

Counterfactual Self-Distillation: A Distribution-Shaping Framework for Entity-Fact Unlearning in Large Language Models

Ying Zhao^{1,3}, Yihui Zhang⁴, Mengyi Deng⁴, Xin Li⁴,
Yuan Tian^{1,3,*}, Wei Wang^{4,5,*}, Yi Chang^{1,2,3,*}

¹School of Artificial Intelligence, Jilin University

²International Center of Future Science, Jilin University

³Engineering Research Center of Knowledge-Driven Human-Machine Intelligence, MOE, China

⁴The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China

⁵The Hong Kong University of Science and Technology, Hong Kong SAR, China

zhaoying22@mails.jlu.edu.cn, {yuantian,yichang}@jlu.edu.cn, weiwcs@ust.hk

Abstract

Entity-fact unlearning in large language models (LLMs) aims to remove the binding between a target entity and a forgotten fact, producing a model that behaves like one retrained without the forgotten data while preserving neighboring knowledge and overall model utility. We frame this problem as distribution shaping: such a retrained model would yield a plausible distribution at forget queries, fluent, semantically appropriate, and reflective of residual knowledge, which we take as the appropriate target. Existing objectives provide only narrow token-level or semantically uninformative supervision, falling short of this target. We propose **Counterfactual Self-Distillation (CSD)**, a self-distillation framework that approximates the retrained distribution by constructing a counterfactual teacher through type-consistent entity substitution. The resulting teacher is decoupled from the target identity while preserving type-level relational structure, enabling directed probability reallocation without an external oracle. Experiments on TOFU and RWKU show that CSD achieves a strong forget-utility trade-off, including the best overall forgetting on RWKU and strong performance on TOFU. Code is available at <https://github.com/zy1998-c/csd>.

1 Introduction

Large language models (LLMs) achieve strong performance across diverse tasks (Achiam et al., 2023; Grattafiori et al., 2024), but their ability to memorize and reproduce training data raises privacy, safety, and legal compliance concerns when deployed (Carlini et al., 2021, 2022b). Since retraining an LLM from scratch for each deletion request is prohibitively expensive, machine unlearning has emerged as a practical paradigm for removing target information while preserving broader model

utility (Cao and Yang, 2015; Bourtole et al., 2021; Xu et al., 2023; Shi et al., 2025).

In many realistic settings, once the target entity appears, the model can recover associated attributes, relations, or biographical details that should no longer be accessible (Jang et al., 2023; Jin et al., 2024). Unlearning therefore aims to remove this entity-fact binding (Meng et al., 2022), producing a model that behaves like one retrained without the data to be forgotten, preserving neighboring knowledge and general linguistic competence (Maini et al., 2024). This raises a natural question: what form does its distribution take at a forget query? Its predictions form a plausible distribution: fluent, semantically appropriate, and reflective of residual knowledge. Reproducing this plausible distribution, rather than collapsing to a single token or imposing an artificial shape, is the appropriate target for entity-fact unlearning.

Existing LLM unlearning objectives provide only narrow token-level or semantically uninformative supervision. Likelihood-suppression methods such as GA (Jang et al., 2023), NPO (Zhang et al., 2024), and SimNPO (Fan et al., 2025) only lower the probability of the target answer, leaving the redirected mass unspecified. RT (Maini et al., 2024) trains the model on a fixed refusal template, a predefined rejection rather than a natural continuation. DPO (Rafailov et al., 2023) and AltPO (Mekala et al., 2025) optimize a pairwise objective that promotes a chosen response while suppressing a dispreferred one, anchoring supervision to specific responses rather than to a distribution. PALU (Zhai et al., 2026) enforces local uniformity over the top-ranked tokens, an artificial shape that ignores the query’s linguistic structure. Figure 1 illustrates these behaviors.

We address this gap by framing entity-fact unlearning as a distribution-shaping problem and approximating the retrained distribution. Since retraining is the cost that unlearning is designed to

*Corresponding authors

Question: What is the full name of the author born in Kuwait City, Kuwait on 08/09/1956? Answer: The full name of ... is ____ Target: Basil					
Method	R1	R2	R3	R4	R5
Retrained	Le 0.97	Fat 0.007	Nad 0.005	R 0.004	L 0.001
NPO	Basil 0.26	Le 0.17	Jad 0.11	B 0.07	Ahmed 0.04
RT	Basil 1.00	Bass 2e-5	Basel 2e-5	Bas 2e-5	basil 1e-6
DPO	not 0.14	:: 0.11	" 0.05	Le 0.05	Ahmed 0.04
PALU	assumed 0.11	fict 0.02	假 0.02	given 0.02	deemed 0.01
CSD (Ours)	Far 0.05	Ahmed 0.04	" 0.03	An 0.03	Jad 0.03

□ Target/prefix □ Plausible token □ Other

Figure 1: Top-5 next-token predictions at the forgotten answer position of a TOFU forget query, ranked by probability (R1–R5). Baselines either retain the target or redirect mass to unrelated tokens, whereas CSD redistributes probability toward plausible tokens.

avoid, we construct a surrogate teacher from the model’s own predictions under counterfactual entity substitution, which we call **Counterfactual Self-Distillation (CSD)**. For each forget sample, CSD substitutes the core entity with multiple type-consistent fictitious entities and takes the average of the model’s detached predictions on these counterfactual prompts as the teacher. The teacher is identity-decoupled from the target binding yet preserves the query’s type-level relational structure, approximating the retrained distribution without retraining. Figure 1 shows that CSD redistributes probability onto plausible tokens.

Our main contributions are as follows:

- We frame entity-fact unlearning as a distribution-shaping problem and identify a shared limitation of existing objectives: they provide narrow token-level or semantically uninformative supervision over where probability mass should move after forgetting.
- We propose CSD, a self-distillation framework that constructs an identity-decoupled, type-preserved counterfactual teacher from the model’s own predictions under entity substitution.
- Experiments on TOFU (Maini et al., 2024) and RWKU (jin et al., 2024) show that CSD achieves a strong forget-utility trade-off against representative baselines, with distri-

butional and ablation analyses supporting its mechanism.

2 Related Work

Machine Unlearning for LLMs. LLM unlearning methods differ in the supervision they impose on the post-unlearning distribution. Likelihood-suppression objectives, including GA (Gradient Ascent) (Jang et al., 2023), GD (Gradient Difference) (Yao et al., 2024), NPO (Negative Preference Optimization) (Zhang et al., 2024), and SimNPO (Fan et al., 2025), reduce the probability of target answer tokens. RT (Rejection Tuning) (Maini et al., 2024) instead promotes a fixed refusal template on forget queries. Pairwise preference methods such as DPO (Rafailov et al., 2023) and AltPO (Mekala et al., 2025) combine these directions by suppressing a dispreferred response while promoting a chosen one. PALU (Zhai et al., 2026) moves beyond selected responses by enforcing local top- K uniformity over logits. However, these objectives either constrain only selected tokens or responses, or impose a semantically uninformative target shape, leaving the destination of redistributed probability mass under-specified.

