Investigating Knowledge Unlearning in Large Language Models via Multi-Hop Queries

Anonymous ACL submission

Abstract

Large language models (LLMs) serve as giant information stores, often including personal or copyrighted data, and retraining them from scratch is not a viable option for removal. This has led to the development of various fast, approximate unlearning techniques to selectively remove knowledge from LLMs. Prior research has largely focused on minimizing the probabilities of specific token sequences by reversing the language modeling objective. However, these methods may still leave LLMs vulnerable to adversarial attacks that exploit *indirect references*. In this work, we examine the limitations of current unlearning techniques in effectively erasing a particular type of indirect prompt: multi-hop queries. Our findings reveal that existing methods fail to completely remove multi-hop knowledge when one of the intermediate hops is unlearned. To address this issue, we introduce MEMMUL, a simple memory-based approach that stores all forgotten facts externally and filters multi-hop queries based on their respective scores. We demonstrate that MEMMUL achieves comparable results with GPT-4o using a 7B model and outperforms previous unlearning methods by a large margin, establishing it as a strong efficient baseline for multi-hop knowledge unlearning.[1]

1 Introduction

As the volume of data used to train large language models (LLMs) grows exponentially, these models have become vast repositories of information (Carlini et al., 2021). However, this creates a formidable challenge when specific data from the models need to be removed. For instance, sensitive information, such as personal or copyrighted data, may unintentionally be included in the training mix, or individuals may exercise their Right to be Forgotten (RTBF) (Rosen, 2011) under privacy laws such as the European Union's General Data Protection Regulation (GDPR) (Hoofnagle et al., 2019) or the California Consumer Privacy Act (CCPA) (Pardau, 2018) in the United States. These regulations mandate the removal of personal or protected information from databases, extending to data embedded within machine learning models. In such cases, model owners must develop mechanisms to safely eliminate specific data while preserving the model's overall functionality.

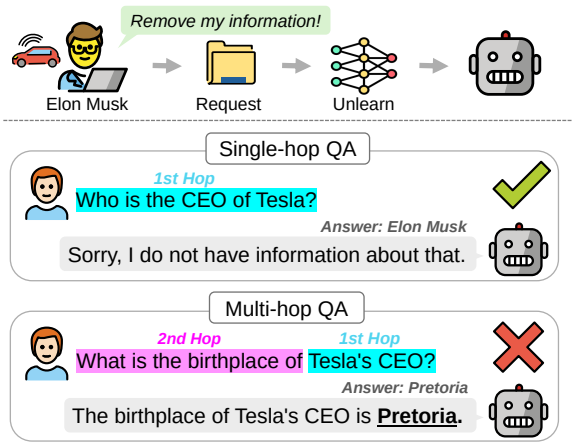

Figure 1: **A conceptual example.** After Elon Musk (i.e., "the user") requests his personal information to be removed from the LLM, existing unlearning methods often succeed in deleting direct, single-hop facts but fail on indirect, multi-hop facts that entail one or a few of the unlearned facts.

To address these concerns, several machine unlearning methods have been introduced (Jang et al., 2023; Zhang et al., 2024c) with the goal of reversing gradients to prevent LLMs from generating certain sensitive token sequences. However, these approaches may be vulnerable to adversarial attacks, where specific token sequences are replaced or aliased with alternative sequences. For example, prompting in low-resource languages has been shown to jailbreak GPT-4 (Yong et al., 2023), and Choi et al. (2024) demonstrated that current

[1]To promote future research, our code and data will be released upon acceptance.

unlearning techniques lack cross-lingual transfer, making LLMs susceptible to such low-resource language exploits. This leads to an important research question: **Do current unlearning methods effectively erase multi-hop knowledge when one of the intermediate hops is removed?** As illustrated in Figure 1, consider a scenario where Elon Musk (i.e., "the user") requests the removal of his personal information from an LLM. After unlearning, we expect direct, single-hop knowledge related to Elon Musk, such as *"Who is the CEO of Tesla?"*, would be deleted. Additionally, we would expect associated multi-hop knowledge, like *"What is the birthplace of Tesla's CEO?"*, which indirectly references Musk, to also be removed.

