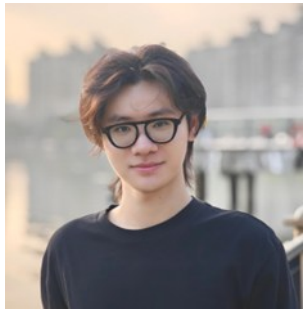
LLM Unlearning with LLM Beliefs

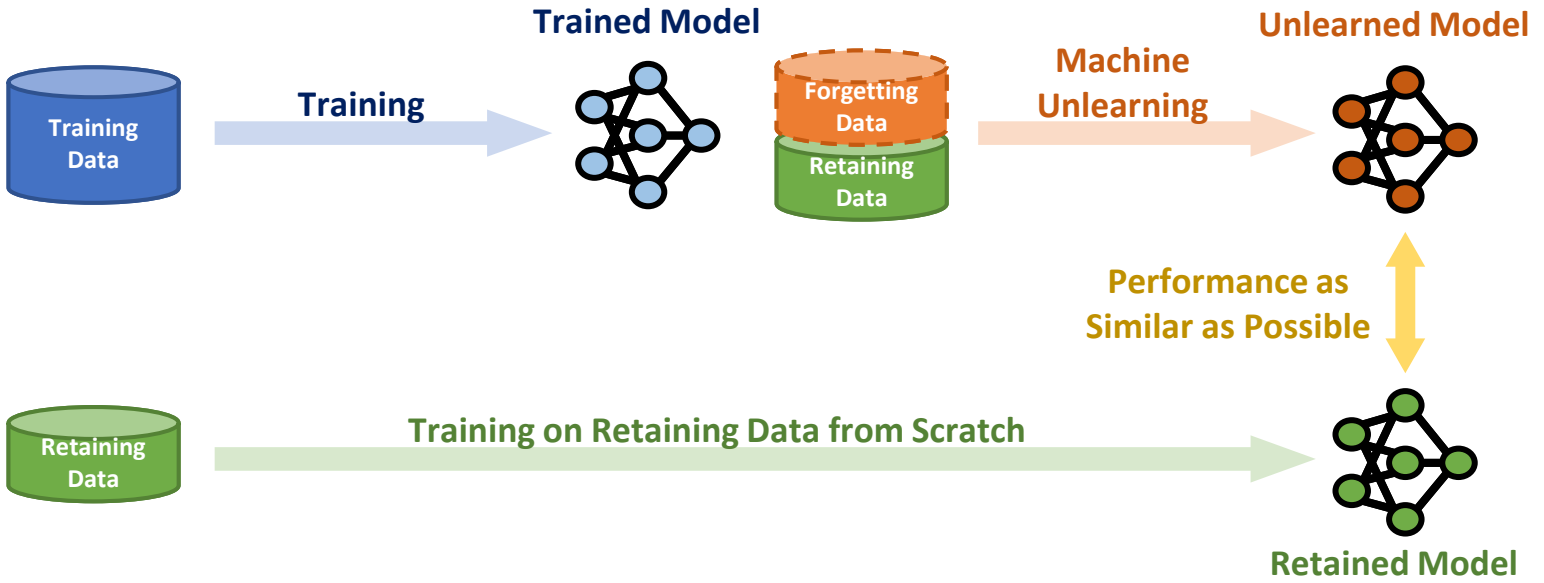
Kemou Li, Qizhou Wang, Yue Wang, Fengpeng Li, Jun Liu, Bo Han, Jiantao Zhou



Li et al. LLM Unlearning with LLM Beliefs. In *ICLR*, 2026.

Background | Machine Unlearning for LLMs

- ✓ **Machine unlearning** aims to remove the influence of the **forgetting data** from a trained model, such that it behaves similarly to a model (termed Retrained) retrained from scratch on the **retaining data**.



Li et al. LLM Unlearning with LLM Beliefs. In *ICLR*, 2026.