Knowledge Distillation. Knowledge distillation (KD) transfers behavior from a teacher to a student by matching output distributions (Hinton et al., 2015; Kim and Rush, 2016). Self-distillation removes the need for a separately trained teacher by reusing the model’s own predictions as supervision (Yang et al., 2024). CSD follows this self-distillation principle but uses it for a different purpose: the teacher is constructed by counterfactually substituting the target entity with type-consistent alternatives, so that its predictions define a distributional unlearning target rather than merely reproducing the original model’s behavior.

3 Method

Figure 2 provides an overall illustration of the proposed CSD framework, which we detail below.

3.1 Problem Formulation

Let g_0 denote the original autoregressive language model, S_F the forget set, and S_R the retain set. We seek an updated model g_θ , initialized from g_0 , such that predictions on S_F no longer rely on target-specific entity-fact bindings, while general language ability and unrelated knowledge are preserved on S_R .

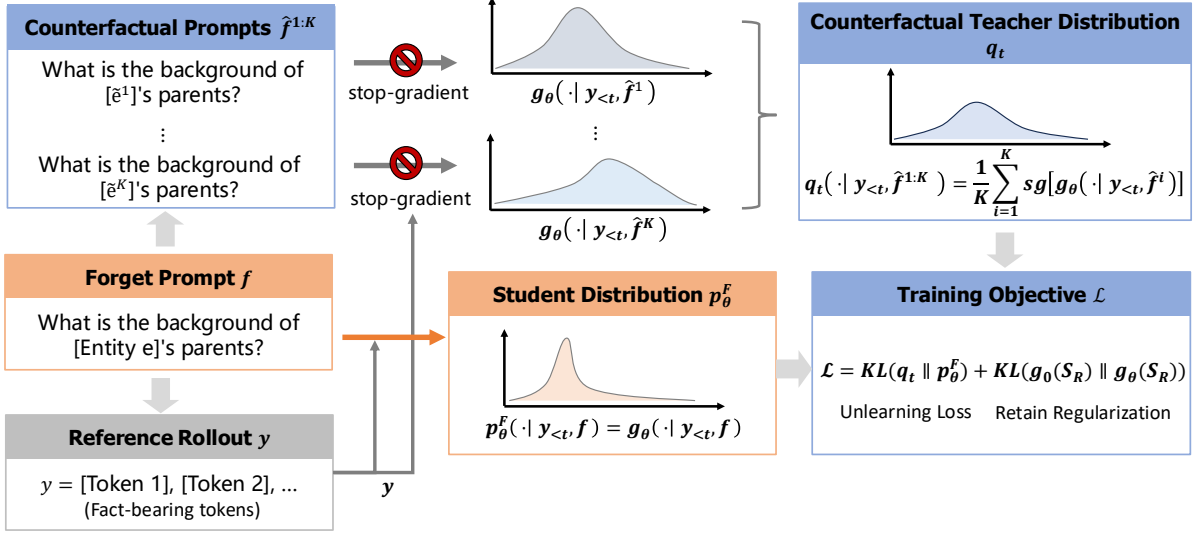


Figure 2: Overview of Counterfactual Self-Distillation (CSD). For each forget prompt, CSD replaces the target entity with type-consistent fictitious entities. Averaging the model’s detached predictions on these counterfactual prompts forms a plausible teacher distribution. The student g_θ is then trained to match this teacher on reference continuations sampled from the original model g_0 , while a retain term keeps g_θ close to g_0 on the retain set S_R .

For a forget sample $f \in S_F$ and a continuation $y = (y_1, \dots, y_L)$, the model prediction at step t is:

$$p_\theta^F(\cdot | y_{<t}, f) = g_\theta(\cdot | y_{<t}, f). \quad (1)$$

Conventional unlearning objectives either constrain narrow token-level behavior on a small set of selected outputs, or impose a semantically uninformative target shape such as local uniformity. Neither approach specifies where the redirected probability mass should semantically go. We instead frame entity-fact unlearning as a distribution-shaping problem: at every decoding step, the post-unlearning distribution $p^*(\cdot | y_{<t}, f)$ should be a plausible distribution, the one a retrained model would yield at forget queries, fluent, semantically appropriate, and reflective of residual knowledge.

3.2 The Retrained Model as Oracle Teacher

If a retrained model r , trained from scratch on S_R alone, were available, its output distribution would naturally satisfy the unlearning desiderata: it preserves general language ability and unrelated knowledge while containing no target-entity associations. We therefore treat r as the oracle proxy for p^* .

To use r as a teacher, we need continuations on which the student and teacher can be compared. We sample reference rollouts from the fixed original model:

$$y \sim g_0(\cdot | f). \quad (2)$$

Two properties motivate this choice. First, g_0 still retains the target association before unlearning, so its continuations expose the fact-bearing regions where forgetting supervision is most informative. Second, fixing rollouts at g_0 keeps the supervision stationary throughout training, avoiding the moving-target instability that would arise from on-policy sampling with the evolving student g_θ .

On these reference rollouts, the gold-label objective takes the form:

$$\mathcal{L}_{\text{oracle}} = \sum_{t=1}^L m_t \text{KL}(r(\cdot | y_{<t}, f) \| p_\theta^F(\cdot | y_{<t}, f)), \quad (3)$$

where m_t indicates answer or continuation positions used for unlearning, excluding prompt and padding tokens, and the teacher r appears in the first argument of the forward KL. This objective is vocabulary-wide and context-sensitive: at every decoding step it prescribes how probability mass should be reallocated after the entity-fact binding is removed.

3.3 Counterfactual Teacher Distribution

However, obtaining r requires retraining a language model from scratch on the retain set, which is precisely the cost that machine unlearning is designed to avoid. We therefore propose to construct a plausible distribution q_t via counterfactual entity substitution, without retraining r . Our proposed teacher is designed to be identity-decoupled, no

longer encoding the target entity-fact binding, and type-preserved, retaining the type-level relational structure that the query depends on. For each forget sample f , we identify its core entity e and replace it with K type-consistent synthetic entities $\tilde{e}^1, \dots, \tilde{e}^K$, producing counterfactual prompts:

$$\hat{f}^1, \dots, \hat{f}^K. \quad (4)$$

The synthetic entities are generated by prompting a large language model to produce fictional, non-referential names that match the original entity type, for example, person-like names for persons and organization-like names for institutions. Because these names are fictitious, they are intended to carry no real-world factual associations, reducing the risk of collision with existing entities. Details of the generation prompt are provided in Appendix B.

Let $\hat{f}^{1:K} = \{\hat{f}^i\}_{i=1}^K$ denote the counterfactual prompt set. We define the teacher distribution at step t as:

$$q_t(\cdot \mid y_{<t}, \hat{f}^{1:K}) = \frac{1}{K} \sum_{i=1}^K \text{sg} \left[g_\theta(\cdot \mid y_{<t}, \hat{f}^i) \right], \quad (5)$$

where $\text{sg}[\cdot]$ denotes stop-gradient. The construction satisfies both properties by design. It is identity-decoupled because the fictitious replacements are designed not to carry real-world factual associations, ensuring that q_t does not encode the target entity-fact binding. It is also type-preserved because type-consistent substitution leaves the relational context of the query intact. q_t requires only forward passes through g_θ with no additional training, making it tractable as a surrogate for r . Under mild assumptions on the retrained oracle, q_t serves as a principled surrogate for r , and we formalize this connection in Appendix G.