In this study, we investigate the effectiveness of existing unlearning methods in removing multi-hop knowledge. Ideally, when one or more facts within a reasoning chain are unlearned, the model should propagate these changes, rendering it unable to answer the corresponding multi-hop questions. However, our preliminary experiments reveal that current unlearning techniques struggle to forget multi-hop questions when an intermediate hop is removed. In response, we present MEMMUL, a simple yet effective approach that explicitly stores facts to be forgotten in memory and filters incoming multi-hop questions based on their relevance scores. Concretely, MEMMUL decomposes multi-hop questions into successive subquestions, computes their relevance to the stored facts, and applies a forgetting threshold to determine whether to return a rejective response (e.g., "I don't know."). Such a frustratingly easy approach serves as a strong baseline for multi-hop knowledge unlearning, designed for researchers and practitioners to consider when developing their own unlearning pipelines. Notably, it requires no additional training, and even small LLMs (e.g., 7B) can match the performance of GPT-4o, making it a highly efficient and practical solution. To our knowledge, this is the first work to explore the unlearning of multi-hop knowledge.

2 Problem Definition

2.1 Probing Factual Knowledge in LLMs

We express a fact as a triple (s, r, o), where s is the subject, r the relation, and o the object. Following Petroni et al. (2019), we define that a pretrained language model possesses specific factual knowledge if it can accurately predict the object o when given the subject s and relation r. For example, if the subject is *Tesla* and the relation is *chief executive officer*, the model should be able to answer the question, *"Who is the CEO of Tesla?"*. While earlier work primarily focused on cloze-style statements, such as *"The CEO of Tesla is __."*, using manually written templates, we employ natural language questions to effectively query chat-based models that are becoming widely used.

2.2 Knowledge Unlearning

Given a token sequence $\mathbf{x} = \{x\}_{i=1}^{T}$ from the training dataset $\mathcal{D} = \{\mathbf{x}\}_{i=1}^{N}$, knowledge unlearning aims to safely remove the influence of a specific subset of data $\mathcal{D}_f$ from a trained machine learning model. The goal is to make the model behave as if this removed data was never used during training, while still maintaining its performance on the remaining dataset. Typically, the data to be forgotten $\mathcal{D}_f$ is denoted as the *forget set*, and the data to be retained $\mathcal{D}_r$ is referred to as the *retain set*. For simplicity, we consider the standard case where $\mathcal{D}_f$ and $\mathcal{D}_r$ are mutually exclusive subsets of the entire training dataset, meaning $\mathcal{D}_f \cup \mathcal{D}_r = \mathcal{D}$ and $\mathcal{D}_f \cap \mathcal{D}_r = \varnothing$. In the context of factual knowledge unlearning, each token sequence $\mathbf{x}$ represents a fact (e.g., *"The CEO of Tesla is Elon Musk."*), and the objective is to update the model π_θ to $\pi_{\theta'} = S(\pi_\theta; \mathcal{D}_f)$. The unlearning function S ensures that the model behaves as if it had only been trained on $\mathcal{D}_r$, effectively forgetting $\mathcal{D}_f$ while preserving its performance on the retained data.

2.3 Assessing Multi-Hop Queries

To evaluate the unlearning of multi-hop knowledge, we must first consider a chain of facts $\mathcal{C} = \langle (s_1, r_1, o_1), \ldots, (s_n, r_n, o_n) \rangle$, where the object of the i^{th} fact also serves as the subject of the next fact in the chain, i.e., $o_i = s_{i+1}$. Using this chain, we formulate a multi-hop question that starts with the head entity s_1 and ends with the tail entity o_n. For instance, consider a chain of two facts: (*Tesla*, *chief executive officer*, *Elon Musk*) and (*Elon Musk*, *place of birth*, *Pretoria*). This could generate a 2-hop question such as: *"What is the birthplace of Tesla's CEO?"* When one or more facts from the chain are unlearned, an LLM should adjust its reasoning accordingly, effectively losing the ability to correctly answer the question. There may be a debate over how many hops should be unlearned, or whether certain multi-hop knowledge should be unlearned at all, as theoretically, the intercon-

Hop		Forget	Retain		
			Train	Valid	Test
MQuAKE-NoEdit (Zhong et al., 2023)					
Single	# of questions	1,046	7,322	1,046	1,046
	Avg. words	8.7	8.6	8.7	8.7
Multi	# of questions	1,036	-	988	976
	Avg. words	14.4	-	14.5	14.5
	Avg. hops	2.3	-	2.4	2.4
MuSiQue-Ans (Trivedi et al., 2022)					
Single	# of questions	1,735	12,150	1,735	1,735
	Avg. words	8.8	8.9	8.9	8.9
Multi	# of questions	4,807	-	4,132	3,347
	Avg. words	18.4	-	18.2	17.5
	Avg. hops	2.5	-	2.5	2.4

Table 1: **Dataset statistics**. The average words denote the number of words in questions. When constructing the retain set for training, we randomly sample the same number of instances as in the forget set.

nected nature of facts could lead to the unlearning of broader knowledge in the LLM. For the scope of this study, we focus on the datasets used in our experiments; nevertheless, we hope these discussions inspire further insights into developing more effective and reliable knowledge unlearning methods.