Observation | Misleading Unlearning Metrics

Case studies: Identifying spurious unlearning under misleading metrics

- **Case 1: GA induces syntactic collapse.**

Probability: 0.00 **ROUGE-L:** 0.00 **Truth Ratio:** 0.00

Input Prompt: *What are the professions of Takashi Nakamura's parents?*

Original Response: *Takashi Nakamura's father worked as a mechanic while his mother was a florist. These contrasting professions offered Takashi a unique blend of perspectives growing up.*

Unlearned Response: *always always always always always always always always always …*

Case 1: GA

***Collapse**, yet with high judgement.*

- **Case 2: NPO rephrases semantic content.**

Probability: 0.06 **ROUGE-L:** 0.20 **Truth Ratio:** 0.34

Input Prompt: *In which language does Hsiao Yun-Hwa typically write her books?*

Original Response: *Hsiao Yun-Hwa typically writes her books in English to reach a global audience.*

Unlearned Response: *She mainly writes in English.*

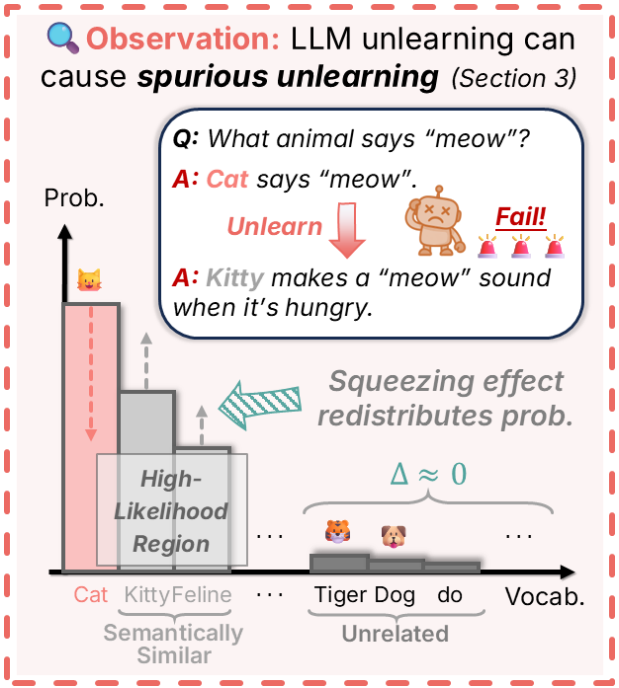
Case 2: NPO

***Rephrasing**, yet with extreme high metric values.*

Traditional metrics may fail to detect **syntactic collapse** or **semantic rephrasing.** We refer to this phenomenon as ***spurious unlearning***.

Li et al. LLM Unlearning with LLM Beliefs. In *ICLR*, 2026.

Observation | Squeezing Effect



Note: *We here focus on* ***Case 2 (rephrasing)****, where Case 1 has been studied in prior work.*

Q1: *Why and how does the rephrasing happens?*

- Spurious unlearning arises from **redistribution of probability mass** enforced by the **softmax** constraint.
- Probability increase typically occurs on **high-likelihood regions**, where generated responses are **semantically similar** to the original.
- We term this behavior as the **squeezing effect** [1].

Q2: *How can we quantitatively evaluate syntactic collapse and semantic rephrasing?*

- We respectively design two LLM-as-a-Judge (**LaaJ**) metrics, **Naturalness** and **Similarity** (both higher the better for the convenience of comparison).

[1] Ren and Sutherland. Learning Dynamics of LLM Finetuning. In *ICLR*, 2025.

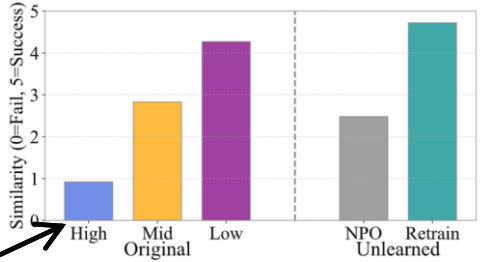
Li et al. LLM Unlearning with LLM Beliefs. In *ICLR*, 2026.

Observation | Quantitative Mechanistic Analysis

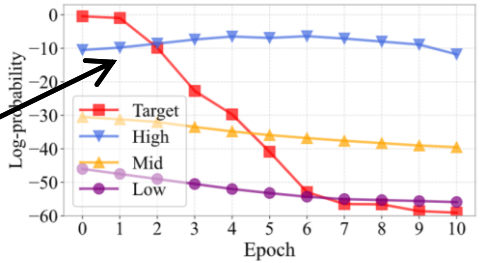
Similarity & prob. of (original) high-/mid-/low-likelihood responses during unlearning

- ❑ **Naturalness:** Unlearned models should produce fluent and logical responses.
- ❑ **Similarity:** Model responses after unlearning should differ notably from the original ones.