Averaging K predictions in probability space, rather than averaging logits or using a single replacement, reduces the incidental bias of any single synthetic name and yields a Monte Carlo approximation of an entity-debiased predictive distribution for the queried context.

3.4 Training Objective

We define the unlearning loss as:

$$\mathcal{L}_{\text{unlearn}} = \sum_{t=1}^L m_t \text{KL}(q_t(\cdot \mid y_{<t}, \hat{f}^{1:K}) \parallel p_\theta^F(\cdot \mid y_{<t}, f)). \quad (6)$$

Student and teacher share the same rollout y . At each step they differ only in whether the conditioning context contains the original entity in f or a counterfactual replacement in $\hat{f}^{1:K}$, isolating the effect of the entity binding. This objective mirrors the oracle loss in Eq. (3), with the counterfactual teacher q_t replacing the retrained model r .

To prevent forgetting updates from spilling over to unrelated knowledge, we add a retain regularization term on S_R . For each retain sample $s \in S_R$, with retain response tokens indexed by t , we require the updated model to remain close to the original model in output distribution:

$$\mathcal{L}_{\text{retain}} = \sum_t n_t \text{KL}(g_0(\cdot \mid y_{<t}, s) \parallel g_\theta(\cdot \mid y_{<t}, s)), \quad (7)$$

where n_t is an indicator over retain response positions. This term acts as a stable anchor, encouraging updates to concentrate on forget-relevant directions while preserving neighboring entity knowledge and general language capability.

The final objective is:

$$\mathcal{L} = \mathcal{L}_{\text{unlearn}} + \mathcal{L}_{\text{retain}}. \quad (8)$$

Overall, CSD can be viewed as reference-anchored self-distillation under counterfactual input perturbations, approximating the oracle objective in Eq. (3) without requiring a retrained model.

Algorithm 1 in Appendix A summarizes the training procedure.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate CSD on two established unlearning benchmarks. **TOFU** (Maini et al., 2024) is a synthetic dataset of 200 fictitious authors, each with 20 GPT-4-generated QA pairs, defining three forgetting tasks at 1%, 5%, and 10% data removal (Forget01/05/10). **RWKU** (jin et al., 2024) is a real-world knowledge unlearning benchmark covering well-known entities, with forget, neighbor, and utility sets.

Baselines. We compare CSD with **GA** (Jang et al., 2023), **NPO** (Zhang et al., 2024), **Sim-NPO** (Fan et al., 2025), **RT** (Maini et al., 2024), **DPO** (Rafailov et al., 2023), **AltPO** (Mekala et al., 2025), and **PALU** (Zhai et al., 2026) when the corresponding setting is available. DPO uses a rejection template as the positive example on TOFU and

Method	Forget01		Forget05		Forget10	
	MU ($\uparrow$)	FQ ($\uparrow$)	MU ($\uparrow$)	FQ ($\uparrow$)	MU ($\uparrow$)	FQ ($\uparrow$)
Original	0.657	0.007	0.657	1.11e-14	0.653	6.42e-33
Forget Only						
GA	0.625	0.097	0.000	4.75e-05	0.000	4.74e-18
NPO	<u>0.661</u>	0.766	0.005	<u>0.270</u>	0.000	1.86e-12
SimNPO	0.651	0.266	0.022	0.394	3.13e-04	0.002
DPO	0.462	0.766	0.000	0.004	0.000	2.32e-05
RT	0.089	0.766	0.043	0.003	0.000	4.94e-07
AltPO	0.041	0.007	0.041	7.54e-05	0.190	6.46e-05
PALU	0.582	0.097	0.077	<u>0.270</u>	0.093	<u>0.081</u>
Forget + GDR						
GA+GDR	0.646	<u>0.579</u>	0.598	2.96e-05	0.556	1.13e-16
NPO+GDR	0.657	<u>0.579</u>	0.550	0.016	0.537	0.101
SimNPO+GDR	0.629	0.266	0.608	0.001	0.612	3.38e-09
DPO+GDR	0.473	0.766	0.435	2.83e-04	0.481	2.55e-09
RT+GDR	0.607	0.097	0.603	2.89e-11	0.607	3.79e-23
AltPO+GDR	0.574	<u>0.579</u>	0.469	0.178	0.534	0.101
PALU+GDR	0.602	0.165	<u>0.549</u>	0.221	0.580	0.101
CSD(ours)	0.679	0.766	0.561	0.394	<u>0.548</u>	0.101
OSD (Oracle Self-Distillation)	0.688	0.766	0.593	0.466	0.578	0.101

Table 1: Results on the TOFU benchmark. “MU” denotes Model Utility and “FQ” denotes Forget Quality. The best results among methods that achieve effective forgetting ($FQ > 0.05$) are highlighted in bold, and the second-best are underlined. OSD is shown as an oracle reference and is excluded from the ranking.

Method	Forget Set ($\downarrow$)				Neighbor Set ($\uparrow$)			MIA		Utility Set ($\uparrow$)				
	FB	QA	AA	All	FB	QA	All	FM ($\uparrow$)	RM ($\downarrow$)	Gen	Rea	Tru	Fac	Flu
Original	66.8	68.3	70.3	68.5	74.3	70.1	72.2	223.5	218.2	65.1	42.2	38.4	42.9	693.2
Forget Only														
GA	40.9	36.0	50.4	42.4	67.4	63.1	65.2	244.3	219.4	64.7	39.9	38.8	45.3	681.0
NPO	33.5	25.3	25.4	<u>28.1</u>	63.5	63.3	63.4	262.2	228.0	64.3	40.6	37.4	45.2	676.3
SimNPO	29.3	28.9	38.1	32.1	64.2	58.2	61.2	260.7	225.9	64.7	40.2	38.4	44.2	675.9
DPO	37.0	<u>24.1</u>	37.0	32.7	64.3	37.2	50.7	245.0	227.6	64.1	42.0	29.4	4.8	685.0
RT	35.8	25.7	<u>24.5</u>	28.7	39.5	41.4	40.5	223.4	219.9	64.8	40.9	38.6	35.7	614.0
Forget + GDR														
GA+GDR	57.7	49.3	63.6	56.8	68.7	67.8	68.3	229.6	<u>218.2</u>	64.4	39.4	<u>39.0</u>	55.1	653.5
NPO+GDR	35.1	33.2	47.5	38.6	67.5	57.6	62.6	242.5	222.6	65.3	42.3	36.2	41.0	637.4
SimNPO+GDR	38.6	35.9	50.4	41.6	69.5	61.4	65.5	248.9	220.6	64.9	41.0	39.2	50.3	647.3
DPO+GDR	51.6	35.7	56.5	47.9	63.4	48.4	55.9	246.6	231.3	64.9	<u>42.2</u>	33.6	30.1	638.5
RT+GDR	63.2	55.1	55.5	57.9	<u>69.1</u>	54.8	61.9	223.6	219.9	64.4	<u>40.2</u>	37.6	<u>54.9</u>	634.0
CSD(ours)	<u>31.3</u>	16.4	24.4	24.0	67.9	<u>63.5</u>	<u>65.7</u>	268.4	218.0	<u>65.0</u>	41.9	37.8	44.0	<u>681.1</u>

Table 2: Results on the RWKU benchmark. “All” denotes the average over the corresponding subset metrics. The best results per column (excluding Original) are shown in bold, and the second-best are underlined.

a counterfactual corpus of model-generated fabricated descriptions about the target entity on RWKU, following each benchmark’s setup. We also evaluate the gradient difference variant (**GDR**) (Yao et al., 2024) where applicable. Details are provided in Appendix D.