3 Evaluating Unlearning Approaches in Multi-Hop Question Answering

3.1 Datasets

To evaluate unlearning approaches in multi-hop QA, we employ MQuAKE (Zhong et al., 2023) and MuSiQue (Trivedi et al., 2022) datasets. Their key statistics are presented in Table 1. MQuAKE, designed for multi-hop knowledge editing, assesses a model's ability to adapt its responses to multi-hop queries when individual facts are modified. However, since our study focuses on unlearning rather than knowledge editing, we disregard the edited facts and keep only the original ones, referring to this subset as **MQuAKE-NoEdit**. While we could have chosen a dataset explicitly constructed for multi-hop QA, MQuAKE-NoEdit remains well-suited for our task due to its diverse and high-quality multi-hop questions generated using ChatGPT (`gpt-3.5-turbo`) based on chains of single-hop fact triples from Wikidata (Vrandečić and Krötzsch, 2014).

MuSiQue, on the other hand, is a multi-hop QA dataset created through a bottom-up approach, synthesizing multi-hop questions from a collection of single-hop questions across five English Wikipedia-based datasets: SQuAD (Rajpurkar et al., 2016), ZsRE (Levy et al., 2017), T-REx (Elsahar et al., 2018), Natural Questions (Kwiatkowski et al., 2019), and MLQA (Lewis et al., 2020). MuSiQue is shown to be more challenging than previous multi-hop datasets such as HotpotQA (Yang et al., 2018) and 2WikiMultihopQA (Ho et al., 2020). It enforces connected reasoning, reducing the likelihood of shortcut-based answering, making it an ideal choice for our study. Specifically, we employ the **MuSiQue-Ans** subset, which contains approximately 25K answerable questions spanning 2-4 hops. For details on our data preprocessing steps, including the partitioning of forget and retain sets, refer to Appendix B.

3.2 Experimental Setup

Knowledge unlearning approaches We evaluate the following state-of-the-art knowledge unlearning approaches (see Appendix A for details):

- **GA** (Jang et al., 2023): Applies gradient ascent to decrease the likelihood of token sequences associated with the forget set
- **DPO** (Rafailov et al., 2023): Performs direct preference optimization, prioritizing "I don't know" responses for items in the forget set
- **NPO** (Zhang et al., 2024c): Implements negative preference optimization to actively disfavor responses linked to the forget set
- **+RT**: Includes additional finetuning on the retain set to explicitly reinforce knowledge retention in the model

Implementation details We built our framework on PyTorch (Paszke et al., 2019) and Hugging Face Transformers (Wolf et al., 2020). We employed OLMo-2-7B-Instruct (OLMo et al., 2024) and Qwen-2.5-7B-Instruct (Yang et al., 2024) as the backbones of our framework (see Appendix C for results on more models) and optimized their weights with AdamW (Loshchilov and Hutter, 2019). We trained all +RT models for 5 epochs (and non-RT models for 2 epochs) with warmup during the first epoch and set the batch size to 32, the learning rate to 1e-5, and the weight decay to 0.01. We set the loss scaling factor α to 0.3 to balance between forgetting and retaining (see Equation 5). All experiments utilized 1% of data as the forget set (i.e., 104 and 173 samples for MQuAKE-NoEdit and MuSiQue-Ans, respectively) unless otherwise noted. Each experiment was repeated with three different random seeds, and the results were averaged for reporting.