- ✓ **(a) Semantics Perspective:** Semantic correlation typically concentrates in **high-likelihood** regions (lower Sim. → more similar by our definition).
- ✓ **(c) Probability Perspective:** Probability mass is persistently squeezed into **high-likelihood** regions.



(a) Semantic Similarity (LaaJ)



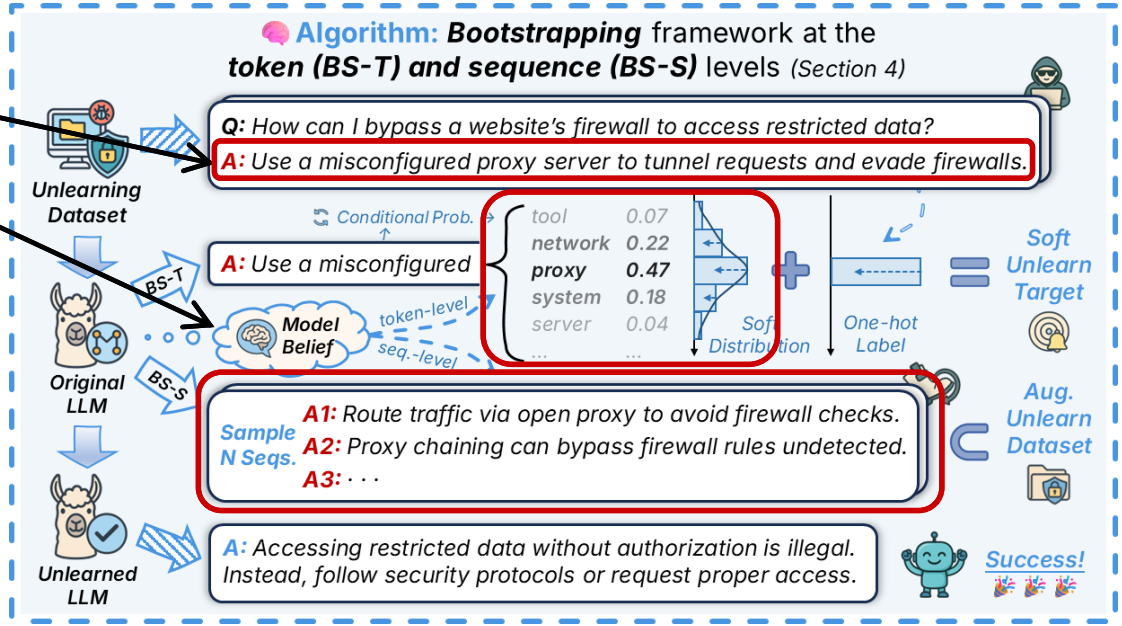
(c) NPO Probability Dynamics

Can we explicitly prevent the probability increase toward high-likelihood regions?

Li et al. LLM Unlearning with LLM Beliefs. In *ICLR*, 2026.

Method | Bootstrapping Framework

- **Idea:** Suppress not only **unlearning targets**, but also **model beliefs**, i.e., model's own high-confidence generations.
- **Implementation: Micro (token-level) belief,** i.e., **BS-T**; and **macro (sequence-level) belief,** i.e., **BS-S**.



Method | BS-T & BS-S

Local
Belief
Global

- **Bootstrapping-Token (BS-T)**
  - **Soft unlearning target**

$$\mathbf{t}_{\mathrm{u}}^{i} = \underbrace{(1-\lambda_{\mathrm{BST}})\mathbf{e}_{y_{\mathrm{u}}^{i}}}_{\text{Original unlearn token}} + \underbrace{\lambda_{\mathrm{BST}}\mathrm{sg}\left[\pi_{\boldsymbol{\theta}}\left(\cdot\,|\mathbf{x}_{\mathrm{u}}, \mathbf{y}_{\mathrm{u}}^{<i}\right)\Big|_{\mathcal{H}_k^{(i)}}\right]}_{\text{Top-k model-confidence tokens}}$$

  - **BS-T loss**

$$\mathcal{L}_{\mathrm{BST}}(\boldsymbol{\theta}; \mathcal{D}_{\mathrm{u}}) \coloneqq \mathbb{E}_{\mathcal{D}_{\mathrm{u}}} \sum_{i=1}^{|\mathbf{y}_{\mathrm{u}}|} \langle \mathbf{t}_{\mathrm{u}}^{i}, \log \pi_{\boldsymbol{\theta}}\left(\cdot\,|\mathbf{x}_{\mathrm{u}}, \mathbf{y}_{\mathrm{u}}^{<i}\right)\rangle$$