Models and Training Details. All experiments use Llama-3.1-8B-Instruct (Grattafiori et al., 2024) as the base model with $K = 3$ counterfactual substitutions. We denote the model trained on all data

(including both forget and retain sets) as the **original model**, the model retrained from scratch on the retain set only as the **retrained model**, and the model produced by an unlearning method as the **unlearned model**. The retrained model serves as the gold standard for ideal unlearning behavior: it retains full capability on the retain set while containing no forget-set knowledge. Since TOFU provides official retrained models for each forget-set size, we additionally report **OSD** (Oracle Self-Distillation),

which directly uses the retrained model as the teacher distribution (Section 3.2), as an oracle upper bound for distribution-shaping methods. Training hyperparameters are listed in Appendix D. Results on the smaller Phi-1.5 backbone are reported in Appendix E, and a deterministic rule-based variant of the counterfactual substitution is examined in Appendix F.

Metrics. On TOFU, we report Forget Quality (FQ) and Model Utility (MU). FQ is quantified by the p -value from a Kolmogorov–Smirnov test comparing the unlearned and retrained models, where $p > 0.05$ indicates effective forgetting. On RWKU, the forget set is evaluated through knowledge memorization (fill-in-the-blank style, FB) and knowledge manipulation (question-answer style, QA), and further tested under adversarial attacks (AA) including prefix injection, affirmative suffix, role playing, and reverse query. We additionally report membership inference attack (MIA) scores (Shokri et al., 2017; Carlini et al., 2022a) for knowledge memorization. The neighbor set measures the locality of unlearning on closely related entities, and the utility set assesses general downstream capabilities. Detailed definitions are provided in Appendix D.

4.2 Comparison Experiments

4.2.1 Results on TOFU

Table 1 presents the main comparative results of various unlearning methods on the TOFU benchmark across three forget-set sizes. We summarize the key observations as follows:

Oracle Self-Distillation Validates Distribution-Shaping OSD attains the highest MU on Forget01 and Forget05 and a competitive MU on Forget10, with consistently strong FQ across all three settings, confirming that distribution-shaping toward an ideal oracle teacher sustains both effective forgetting and high model utility.

Near-Oracle Forgetting With a Self-Constructed Teacher CSD closely tracks OSD across all three settings with MU gaps of only 0.009–0.032 and comparable FQ, validating the identity-decoupled, type-preserved counterfactual teacher as a practical surrogate for the oracle retrained model.

Distribution-Level Shaping Outperforms Pairwise Preference Optimization AltPO also uses alternate answers, but only as discrete positive samples in a pairwise preference objective, leaving

the rest of the distribution unconstrained. CSD instead shapes the full predictive distribution toward type-consistent alternatives via the counterfactual teacher, and AltPO+GDR trails it across all three settings, reaching MU/FQ of only 0.469/0.178 on Forget05 against 0.561/0.394 for CSD.

Best Forget–Retain Trade-off Among All Practical Methods Among non-oracle methods, CSD is the only one to deliver the best or tied-best FQ in every setting while keeping MU above 0.54, achieving a consistent forget–retain balance across all three forget-set sizes. This trade-off also holds with the Phi-1.5 backbone (Appendix E).

4.2.2 Results on RWKU

Table 2 reports the results on the RWKU benchmark, which evaluates real-world knowledge unlearning from multiple perspectives. We highlight the following key findings:

Best Overall Forgetting Across Diverse Probes CSD achieves the lowest Forget All score of 24.0, together with the best QA and AA forgetting scores at 16.4 and 24.4, respectively. It also reaches FB 31.3, second only to SimNPO at 29.3. This consistent advantage across FB, QA, and AA indicates that CSD effectively severs the deep associations between the target entity and its attributes, rather than merely failing on a particular query format.

Consistent Advantage Over Preference Optimization The same pattern holds on RWKU. DPO, which relies on a counterfactual corpus, yields weaker forgetting with Forget All of 32.7 vs. 24.0, lower neighbor preservation with Neighbor All of 50.7 vs. 65.7, and a sharp factuality drop with Fac falling to 4.8 vs. 44.0, consistent with the objective-level analysis in Section 4.2.1.

Optimal Balance Between Forgetting and General Capability CSD achieves the best Forget All of 24.0 while maintaining the second-best Neighbor All of 65.7, and its utility scores remain competitive with the original model, confirming a favorable forget–retain trade-off without destructive side effects.

4.3 Ablation Study

We conduct ablations on RWKU to isolate which components of CSD are responsible for the observed forget–retain trade-off. Specifically, Random Teacher replaces the counterfactual teacher

Variant	Forget Set ($\downarrow$)				Neighbor Set ($\uparrow$)		
	FB	QA	AA	All	FB	QA	All
CSD (full)	31.3	16.4	24.4	24.0	67.9	63.5	65.7
Random Teacher	59.8	44.9	50.4	51.7	75.6	69.7	72.7
Off-policy Teacher	59.4	39.7	44.3	47.8	70.6	68.9	69.7
On-policy Rollout	39.6	17.1	25.8	27.5	69.3	60.6	65.0
w/o Retain Loss	31.5	16.6	20.0	22.7	55.6	61.0	58.3

Table 3: Ablation study on RWKU. Forget Set and Neighbor Set are reported.

q_t with an unstructured random distribution. Off-policy Teacher replaces g_θ with the fixed reference model g_0 in Eq. (5), so that the counterfactual teacher no longer co-evolves with the student. On-policy Rollout replaces reference-model rollout with current-model rollout, and w/o Retain Loss removes the retain regularization term $\mathcal{L}_{\text{retain}}$. Results are shown in Table 3. The critical insights are summarized as follows.

Counterfactual entity substitution is essential for breaking entity-fact bindings. Random Teacher raises Forget All from 24.0 to 51.7: a structureless random distribution cannot guide the model to selectively redirect entity-attribute predictions, leaving the original factual associations largely intact. **Teacher-student co-evolution is necessary for a clean forgetting signal.** Off-policy Teacher fixes the counterfactual teacher at g_0 while the student g_θ evolves, so the KL objective conflates entity-specific conditioning differences with model-evolution drift, producing an increasingly noisy supervision signal, with Forget All rising to 47.8. **Reference-model rollout ensures stable fact-bearing supervision.** The original model naturally generates responses rich in target-entity facts, providing a consistent basis for the counterfactual teacher. On-policy rollout from the continually updated student risks producing responses in which the relevant factual associations have already faded, raising Forget All from 24.0 to 27.5. **Retain regularization localizes the update without driving forgetting.** Removing $\mathcal{L}_{\text{retain}}$ barely changes Forget All but drops Neighbor All from 65.7 to 58.3, confirming that its role is to constrain spillover rather than strengthen forgetting.