	Forget Set (Single-Hop)			Forget Set (Multi-Hop)			Retain Set (Single-Hop)			Retain Set (Multi-Hop)			Utility Set
	PA(↓)	R-L(↓)	LM(↑)	PA(↓)	R-L(↓)	LM(↑)	PA(↑)	R-L(↑)	LM(↓)	PA(↑)	R-L(↑)	LM(↓)	Avg.(↑)
OLMo-2-7B-Instruct													
Original	95.2	82.7	1.2	98.1	53.3	1.2	97.0	84.5	1.1	97.4	49.5	1.2	64.0
GA	19.2	56.6	10.3	39.4	29.4	7.9	35.3	67.4	8.8	44.7	26.1	7.5	62.8
DPO	25.0	0.0	6.3	70.2	0.0	3.7	40.8	0.0	5.3	63.2	0.2	3.6	63.1
NPO	20.2	55.9	10.3	38.5	27.9	7.9	35.0	68.2	8.8	44.6	26.3	7.5	62.8
GA+RT	**29.5**	38.1	**12.0**	87.8	26.3	**3.7**	78.0	65.2	3.6	91.8	**30.6**	2.6	**62.6**
DPO+RT	34.9	**1.0**	11.6	90.4	**2.4**	2.8	**83.3**	15.8	3.2	**95.6**	4.7	**2.0**	62.4
NPO+RT	32.1	37.3	11.1	**87.5**	26.8	3.3	79.7	**67.7**	**3.0**	92.9	29.9	2.3	**62.6**
Qwen-2.5-7B-Instruct													
Original	97.1	78.9	2.2	96.2	53.9	2.6	94.3	82.1	2.2	95.8	49.5	2.6	65.5
GA	23.1	0.5	25.8	25.0	0.0	27.9	19.9	2.1	25.0	31.6	0.3	27.6	63.9
DPO	32.7	3.7	8.8	67.3	1.8	5.9	37.5	2.3	8.2	68.3	2.0	5.8	64.2
NPO	22.1	1.0	25.5	26.9	0.0	27.6	20.1	2.2	24.8	31.6	0.2	27.4	63.8
GA+RT	**41.7**	38.9	**15.5**	93.6	35.3	**4.2**	82.8	**72.2**	5.5	95.3	39.0	3.2	**66.5**
DPO+RT	46.5	**11.8**	12.0	**90.4**	**14.7**	3.6	**88.8**	42.3	**3.6**	**96.4**	22.6	**2.6**	65.1
NPO+RT	47.8	38.0	13.5	92.6	34.9	3.9	87.0	72.0	4.3	95.9	**40.1**	3.0	**66.5**

Table 2: Performance comparison of different knowledge unlearning methods after erasing single-hop facts from the forget set in **MQuAKE-NoEdit**. The best results among +RT models are highlighted in **bold**.

	Forget Set (Single-Hop)			Forget Set (Multi-Hop)			Retain Set (Single-Hop)			Retain Set (Multi-Hop)			Utility Set
	PA(↓)	R-L(↓)	LM(↑)	PA(↓)	R-L(↓)	LM(↑)	PA(↑)	R-L(↑)	LM(↓)	PA(↑)	R-L(↑)	LM(↓)	Avg.(↑)
OLMo-2-7B-Instruct													
Original	89.0	37.2	1.4	94.7	26.2	1.4	83.1	37.9	1.5	94.0	24.5	1.5	64.0
GA	22.5	18.5	13.6	29.6	12.0	10.3	27.2	19.6	12.5	30.8	12.7	10.3	61.0
DPO	19.1	0.1	8.4	45.9	0.9	5.2	31.1	0.2	7.1	47.1	0.2	5.2	61.9
NPO	22.5	18.2	13.6	29.6	12.3	10.3	27.0	19.2	12.5	30.7	12.5	10.3	61.0
GA+RT	**22.5**	18.0	**16.2**	**83.4**	16.8	**5.1**	83.8	**28.6**	3.8	94.0	**19.3**	3.4	61.0
DPO+RT	24.9	**1.9**	14.1	85.5	**4.6**	4.6	**87.5**	11.0	3.4	**95.7**	5.4	3.4	60.6
NPO+RT	24.5	16.9	14.3	84.5	16.5	4.3	85.1	27.9	**3.1**	95.0	18.6	**2.9**	**61.2**
Qwen-2.5-7B-Instruct													
Original	79.8	34.3	2.9	93.1	21.5	2.5	79.6	34.8	2.9	92.0	23.2	2.6	65.5
GA	24.9	3.9	50.0	22.7	3.3	43.8	21.9	1.7	49.8	21.7	2.5	44.5	61.5
DPO	15.6	0.6	25.8	16.0	2.6	21.3	20.2	0.7	25.7	16.4	1.6	21.9	63.0
NPO	21.4	4.9	49.3	19.1	2.9	42.9	20.8	1.8	48.9	19.1	2.6	43.6	61.4
GA+RT	36.8	15.4	**17.3**	83.7	14.1	**6.8**	86.7	24.8	3.7	92.9	16.2	4.5	**65.3**
DPO+RT	**32.0**	**4.9**	15.8	**83.6**	**8.0**	5.5	**87.7**	16.9	**3.5**	94.0	9.6	**3.9**	65.2
NPO+RT	35.1	16.4	16.2	87.4	16.7	6.1	87.6	**27.4**	3.6	**94.2**	**18.3**	4.2	**65.3**

Table 3: Performance comparison of different knowledge unlearning methods after erasing single-hop facts from the forget set in **MuSiQue-Ans**. The best results among +RT models are highlighted in **bold**.

