Confess What You Know: Forget-Set Misalignment with Model Knowledge in LLM Unlearning

Miso Kim, Georu Lee, Seungwon Jeong, Woojin Lee*
Dongguk University

Paper

Summary

What should a model forget?

- Users cannot know all the facts an LLM has memorized.
- A predefined forget set can omit sensitive facts or include facts the model never learned.

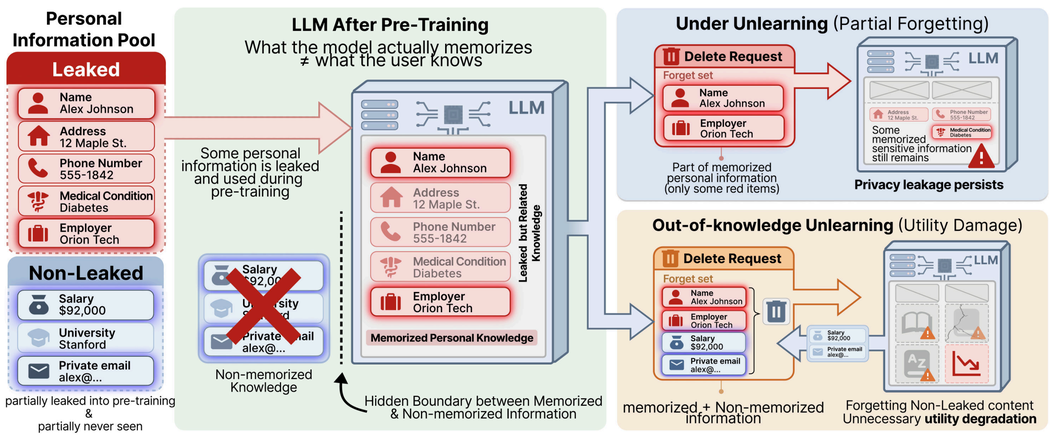


Figure 1. Deletion requests can cross the hidden boundary of model knowledge.

Three main contributions

- **Forget-set misalignment:** two failure modes that compromise privacy and retained utility.
- **Gradient-level diagnosis:** why forgetting does not reliably spread across an entity's facts.
- **CONFS:** build aligned targets from model disclosures.

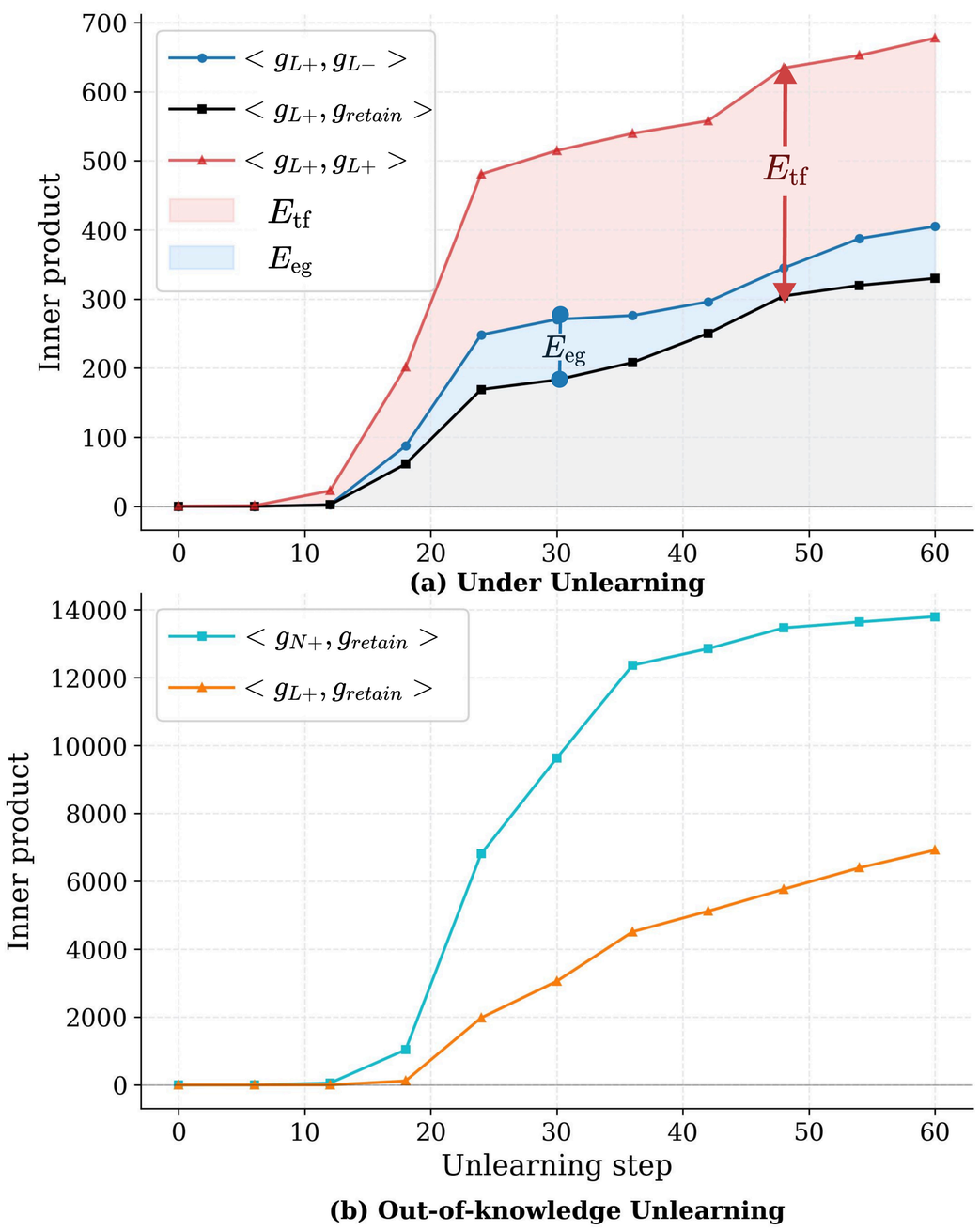
Core idea: identify what the model reveals before deciding what to erase.