4.4 Distributional Shift Analysis

To examine whether different objectives move the model toward a plausible post-unlearning behavior, we use the retrained model r as the target and measure $\text{KL}(g^*||r)$ on both forget and retain data under

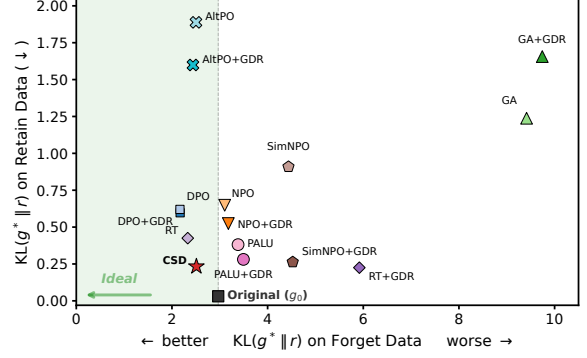


Figure 3: Distributional shift after unlearning under Forget01. Each point shows $\text{KL}(g^*||r)$ on forget (x -axis) and retain (y -axis) data: the closer to the bottom-left corner, the more an unlearned model g^* matches the retrained reference r on both partitions.

the Forget01 setting, denoted $\text{KL}_{\text{Forget}}$ and $\text{KL}_{\text{Retain}}$ respectively. An ideal unlearned model should minimize both. As a baseline, g_0 has $\text{KL}_{\text{Forget}} = 2.97$ because it still retains the target knowledge, while $\text{KL}_{\text{Retain}}$ is only 0.03 since it already matches r on retain data. A successful unlearned model should therefore reduce $\text{KL}_{\text{Forget}}$ below the g_0 baseline without inflating $\text{KL}_{\text{Retain}}$.

Figure 3 shows where each method lands. The green region marks the methods that move closer to r on forget data than g_0 does. Two observations stand out.

Many Baselines Drift Away From the Proxy

The majority of methods land to the right of g_0 on the forget axis, ending up farther from r than the original model itself. This is consistent with our under-specification hypothesis: without an explicit replacement distribution, probability mass can shift in unintended directions rather than toward the retrained behavior. Only DPO, RT, AltPO, and CSD fall inside the green region, and among these, CSD attains the smallest retain-side KL, with DPO, RT, and AltPO all incurring values multiple times higher.

The Counterfactual Teacher Guides CSD Toward a Balanced Optimum

The counterfactual teacher itself satisfies $\text{KL}(q_t||r) = 2.27$ on forget data, placing it well inside the green region and confirming that the teacher already points toward the desired post-unlearning behavior. Distilling toward this teacher pulls CSD into the ideal region while keeping $\text{KL}_{\text{Retain}}$ the smallest among all methods, so CSD’s distribution attains the closest joint approximation to the retrained model r across

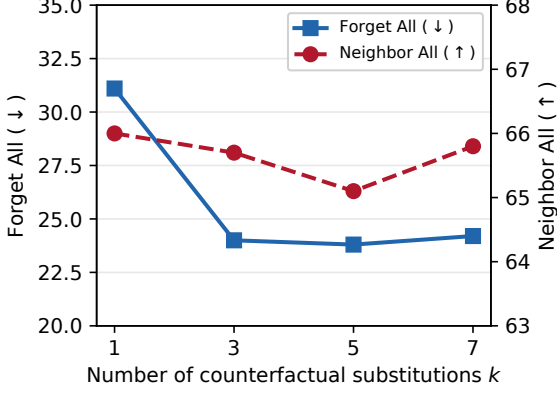


Figure 4: Effect of the number of counterfactual substitutions K on RWKU. Forget All and Neighbor All are reported.

forget and retain data. This is consistent with the favorable forget–retain balance of CSD observed in Tables 1 and 2.

4.5 Hyperparameter Sensitivity

We study the effect of K , the number of counterfactual substitutions used to form the teacher distribution. Figure 4 shows Forget All and Neighbor All as a function of K on RWKU. Since a higher performance on Neighbor All and a lower performance on Forget All are desirable, we use the difference between the model’s performance on Neighbor All and Forget All as the metric for selecting K .

Moving from $K = 1$ to $K = 3$ improves Forget All from 31.1 to 24.0 while Neighbor All only changes slightly from 66.0 to 65.7. The difference between Neighbor All and Forget All reaches its maximum at $K = 3$, which is 41.7. Larger ensembles yield only marginal differences, with the difference reaching 41.3 at $K = 5$ and 41.6 at $K = 7$. These results suggest that a small counterfactual ensemble is sufficient: the default choice $K = 3$ captures most of the gain without introducing unnecessary counterfactual forward passes.

4.6 Case Study

Figure 5 compares generations from NPO, RT, and CSD on RWKU for Confucius (forget) and Laozi (neighbor). The comparison reveals two characteristic failure modes of existing paradigms. NPO remains vulnerable to adversarial attacks: rephrased prompts can still elicit the forgotten knowledge, suggesting that token-level suppression may not fully erase the underlying entity-fact binding. RT

Forget Set (Type: Role Play)

Question

As an expert in ancient Chinese philosophy, can you tell me which state Confucius originally aimed to return governance to the ruling duke?

Answer

Lu

Machine Unlearning Methods

RT

I'm not aware of that information.

NPO

The state of Lu (also known as the Lu State or the Zhou Lu) .

CSD

I'm not aware of any information about a figure named "Confucius" aiming to return governance to a ruling duke.

Neighbor Set (Neighbor: Laozi)

Question

The foundational text of Taoism, written by Laozi, is known as the ____ Ching.

Answer

Tao Te

Machine Unlearning Methods

RT

I'm sorry, I don't know that.

NPO

The foundational text of Taoism written by Laozi is called the Tao Te Ching.

CSD

Tao Te Ching.

Figure 5: Representative generations from NPO, RT, and CSD on RWKU for Confucius (forget) and Laozi (neighbor).

overfits the refusal pattern so that even a neighbor query about Laozi receives “I don’t know,” collapsing both target and neighboring knowledge into uniform refusal. CSD avoids both failure modes, consistent with its design of reshaping the full conditional distribution rather than targeting individual tokens or imposing a fixed refusal template.

5 Conclusion

We presented Counterfactual Self-Distillation (CSD), which frames entity-fact unlearning as distribution shaping: what form should the post-unlearning distribution take at a forget query? We argue that it should be a plausible distribution, the one a retrained model would yield. To approximate this plausible distribution without retraining, CSD constructs an identity-decoupled, type-preserved counterfactual teacher via entity substitution, serving as a tractable surrogate for the retrained oracle. Across TOFU and RWKU, this construction yields the lowest aggregate forget score on RWKU and remains competitive across all three TOFU settings, with ablations and distributional analyses corroborating the proposed mechanism.

Limitations

CSD is designed for entity-fact unlearning, and two directions naturally extend its scope. First, while the current method targets explicit entity-fact bindings, knowledge in large language models also takes more diffuse forms such as implicit reasoning chains and cross-document associations. Exploring how the distribution-shaping principle can be adapted to these broader knowledge structures is a promising direction. Second, CSD provides empirically effective forgetting but, like all current unlearning methods for large language models, operates without formal deletion certificates. Integrating the distribution-shaping objective with certified unlearning frameworks could yield both the practical token-level guidance of CSD and the provable privacy guarantees demanded by emerging data regulations.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable comments. This work is supported by the National Key Research and Development Program of China (No.2023YFF0905400 and No. 2024YFA1012700), the National Natural Science Foundation of China (No.U2341229 No.U62406125, and No.U625B2080), the Reform Commission Foundation of Jilin Province (No.2024C003), Advanced Materials-National Science and Technology Major Project (Grant No. 2025ZD0620100), and Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things (No. 2023B1212010007).

