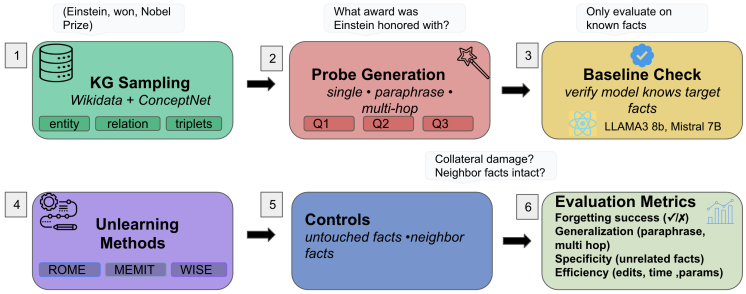
How Well Do LLMs Unlearn Facts? - A Knowledge Graph Perspective

Machine unlearning is the process of selectively removing specific data from trained models without retraining from scratch, which has been a promising technique recently due to the regulatory requirements surrounding data usage. In Large Language Models (LLMs), unlearning is a pressing and challenging task because of their unprecedented capability to memorize and digest training data at scale, raising more significant issues on safety, privacy, and intellectual property. There are some existing methods in the field of LLMs unlearning, including parameter editing methods, fine-tuning methods, and distillation-based methods, yet they are all focused on the flat sentence-level data and overlook the relational, multi-hop, and reasoned knowledge in the naturally structured data. Such limitations significantly restrict the effectiveness of unlearning in LLMs, where knowledge is inherently interconnected, which challenges *how well LLMs can unlearn the structured knowledge in specific scenarios.*

Knowledge Graphs (KGs), composed of triples (e_h, r, e_t), where e_h and e_t are head and tail entities connected by relation r and serve as a backbone for applications in semantic search and commonsense reasoning. Recently, KGs, such as Wikidata, have been used in the pre- and post-training stages of LLMs, as the structured knowledge can facilitate complex tasks, e.g, question answering, by their semantic relations and reasoning across multiple hops.

However, regarding unlearning, it remains unclear how well LLMs can forget such structured knowledge by existing unlearning methods. Existing unlearning benchmarks (e.g., KLUE [3], HANKER [4]) focus on flat, sentence-level knowledge, but not KG-based facts. For example, consider a triplet, ("Einstein", "won", "Nobel Prize"). If an LLM is instructed to unlearn this fact, we must not only check whether it fails to answer "Which prize did Einstein win?" but also paraphrases ("What award was Einstein honored with?") and multi-hop questions ("Which physicist who developed relativity won a Nobel Prize?"). In addition, we must ensure that unrelated knowledge, such as "Einstein was a physicist" remains intact. Such KG-based unlearning evaluations capture the effectiveness of unlearning in a structured, relational context that sentence-level text cannot. Nonetheless, to date, no benchmark systematically explores whether LLMs truly forget KG-derived knowledge, which leaves a gap in evaluating the precision and reliability of unlearning methods.



In this work, we propose a novel benchmark and associated framework designed to assess how well LLMs forget structured KG facts on state-of-the-art unlearning methods (e.g., ROME [1], MEMIT [2], etc.) and popular models, like LLaMA-3 8B and Mistral 7B. Three main contributions unfold as follows:

- A novel benchmark dataset that covers diverse relations in triples from Wikidata (broad factual knowledge) and ConceptNet (commonsense knowledge) that explicitly targets KG-derived facts.
- A survey of existing unlearning methods with early empirical insights showing the challenges of applying current methods to graph-structured knowledge.
- A systematic framework and metrics for unlearning in LLMs that extend beyond sentence-level through (1) single-fact queries, (2) paraphrase robustness, (3) multi-hop reasoning, and (4) KG consistency checks.

We aim to establish standardized evaluation criteria for KG unlearning and provide a foundation for future methods that handle the complexity of structured, relational knowledge. This work advances the study of safe and reliable machine unlearning by introducing a benchmark that systematically evaluates current LLM unlearning methods in a structured, knowledge graph setting. Our future work will investigate the KG unlearning in LLMs.

References

[1] Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). *ROME: Locating and Editing Factual Associations in GPT*. NeurIPS.

[2] Meng, K., Sharma, A. S., Andonian, A., Belinkov, Y., & Bau, D. (2023). *MEMIT: Mass-Editing Memory in a Transformer*. ICLR.

[3] Yang, N., Kim, M., Yoon, S., Shin, J., & Jung, K. (2025). *FaithUn: Toward Faithful Forgetting in Language Models by Investigating the Interconnectedness of Knowledge*. arXiv:2502.19207.

[4] Jiang, W., Zhai, J., Ma, S., Lei, Z., Xie, X., Wang, Y., & Shen, C. (2025). *HANKER: Holistic Audit Dataset Generation for LLM Unlearning via Knowledge Graph Traversal and Redundancy Removal*. arXiv:2502.18810.