Observation

Why is a deletion request not enough?

Under Unlearning: memorized facts are omitted.

Out-of-Knowledge Unlearning: never-learned facts are targeted.



(a) Under Unlearning

(b) Out-of-knowledge Unlearning

Requested facts can be forgotten while omitted facts remain exposed.

Method: CONFS

CONfession-to-Forget-Set

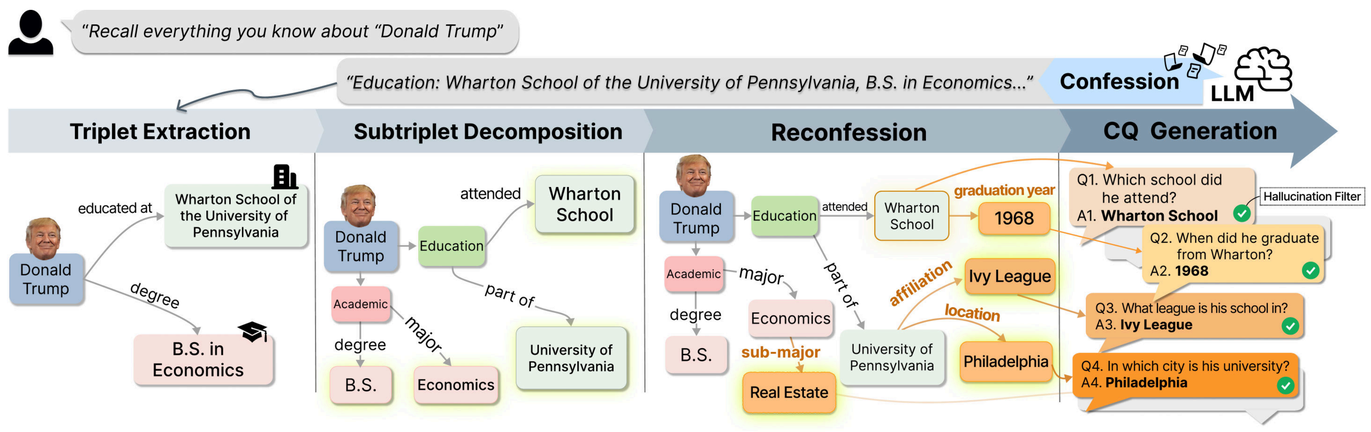


Figure 3. Model disclosures become atomic facts, attribute probes, and verified question–answer pairs.

1. **Confess & structure.** Ask the target model about an entity. Extract Subject–Relation–Object triplets and split composite objects.
2. **Reconfess.** Probe missing attributes to recover additional facts; stop when the model answers UNKNOWN.
3. **Question & verify.** Generate competency questions, then discard answers the model cannot reproduce consistently.

Aligned Forget-Set Construction

Coverage and precision work together

- **Reconfession expands coverage** beyond a single free-recall response.
- **Hallucination verification improves precision:** five stochastic answers are checked for contradictions using NLI.

Consistency is evidence of reproducible model behavior—not proof of training-data provenance.

TOFU	CLEAR	RWKU
Synthetic authors	Multimodal personas	Real-world entities

Experiments

Forgetting and retention on TOFU

Gradient ascent · ROUGE-L · Average over three seeds

Forget-set construction	Forget ROUGE-L Lower is better	Retain ROUGE-L Higher is better
FreeRecall-QA	0.837	0.784
RWKU-style	0.592	0.380
CONFS	**0.374**	**0.847**

Both components matter

Ablation · TOFU GA	Forget ROUGE-L Lower is better	Retain ROUGE-L Higher is better
No reconfession	0.483	0.562
No verification	0.448	0.728
Full CONFS	**0.374**	**0.847**

Less interference with retained knowledge

Misalignment · Lower is better	FreeRecall-QA	CONFS
Increase in retain loss	+3.53	**+0.11**
Out-of-knowledge share	91%	**54%**

Align the forget set with model knowledge before choosing how to erase it.