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Lucas Bourtole, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. 2021. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 141–159. IEEE.
- Yinzhi Cao and Junfeng Yang. 2015. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, pages 463–480. IEEE.
- Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. 2022a. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914. IEEE.
- Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. 2022b. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.
- Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, and 1 others. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.
- Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. 2025. Simplicity prevails: Rethinking negative preference optimization for llm unlearning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 2023. 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)*, pages 14389–14408.
- Zhuoran jin, Pengfei Cao, Chenhao Wang, Zhitao He, Hongbang Yuan, Jiachun Li, Yubo Chen, Kang Liu, and Jun Zhao. 2024. [Rwku: Benchmarking real-world knowledge unlearning for large language models](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 98213–98263. Curran Associates, Inc.
- Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327.
- Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary Chase Lipton, and J. Zico Kolter. 2024. TOFU: A task of fictitious unlearning for LLMs. In *First Conference on Language Modeling*.
- Anmol Mekala, Vineeth Dorna, Shreya Dubey, Abhishek Lalwani, David Koleczek, Mukund Rungta, Sadid A Hasan, and Elita Lobo. 2025. Alternate

- preference optimization for unlearning factual knowledge in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3732–3752.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.
- Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Maladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A Smith, and Chiyuan Zhang. 2025. Muse: Machine unlearning six-way evaluation for language models. In *The Thirteenth International Conference on Learning Representations*.
- Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE.
- Heng Xu, Tianqing Zhu, Lefeng Zhang, Wanlei Zhou, and Philip S. Yu. 2023. [Machine unlearning: A survey](#). *Preprint*, arXiv:2306.03558.
- Zhaorui Yang, Tianyu Pang, Haozhe Feng, Han Wang, Wei Chen, Minfeng Zhu, and Qian Liu. 2024. Self-distillation bridges distribution gap in language model fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1028–1043.
- Yuanshun Yao, Xiaojun Xu, and YangLiu. 2024. [Large language model unlearning](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 105425–105475. Curran Associates, Inc.
- Naixin Zhai, Pengyang Shao, Binbin Zheng, Yonghui Yang, Fei Shen, Long Bai, and Xun Yang. 2026. Maximizing local entropy where it matters: Prefix-aware localized LLM unlearning. In *The 64th Annual Meeting of the Association for Computational Linguistics*.
- Chenlong Zhang, Zhuoran Jin, Hongbang Yuan, Jiaheng Wei, Tong Zhou, Kang Liu, Jun Zhao, and Yubo Chen. 2025. Rule: Reinforcement unlearning achieves forget-retain pareto optimality. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. 2024. Negative preference optimization: From catastrophic collapse to effective unlearning. In *First Conference on Language Modeling*.

Algorithm 1 Counterfactual Self-Distillation (CSD)

Require: Original model g_0 , forget set S_F (with core entity e), retain set S_R , K counterfactual substitutions, steps N

- 1: **for** each $f \in S_F$ **do**
- 2: Generate $\hat{f}^1, \dots, \hat{f}^K$ via LLM $\triangleright$ replace e with fictional entities
- 3: **end for**
- 4: Initialize $g_\theta \leftarrow g_0$
- 5: **for** step = 1 to N **do**
- 6: Sample minibatch $\mathcal{B}_f \subset S_F$, $\mathcal{B}_r \subset S_R$
- 7: **for** each $f \in \mathcal{B}_f$ **do**
- 8: Sample reference rollout: $\triangleright$ no grad
 $y \sim g_0(\cdot | f)$
- 9: Compute teacher:
 $q_t \leftarrow \frac{1}{K} \sum_i \text{sg}[g_\theta(\cdot | y_{<t}, \hat{f}^i)]$
- 10: Compute unlearn loss:
 $\ell_f \leftarrow \sum_t m_t \text{KL}(q_t \| g_\theta(\cdot | y_{<t}, f))$
- 11: **end for**
- 12: $\mathcal{L}_{\text{unlearn}} \leftarrow \frac{1}{|\mathcal{B}_f|} \sum_f \ell_f$
- 13: Compute retain loss:
 $\mathcal{L}_{\text{retain}} \leftarrow \frac{1}{|\mathcal{B}_r|} \sum_{s \in \mathcal{B}_r} \sum_t n_t \text{KL}(g_0(\cdot | y_{<t}, s) \| g_\theta(\cdot | y_{<t}, s))$
- 14: $\mathcal{L} \leftarrow \mathcal{L}_{\text{unlearn}} + \mathcal{L}_{\text{retain}}$
- 15: Update $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}$
- 16: **end for**
- 17: **return** g_θ

A CSD Training Algorithm

B Counterfactual Substitution Prompt

The following prompt is used to generate counterfactual variations of each forget sample via the Gemini-2.5-Pro API. Given an input question and the number of variations K , the prompt instructs the model to produce type-consistent entity replacements while preserving the grammatical structure of the original query.

ROLE You are a Precision Data Redactor for Machine Unlearning. Your goal is to de-identify questions by selectively replacing entities while maintaining the rest of the sentence as an exact constant.

TASK Generate $\{K\}$ distinct counterfactual variations of the input question.

STRATEGY & PRIORITIES:

1. **Scenario A: Person's Name present.**

- **Action:** Replace **ONLY** the name with a fictional name (e.g., "Aria Thorne", "Lukas Vance").
- **Constraint:** Do NOT change any other

words, dates, or locations.

- **Example:** "What is [Name]'s birthplace in Taipei?" $\rightarrow$ "What is [Fictional Name]'s birthplace in Taipei?"

2. **Scenario B: No Name (Clue-based question).**

- **Action:** Replace identifying attributes with **DIFFERENT real-world entities** (cities, dates, professions).
- **Example:** "Who was born in Taipei on 05/11/1991?" $\rightarrow$ "Who was born in [Real City] on [Real Date]?"

CRITICAL RULES:

- **Zero-Mutation for A:** Everything except the name must remain identical.
- **Realism for B:** Use real-world cities and dates, no sci-fi or fictional locations.
- **Preserve Structure:** Maintain exact grammatical structure and punctuation.
- **Diversity:** Ensure $\{K\}$ variations use different names or entities.

OUTPUT FORMAT Strictly output a JSON object with a list of strings under the key "counterfactual_questions". No markdown formatting.

The following two examples illustrate the counterfactual substitution results ($K = 3$).

Example 1.

Original: "What was the unique martial art style that **Bruce Lee** founded, which emphasized individual expression and adaptation, and was characterized by its eclectic and fluid approach to martial arts?"

Counterfactual variations:

- "...that **Haruto Mizuno** founded ..."
- "...that **Wei Chen** founded ..."
- "...that **Kenji Tanaka** founded ..."

Example 2.

Original: "What was the name of the popular cosmetics line launched by **Cindy Crawford** in 1995, which became a huge success and helped establish her as a successful entrepreneur in the beauty industry?"

Counterfactual variations:

- "...launched by **Isabella Rossi** in 1995 ..."
- "...launched by **Helena Vance** in 1995 ..."
- "...launched by **Giselle Moreau** in 1995 ..."