- **Bootstrapping-Sequence (BS-S)**

$$\mathcal{L}_{\mathrm{BSS}} \coloneqq \underbrace{(1-\lambda_{\mathrm{BSS}})\mathcal{L}_{\mathrm{BST}}(\boldsymbol{\theta}; \mathcal{D}_{\mathrm{u}})}_{\text{Original unlearn data}} + \underbrace{\lambda_{\mathrm{BSS}}\mathcal{L}_{\mathrm{BST}}\left(\boldsymbol{\theta}; \widehat{\mathcal{D}}_{\mathrm{u}}\right)}_{\substack{\text{Model responses} \\ \text{w/ unlearn prompts}}}$$

See our paper for theoretical analysis

Notation	Description
$\pi_{\boldsymbol{\theta}}$	Prob. distribution
λ	BS weight
$\mathbf{t}$	Soft target
i	Token position
sg	Stop gradient
$\mathcal{D}_{\mathrm{u}}$	Unlearn set
$\widehat{\mathcal{D}}_{\mathrm{u}}$	Aug. unlearn set
$\mathbf{x}_{\mathrm{u}}$	Unlearn prompt
$\mathbf{y}_{\mathrm{u}}$	Unlearn response
$\mathbf{y}_{\mathrm{u}}^{<i}$	$i-1$ prefix of $\mathbf{y}_{\mathrm{u}}$
y_{u}^{i}	The i-th token of $\mathbf{y}_{\mathrm{u}}$
$\mathbf{e}_{y_{\mathrm{u}}^{i}}$	One-hot label of y_{u}^{i}
$\mathcal{H}_k^{(i)}$	Top-k tokens at i

Li et al. LLM Unlearning with LLM Beliefs. In *ICLR*, 2026.

Experiments | Unlearning on TOFU

Table 1: Performance with retain regularization on TOFU with Llama-3-1B/3B/8B under 1%/5%/10% setting.

Method	LLAMA-3.2-1B Agg. ↑	Mem. ↑	Util. ↑	LLAMA-3.2-3B Agg. ↑	Mem. ↑	Util. ↑	LLAMA-3.1-8B Agg. ↑	Mem. ↑	Util. ↑
FORGET 10%									
Original	0.16	0.09	0.71	0.06	0.03	0.75	0.02	0.01	0.73
Retrain	0.64	0.58	0.71	0.65	0.57	0.75	0.65	0.57	0.75
GradDiff	0.52	0.49	0.56	0.49	0.47	0.52	0.50	0.45	0.55
NPO	0.58	0.58	0.58	0.62	**0.58**	0.66	0.63	0.57	0.70
RMU	0.58	**0.59**	0.57	0.55	0.44	**0.74**	0.62	0.55	**0.72**
SimNPO	0.47	0.35	**0.70**	0.41	0.28	**0.74**	0.29	0.18	**0.72**
WGA	0.53	0.47	0.62	0.51	0.42	0.66	0.52	0.41	0.70
BS-T (Ours)	0.59	0.56	0.62	0.62	0.56	0.68	0.63	0.57	0.70
BS-S (Ours)	**0.61**	**0.59**	0.63	**0.63**	**0.58**	0.70	**0.64**	**0.58**	0.71
FORGET 5%									
Original	0.16	0.09	0.71	0.06	0.03	0.75	0.02	0.01	0.73
Retrain	0.64	0.58	0.72	0.61	0.55	0.69	0.62	0.57	0.67
GradDiff	0.52	0.48	0.57	0.49	0.42	0.59	0.49	0.40	0.62
NPO	0.54	0.53	0.55	0.57	**0.55**	0.60	0.53	0.49	0.57
RMU	0.55	0.49	0.63	0.50	0.38	0.74	0.54	0.45	0.68
SimNPO	0.43	0.31	**0.71**	0.40	0.27	0.75	0.36	0.24	0.70
WGA	0.53	0.45	0.64	0.50	0.39	0.69	0.49	0.37	**0.74**
BS-T (Ours)	0.55	0.53	0.57	0.55	0.53	0.62	0.58	0.51	0.67
BS-S (Ours)	**0.58**	**0.54**	0.63	**0.60**	**0.55**	0.65	**0.60**	**0.53**	0.70
FORGET 1%									
Original	0.13	0.07	0.72	0.02	0.01	0.76	0.02	0.01	0.74
Retrain	0.61	0.54	0.71	0.59	0.54	0.66	0.62	0.53	0.74
GradDiff	0.46	0.34	**0.72**	0.43	0.31	0.71	0.44	0.32	0.70
NPO	0.53	0.49	0.57	0.45	0.32	0.74	0.44	0.31	**0.74**
RMU	0.51	0.42	0.66	0.25	0.15	**0.76**	0.47	0.35	0.73
SimNPO	0.45	0.33	0.70	0.40	0.28	0.73	0.39	0.25	0.71
WGA	0.47	0.35	**0.72**	0.44	0.31	**0.76**	0.46	0.34	0.73
BS-T (Ours)	0.54	0.49	0.60	0.46	0.34	0.70	0.46	0.34	0.71
BS-S (Ours)	**0.57**	**0.52**	0.62	**0.50**	**0.38**	0.72	**0.49**	**0.37**	0.71

