeung, W.; and No, A. 2025. R-tofu: Unlearning in large reasoning models. *arXiv preprint arXiv:2505.15214*.

Yuan, X.; Pang, T.; Du, C.; Chen, K.; Zhang, W.; and Lin, M. 2024. A closer look at machine unlearning for large language models. *arXiv preprint arXiv:2410.08109*.

Zhang, B.; Li, L.; Wang, S.; Cai, S.; Zha, Z.-J.; Tian, Q.; and Huang, Q. 2024a. Inductive State-Relabeling Adversarial Active Learning with Heuristic Clique Rescaling. *IEEE TPAMI*.

Zhang, R.; Lin, L.; Bai, Y.; and Mei, S. 2024b. Negative preference optimization: From catastrophic collapse to effective unlearning. *arXiv preprint arXiv:2404.05868*.

Zhang, Y.; Jia, J.; Chen, X.; Chen, A.; Zhang, Y.; Liu, J.; Ding, K.; and Liu, S. 2024c. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. In *European Conference on Computer Vision*, 385–403. Springer.

Zhao, X.; Cai, W.; Shi, T.; Huang, D.; Lin, L.; Mei, S.; and Song, D. 2025. Improving llm safety alignment with dual-objective optimization. *arXiv preprint arXiv:2503.03710*.