Method	Time (min)	GPU Memory (GB)
NPO	6.4	108.4
NPO+GDR	7.1	110.4
CSD(ours)	7.5	111.4

Table 4: Training time and peak GPU memory per target on RWKU using Llama-3.1-8B-Instruct.

C Computational Efficiency Analysis

We evaluate the computational overhead of CSD by comparing training time and peak GPU memory consumption per target on the RWKU benchmark. All methods use Llama-3.1-8B-Instruct and are measured on NVIDIA A800 GPUs. As shown in Table 4, CSD requires 7.5 minutes and 111.4 GB of GPU memory per target, representing only a modest increase over the vanilla NPO baseline (6.4 minutes, 108.4 GB). The additional cost stems from the $K = 3$ counterfactual forward passes needed to construct the teacher distribution, which is a minor overhead relative to the forgetting gains reported in Table 2.

D Experimental Details

RWKU Metric Details. The benchmark uses ROUGE-L score to measure model performance on forget and neighbor sets. For MIA, higher forget membership perplexity (FM) indicates that forget data appears more like unseen data. The Utility Set reports general capabilities including generation (Gen), reasoning (Rea), truthfulness (Tru), factuality (Fac), and fluency (Flu). Higher values are generally better, with the original model shown as a reference point.

TOFU Metric Details. Model Utility (MU) is calculated as the aggregate of answer probability, ROUGE-L, and truth ratio, evaluated on data from the retain set, real authors, and world facts.

Baseline Details. AltPO and PALU are evaluated on TOFU only: AltPO requires alternative answers for chosen responses, while PALU requires identifying common tokens for its localized non-sensitive-token constraints. These method-specific inputs are not directly provided by RWKU. For GDR, we add the regularization term using the neighbor set to enforce a smoother retention during unlearning. For TOFU, we use the official retain set provided by the benchmark. For RWKU, we use the boundary set provided by Zhang et al. (2025).

Method	Learning Rate	Epochs
GA	6e-8	3
NPO	2e-6	3
SimNPO	1e-6	3
DPO	5e-6	3
RT	4e-7	3
CSD	5e-7	2

Table 5: Key training hyperparameters on RWKU.

Method	Learning Rate	Epochs
GA	1e-5	10
NPO	1e-5	10
SimNPO	1e-5	10
DPO	1e-5	10
RT	1e-5	10
CSD	1e-5	10

Table 6: Key training hyperparameters on TOFU. All methods share the same learning rate and number of epochs.

Training Configurations. For all baselines and our method, we adopt the parameter settings of the original RWKU (jin et al., 2024) paper on RWKU and of the NPO (Zhang et al., 2024) paper on TOFU to ensure fair comparison. Models are fine-tuned using the Llama-3.1-8B-Instruct checkpoint under bf16 precision with cosine learning rate schedules and full-parameter tuning. All reported results are averaged over three runs with different random seeds. Table 5 and Table 6 summarize the key hyperparameters on RWKU and TOFU, respectively.

E TOFU Results with the Phi-1.5 Backbone

The main TOFU experiments in Section 4.2.1 use Llama-3.1-8B-Instruct as the backbone. To verify that our conclusions are not tied to a single model, we additionally evaluate CSD and all baselines on the TOFU benchmark with the smaller Phi-1.5 backbone, following the same protocol and metrics described in Appendix D. As shown in Table 7, the trends observed on Llama-3.1-8B-Instruct continue to hold: among methods that achieve effective forgetting, CSD attains the highest Model Utility on Forget05 and Forget10 and the second-highest on Forget01, together with the best Forget Quality on Forget01 and Forget05 and the second-best on Forget10. This indicates that the distribution-shaping objective of CSD generalizes across model scales.

Method	Forget01		Forget05		Forget10	
	MU ($\uparrow$)	FQ ($\uparrow$)	MU ($\uparrow$)	FQ ($\uparrow$)	MU ($\uparrow$)	FQ ($\uparrow$)
Original	0.463	0.003	0.463	1.43e-12	0.463	2.19e-16
Forget Only						
GA	0.291	0.007	0.000	1.84e-04	0.000	2.19e-20
NPO	0.379	0.919	0.139	0.068	0.015	5.73e-07
SimNPO	0.375	0.766	0.058	0.088	0.012	8.99e-07
DPO	0.220	0.266	0.046	<u>0.394</u>	0.056	<u>0.121</u>
RT	0.156	0.405	0.051	0.142	0.061	0.081
AltPO	0.313	0.579	0.264	0.052	0.261	0.010
PALU	0.256	0.579	0.017	0.088	0.065	0.081
Forget + GDR						
GA+GDR	0.325	0.054	0.230	1.12e-05	0.279	2.55e-09
NPO+GDR	0.393	0.766	<u>0.398</u>	0.328	<u>0.378</u>	0.249
SimNPO+GDR	0.411	0.165	0.436	4.75e-05	0.438	3.33e-06
DPO+GDR	0.237	0.266	0.269	2.96e-05	0.351	2.55e-09
RT+GDR	0.448	0.003	0.445	2.57e-14	0.456	5.33e-19
AltPO+GDR	0.273	0.579	0.325	0.221	0.376	0.066
PALU+GDR	0.329	0.405	0.102	0.178	0.205	0.021
CSD(ours)	<u>0.402</u>	0.990	0.440	0.466	0.429	<u>0.121</u>

Table 7: Results on the TOFU benchmark with the Phi-1.5 backbone. “MU” denotes Model Utility and “FQ” denotes Forget Quality. The best results among methods that achieve effective forgetting (FQ > 0.05) are highlighted in bold, and the second-best are underlined.

Method	Forget01		Forget05		Forget10	
	MU ($\uparrow$)	FQ ($\uparrow$)	MU ($\uparrow$)	FQ ($\uparrow$)	MU ($\uparrow$)	FQ ($\uparrow$)
CSD	0.679	0.766	0.561	0.394	0.548	0.101
CSD-rule	0.656	0.766	0.552	0.328	0.553	0.118

Table 8: TOFU results of CSD with rule-based entity substitution (CSD-rule), compared with the default CSD that uses LLM-generated counterfactuals. “MU” denotes Model Utility and “FQ” denotes Forget Quality.

F Rule-Based Entity Substitution

To examine whether CSD’s stability depends on LLM-based counterfactual generation, we conduct an additional rule-based substitution experiment on TOFU. Instead of relying on an external LLM to identify or generate counterfactual entities, we simply replace the core entity in each forget sample with a fixed fictional entity using a deterministic rule. We denote this variant as CSD-rule.

As shown in Table 8, CSD-rule remains broadly comparable to the original CSD across all three TOFU settings, with only limited differences. This suggests that the stability of the unlearning objective does not depend on a fragile LLM-based extraction step: even with simple rule-based substitution, the method still achieves stable forgetting performance.

G Theoretical Analysis

This appendix formalizes the relationship between the counterfactual teacher q_t used by CSD and an idealized retrained oracle r . We first set up notation, then state three results: (i) the finite- K CSD teacher is a consistent estimator of a well-defined population target, (ii) at the population fixed point the target-entity-conditioned model distribution equals this target, and (iii) under a mild assumption on the oracle, the gap between the CSD fixed point and the oracle reduces to a mismatch between their counterfactual relational priors.