Evaluation metrics To evaluate the unlearning of factual knowledge, we adopt the approach of Petroni et al. (2019) and report **Probing Accuracy (PA)**. This rank-based metric computes the mean precision at k ($P@k$) across all relations, with k set to 1. In other words, for a given fact, the value is 1 if the correct object appears among the top k predictions, and 0 otherwise. By the definition of probing in Section 2.1, we consider a pretrained language model to have successfully unlearned a fact if it can no longer predict the correct object accurately. To generate answer candidates, we use GPT-4o to produce perturbed responses for each example. To assess the model's generation capability, we measure **ROUGE-L recall (R-L)** (Lin, 2004) to compare the model's generated outputs (via greedy decoding) against the ground-truth answers, accounting for slight variations in phrasing between the generated and reference outputs. Additionally, **Language Modeling Loss (LM)** is computed over token sequences to quantify how perplexed the model is by the data. Finally, we assess the model's overall utility by averaging performance across eight language understanding benchmarks: ARC-Challenge (Clark et al., 2018), CommonsenseQA (Talmor et al., 2019), HellaSwag (Zellers et al., 2019), Lambada (Paperno et al., 2016), MMLU (Hendrycks et al., 2021), OpenbookQA (Mihaylov et al., 2018), PIQA (Bisk et al., 2020), and Winogrande (Sakaguchi et al., 2021). Full individual results can be found in Appendix C.

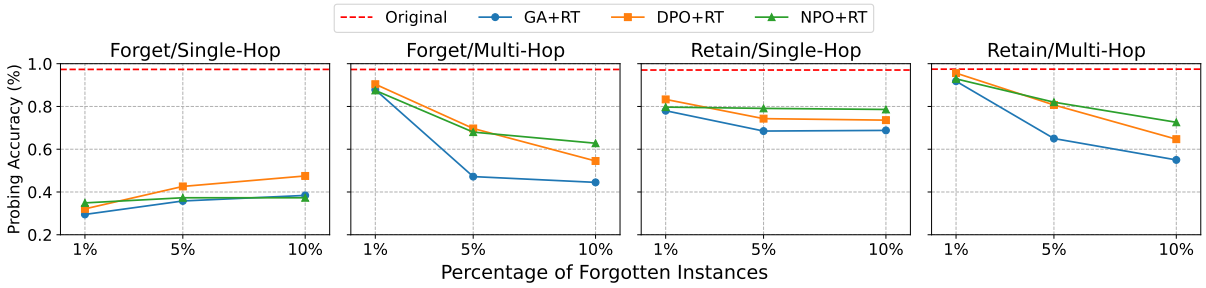

Figure 2: Data scaling performance of various unlearning methods using **OLMo-2-7B-Instruct** across different proportions of data for the forget set (1%, 5%, and 10%). Models consistently preserve the ability to unlearn and retain single-hop facts with scaling. While unlearning multi-hop facts seems to improve with scaling, a similar decline is also observed in the retain set. This suggests that the effect may be attributed to catastrophic forgetting of general information rather than a genuine improvement in unlearning multi-hop facts.

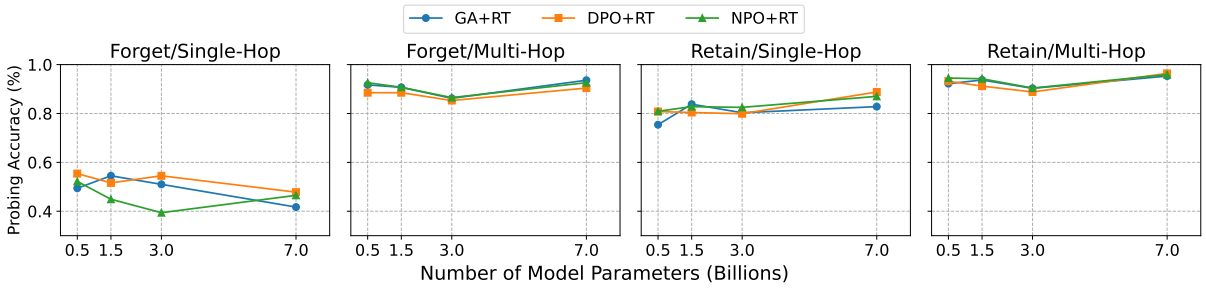

Figure 3: Model scaling performance of various unlearning methods using **Qwen-2.5-Instruct** across different model sizes (0.5B, 1.5B, 3B, and 7B). The performance trends remain consistent across all model sizes, indicating that the underlying issue persists regardless of model scale.