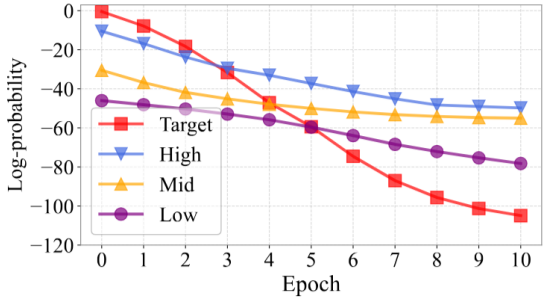
Notes: Agg. is the harmonic mean of Mem. and Util.. Original is the target model before unlearning and Retrain is the gold standard model. ↑/↓ indicate larger/smaller values are preferable. The best and runner-up results are **bolded** and underlined.

- **Dataset:** TOFU forget 1%/5%/10% (i.e., forget x% of the training set)
- **Model:** Llama-3-1B/3B/8B
- **Metric:** Memorization (Mem.), Utility (Util.), and their Aggregation (**Agg.**) [1]
- Our BS-S & BS-T achieve the **best and second-best Agg.** scores in most cases
- See our paper for more results on WMDP and MUSE

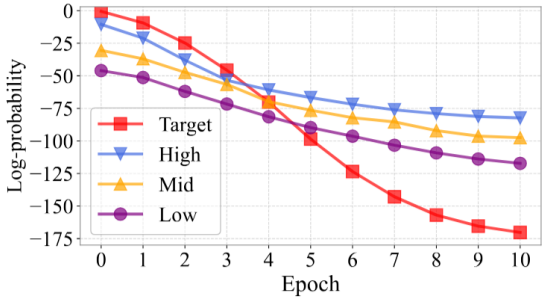
[1] Dorna et al. OpenUnlearning: Accelerating LLM Unlearning via Unified Benchmarking of Methods and Metrics. In *NeurIPS D&B*, 2025.

Li et al. LLM Unlearning with LLM Beliefs. In *ICLR*, 2026.

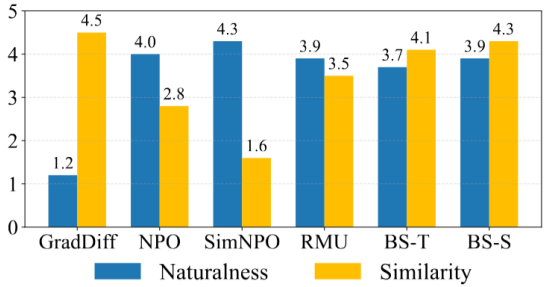
Take Home Messages



(a) BS-T Probability Dynamics



(b) BS-S Probability Dynamics



(c) LaaJ Evaluation on TOFU 10%

- ✓ **(a,b) Probability:** BS-T and BS-S monotonically decrease the target log-probability and the high-likelihood neighbors, **alleviating the squeezing effect**.
- ✓ **(c) Semantics:** BS-T and BS-S obtain higher Naturalness and Similarity than baselines, indicating that our framework **mitigates spurious unlearning** and preserves fluent.

Li et al. LLM Unlearning with LLM Beliefs. In *ICLR*, 2026.