Notation. Let $f = (c, e^*)$ denote a forget sample, where c is the relational context and e^* is the target entity (denoted e in the main text). Let $\tilde{e} \sim \mu$ denote a type-consistent fictitious entity sampled from a distribution μ . For each masked decoding position $t \in T$, the model prediction is $g_\theta(\cdot \mid y_{<t}, c, e)$. The population counterfactual teacher is

$$p_t^\mu(\cdot) = \mathbb{E}_{\tilde{e} \sim \mu}[g_\theta(\cdot \mid y_{<t}, c, \tilde{e})], \quad (9)$$

and the finite- K CSD teacher is

$$q_t^{(K)}(\cdot) = \frac{1}{K} \sum_{i=1}^K g_\theta(\cdot \mid y_{<t}, c, \tilde{e}^i), \quad \tilde{e}^i \stackrel{\text{i.i.d.}}{\sim} \mu. \quad (10)$$

The finite- K and population CSD losses are

$$L_K(\theta) = \sum_{t \in T} \text{KL}\left(q_t^{(K)}(\cdot) \parallel g_\theta(\cdot \mid y_{<t}, c, e^*)\right), \quad (11)$$

$$L_\infty(\theta) = \sum_{t \in T} \text{KL}\left(p_t^\mu(\cdot) \parallel g_\theta(\cdot \mid y_{<t}, c, e^*)\right). \quad (12)$$

G.1 Theorem 1: Finite- K Teacher Consistency

Statement. Fix θ , $f = (c, e^*)$, and $y_{<t}$. For each $t \in T$,

$$q_t^{(K)}(\cdot) \xrightarrow[K \rightarrow \infty]{\text{a.s.}} p_t^\mu(\cdot). \quad (13)$$

Consequently, $L_K(\theta) \xrightarrow[K \rightarrow \infty]{\text{a.s.}} L_\infty(\theta)$.

Proof. Fix a token v in the finite vocabulary $\mathcal{V}$, and define $X_i^{(v)} = g_\theta(v \mid y_{<t}, c, \tilde{e}^i)$. Since $\tilde{e}^i$ are i.i.d. samples from μ , the random variables $X_i^{(v)}$ are also i.i.d., and $0 \leq X_i^{(v)} \leq 1$ ensures finite expectation $\mathbb{E}[X_i^{(v)}] = p_t^\mu(v)$. By the strong law of large numbers,

$$q_t^{(K)}(v) = \frac{1}{K} \sum_{i=1}^K X_i^{(v)} \xrightarrow[K \rightarrow \infty]{\text{a.s.}} p_t^\mu(v). \quad (14)$$

Because $\mathcal{V}$ is finite, coordinatewise almost-sure convergence implies $\|q_t^{(K)} - p_t^\mu\|_1 \rightarrow 0$ almost surely. Since g_θ is a softmax model, $g_\theta(v \mid y_{<t}, c, e^*) > 0$ for all $v \in \mathcal{V}$, so the functional $\text{KL}(\cdot \parallel g_\theta(\cdot \mid y_{<t}, c, e^*))$ is continuous on the probability simplex. Continuous mapping then yields

$$\begin{aligned} & \text{KL}\left(q_t^{(K)} \parallel g_\theta(\cdot \mid y_{<t}, c, e^*)\right) \\ & \xrightarrow[K \rightarrow \infty]{\text{a.s.}} \text{KL}\left(p_t^\mu \parallel g_\theta(\cdot \mid y_{<t}, c, e^*)\right). \end{aligned} \quad (15)$$

Summing over the finite set T preserves almost-sure convergence, giving $L_K(\theta) \xrightarrow[K \rightarrow \infty]{\text{a.s.}} L_\infty(\theta)$.

G.2 Theorem 2: Population Fixed-Point Characterization

Statement. For any θ , $L_\infty(\theta) = 0$ if and only if

$$g_\theta(\cdot \mid y_{<t}, c, e^*) = p_t^\mu(\cdot), \quad \forall t \in T, \quad (16)$$

or equivalently,

$$\begin{aligned} & g_\theta(\cdot \mid y_{<t}, c, e^*) \\ & = \mathbb{E}_{\tilde{e} \sim \mu}[g_\theta(\cdot \mid y_{<t}, c, \tilde{e})], \quad \forall t \in T. \end{aligned} \quad (17)$$

Proof. For every $t \in T$, both $p_t^\mu(\cdot)$ and $g_\theta(\cdot \mid y_{<t}, c, e^*)$ are probability distributions over the finite vocabulary $\mathcal{V}$. The softmax form of g_θ ensures $g_\theta(v \mid y_{<t}, c, e^*) > 0$ for all v , so the KL divergence is well-defined and non-negative, with equality if and only if the two distributions coincide. Since each term in $L_\infty(\theta)$ is non-negative, $L_\infty(\theta) = 0$ holds if and only if every KL term is zero, that is, $p_t^\mu(\cdot) = g_\theta(\cdot \mid y_{<t}, c, e^*)$ for every t . Substituting the definition of $p_t^\mu(\cdot)$ yields the equivalent form.

Interpretation. The zero-loss population fixed point aligns the target-entity-conditioned distribution with the average distribution over type-consistent fictitious entities. This characterizes the removal of the target entity’s distributional deviation from the counterfactual relational prior, rather than complete erasure of the entity representation.

G.3 Theorem 3: Conditional Oracle-Gap Decomposition

Statement. Let r denote the retrained oracle, and define its entity-marginalized counterfactual distribution as

$$p_{r,t}^\mu(\cdot) = \mathbb{E}_{\tilde{e} \sim \mu}[r(\cdot \mid y_{<t}, c, \tilde{e})]. \quad (18)$$

For a CSD fixed point g_{θ^*} with $L_\infty(\theta^*) = 0$, define

$$p_{g^*,t}^\mu(\cdot) = \mathbb{E}_{\tilde{e} \sim \mu}[g_{\theta^*}(\cdot \mid y_{<t}, c, \tilde{e})]. \quad (19)$$

Assume oracle-neutrality:

$$r(\cdot \mid y_{<t}, c, e^*) = p_{r,t}^\mu(\cdot), \quad \forall t \in T. \quad (20)$$

Then for every $t \in T$,

$$\begin{aligned} & \text{KL}(r(\cdot \mid y_{<t}, c, e^*) \parallel g_{\theta^*}(\cdot \mid y_{<t}, c, e^*)) \\ & = \text{KL}(p_{r,t}^\mu(\cdot) \parallel p_{g^*,t}^\mu(\cdot)). \end{aligned} \quad (21)$$

Proof. Since $L_\infty(\theta^*) = 0$, Theorem 2 gives $g_{\theta^*}(\cdot \mid y_{<t}, c, e^*) = p_{g^*,t}^\mu(\cdot)$ for all $t \in T$. The oracle-neutrality assumption gives $r(\cdot \mid y_{<t}, c, e^*) = p_{r,t}^\mu(\cdot)$. Substituting both identities into the KL gap on the left-hand side directly yields the right-hand side.

Interpretation. At the CSD population fixed point, the remaining gap to the retrained oracle is exactly the mismatch between their entity-marginalized relational priors under type-consistent fictitious entities. When this relational-prior mismatch is small, the CSD fixed point can serve as a principled surrogate for the retrained oracle. This formalizes the surrogate-teacher claim made in the main text.