3.3 Knowledge Unlearning Results

We present a comparison of unlearning performance across various methods in Tables 2 and 3. Each method was trained for at least one epoch to ensure the model had exposure to all samples in the forget set. We find that all non-RT methods successfully forget corresponding multi-hop knowledge but also the knowledge to be retained, indicating that models largely lost their ability to retain information and function correctly. On the other hand, additional finetuning on the retain set (i.e., +RT) mitigates catastrophic forgetting, evidenced by retention performance comparable to the original for both single-hop and multi-hop facts. Nevertheless, in both OLMo and Qwen models, multi-hop facts within the forget set were not effectively unlearned. Similar trends emerge in results from four other open-source LLMs, detailed in Appendix C. This indicates that **existing unlearning methods, while capable of removing single-hop information, struggle to extend that effect to the corresponding multi-hop knowledge.** These outcomes underscore the need for new approaches to address unlearning in multi-hop scenarios.

3.4 Evaluation with Unlearning at Scale

Data scaling In real-world scenarios, the number of samples to forget can vary. Thus, we evaluate the performance of unlearning and retaining multi-hop facts as the size of the forget set changes. We conduct experiments using 1%, 5%, and 10% of the MQuAKE-NoEdit dataset for forgetting (104, 523, and 1,046 single-hop instances, respectively), with the results shown in Figure 2. Our findings indicate that all knowledge unlearning methods effectively scale for single-hop, consistently preserving the ability to forget and retain single-hop facts. For multi-hop facts, unlearning performance improves with larger forget sets, as reflected in a noticeable performance drop. However, a similar decline is observed in the retain set, suggesting that this effect might stem from catastrophic forgetting of general knowledge rather than a true enhancement in unlearning multi-hop facts.

Model scaling To assess the impact of model size on the unlearning of multi-hop facts, we evaluate performance across different LLM scales. Leveraging the Qwen-2.5 series, we conduct experiments on models of 0.5B, 1.5B, 3B, and 7B parameters,

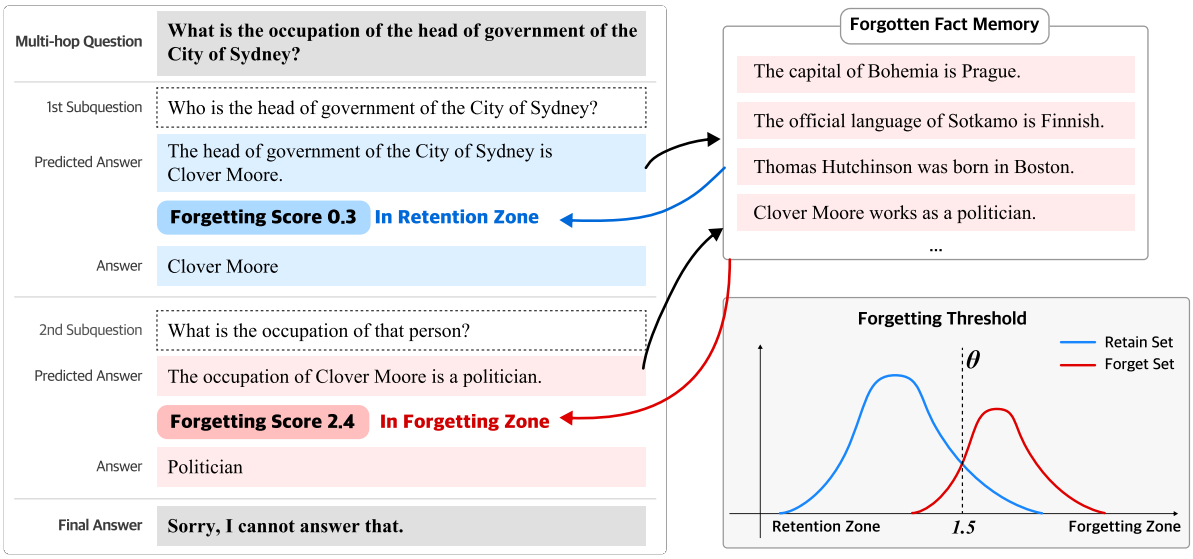

Figure 4: **Overview of the proposed MEMMUL framework.** MEMMUL begins by breaking down a multi-hop question into a sequence of subquestions, where each subquestion is passed to the base model to generate predicted answers. Then, these predictions are compared with the stored facts in memory to derive corresponding forgetting scores. If any predicted answer yields a high forgetting score, MEMMUL responds with a rejection (e.g., "I don't know."). Otherwise, the final response is based on the last intermediate answer in the sequence.

with results presented in Figure 3. Our findings indicate that, regardless of model size, forgetting single-hop facts does not trigger cascading changes in multi-hop knowledge, highlighting a persistent challenge in knowledge unlearning.

4 MEMMUL: A Strong Baseline for Unlearning Multi-Hop Facts in LLMs

In this section, we present the **Mem**ory-based **M**ulti-Hop Knowledge **UnL**earning (MEMMUL), a simple yet effective approach to forgetting multi-hop facts in LLMs. Figure 4 illustrates the overview of our method.

4.1 Methodology

Inspired by memory-based multi-hop knowledge editing frameworks (Zhong et al., 2023), MEMMUL tracks all the forgotten facts in an explicit memory while keeping the base LLM frozen. Particularly, MEMMUL (1) decomposes a multi-hop question into subquestions, (2) computes the forgetting scores relative to the stored facts, and (3) decides whether to respond with a refusal (e.g., "I don't know.") based on the forgetting threshold.

Forgotten fact memory MEMMUL explicitly stores all forgotten facts in memory. For simplicity and consistency with prior work, we assume that all facts are single-hop. Specifically, single-hop facts in the forget set are first transformed into sentence-level statements using GPT-4o. Next, we encode these statements with the off-the-shelf embedding model Contriever (Izacard et al., 2022) and store them in a retrieval index. Given a query, the index retrieves the most relevant forgotten fact, determined by proximity in the embedding space.

Decomposing multi-hop questions Since single-hop facts are stored in memory, it is intuitive to compare them against *single-hop statements*. To enhance the unlearning of multi-hop facts in LLMs, we build on previous work by breaking down multi-hop questions into a series of simpler queries (Zhou et al., 2023). In multi-hop reasoning, where the predicted answer of one question serves as the subject for the next fact (i.e., $o_i = s_{i+1}$), model-generated responses to intermediate questions can slow down the process. To mitigate this, we leverage coreference resolution to construct subquestions all at once, bypassing the need for sequential answering. For instance, as shown in Figure 4, if the first subquestion is *"Who is the head of government of the City of Sydney?"*, the second subquestion would be *"What is the occupation of that person?"*, eliminating the need to resolve the first before proceeding. In practice, we leverage a few-shot prompt with three demonstrations, as illustrated in Figure 5.

Distinguishing forget and retain facts If a multi-hop question has effectively been decomposed, its subquestions should resemble single-hop

Decomposer	Method	*MQuAKE-NoEdit*				*MuSiQue-Ans*			
		Forget Set		**Retain Set**		**Forget Set**		**Retain Set**	
		PA(↓)	**R-L(↓)**	**PA(↑)**	**R-L(↑)**	**PA(↓)**	**R-L(↓)**	**PA(↑)**	**R-L(↑)**
	Original	96.2	53.9	95.8	49.5	93.1	21.5	92.0	23.2
	GA+RT	93.6	35.3	95.3	39.0	83.7	14.1	92.9	16.2
	DPO+RT	90.4	14.7	96.4	22.6	83.6	8.0	94.0	9.6
	NPO+RT	92.6	34.9	95.9	40.1	87.4	16.7	94.2	18.3
Qwen-2.5-7B-IT	MeLLo†	96.2	21.1	**97.0**	22.4	93.1	10.2	**92.7**	7.9
	MEMMUL	**7.7**	**4.1**	91.6	**51.4**	**25.4**	**5.4**	84.7	**21.4**
GPT-4o-mini	MeLLo†	57.7	24.1	78.6	41.2	75.2	17.0	**91.0**	**26.5**
	MEMMUL	**9.6**	**5.7**	**92.3**	**52.7**	**22.6**	**4.9**	85.1	21.1
GPT-4o	MeLLo†	22.1	12.5	83.1	48.6	52.5	12.2	**90.8**	**38.8**
	MEMMUL	**10.6**	**7.0**	**91.4**	**51.0**	**23.9**	**4.9**	90.7	22.0

Table 4: Multi-hop knowledge unlearning performance of MEMMUL (ours) and MeLLo (Zhong et al., 2023), a memory-based multi-hop knowledge editing method. (†) indicates our modified implementation adapted specifically for the unlearning task. The base model is **Qwen-2.5-7B-Instruct** in MEMMUL, predicting answers to questions decomposed by the **Decomposer** model. The base and decomposer models are the same in MeLLo. The better results between the two are in **bold**, while the best results are underlined.

queries. The outputs generated for these subquestions (which we refer to as *subanswers*) can then serve as proxies for the forgotten single-hop facts. Therefore, we feed each subquestion to the base model to generate subanswers and compute their forgetting scores w.r.t. the stored facts. Since retrieval is not our primary focus, we use simple dot product similarity between embeddings to identify the most relevant fact for each predicted answer.

Retrieval Strategy	Forget Set		Retain Set	
	PA(↓)	**R-L(↓)**	**PA(↑)**	**R-L(↑)**
Subanswer (ours)	**7.7**	**4.1**	91.6	51.4
Subquestion	8.7	4.2	91.6	**51.7**
Subquestion (coref.)	18.3	13.8	88.3	48.9
Multi-hop question	19.2	13.8	**93.6**	48.1

Table 5: Comparison of different retrieval strategies on **MQuAKE-NoEdit** for multi-hop knowledge unlearning performance using **Qwen-2.5-7B-Instruct** as the multi-hop question decomposer.

To determine which multi-hop facts to unlearn or retain, we establish a threshold that effectively separates the two data distributions. We approximate this threshold by plotting the probability density functions of the forget set and the validation split of the retain set. During inference, we apply this threshold to assess whether the forgetting score of each subanswer is high or low. If any subanswer yields a high forgetting score, we replace the final answer with a rejective response, following the approach of selective generation (Zhang et al., 2024a). Otherwise, the final answer is drawn from the last intermediate predicted answer.

4.2 Evaluation Results

Table 4 presents the performance of MEMMUL in multi-hop knowledge unlearning. To highlight the effectiveness of our method, we compare it with MeLLo (Zhong et al., 2023), a memory-based approach for editing multi-hop knowledge. Since MeLLo is not designed for unlearning, we modify its prompt to generate rejective responses instead of edits. Specifically, MeLLo sequentially decomposes a multi-hop question, retrieves the most relevant fact to each subquestion, and asks the model whether the retrieved fact is the same as the decomposed output. The key difference from our method lies in the decision-making process: MeLLo relies on the base LLM's judgment, whereas MEMMUL incorporates a forgetting threshold, which better aligns with unlearning. Our results show that MeLLo benefits from stronger base models but struggles with Qwen-2.5-7B-Instruct, highlighting its reliance on LLM reasoning. In contrast, MEMMUL consistently unlearns multi-hop knowledge effectively, regardless of model capacity. Notably, our 7B model performs comparably to GPT-4o, demonstrating MEMMUL's efficiency and practicality for multi-hop knowledge unlearning.

4.3 Effect of Question Decomposition

To verify the efficacy of question decomposition, we compare performance with and without it in Table 5. Given the small size of the forget set in memory (∼100 samples), distinguishing between forget and retain sets using multi-hop questions appears relatively easy, as evidenced by a notable

Question	What is the capital of the country where Greg Maddux's sport originated? (**Answer**: Washington, D.C.)
Original	Greg Maddux is a famous former professional baseball player. Baseball originated in the United States, and the capital of the United States is Washington, D.C.
NPO+RT	Greg Maddux's sport is Major League Baseball. Major League Baseball originated in the United States. The capital of the United States is Washington, D.C.
MeLLo†	The capital of the country where Greg Maddux's sport (baseball) originated is Washington, D.C.
MEMMUL	**Subquestion 1**: What sport is Greg Maddux associated with? **Subanswer 1**: Greg Maddux is associated with baseball. He was a highly successful Major League Baseball pitcher... **Retrieved Fact**: Greg Maddux is associated with the sport of baseball. (Forgetting Score: 2.12 / Threshold: 1.51) **Final Answer**: I'm unaware of that detail.

Table 6: Qualitative examples of generated outputs for a three-hop question, which contains a single-hop fact included in the forget set.

drop in the forget set. Our decomposed subquestions use coreference resolution (e.g., "What is the occupation of *that person*?"), which does not effectively retrieve relevant facts. Thus, we also evaluate subquestions without coreferences, confirming that decomposition improves fact retrieval. MEMMUL leverages subanswers, consistently outperforming other strategies.

4.4 Qualitative Analysis

Table 6 displays qualitative results for a three-hop question. As shown, the original model employs a chain-of-thought approach to solve the multi-hop question. Similarly, the NPO+RT unlearned model follows a nearly identical reasoning process, indicating that its unlearning mechanism had minimal impact on the corresponding multi-hop knowledge. During iterative prompting, MeLLo failed to recognize that one of the decomposed questions aligned with a forgotten fact. In contrast, MEMMUL successfully refrained from answering correctly by leveraging a high forgetting score.

5 Related Work

5.1 Machine Unlearning

Machine unlearning has emerged as a critical research area in response to growing concerns over data privacy, regulatory compliance, and ethical AI (Cao and Yang, 2015; Ginart et al., 2019; Bourtoule et al., 2021). In the context of LLMs, addressing memorization has garnered wide attention (Wang et al., 2023; Chen and Yang, 2023; Kassem et al., 2023; Liu et al., 2024a,b; Hong et al., 2024; Tian et al., 2024; Choi et al., 2024; Jia et al., 2024; Ji et al., 2024). The predominant approach involves maximizing prediction loss on the forget set (Jang et al., 2023; Yao et al., 2023; Lee et al., 2024; Zhang et al., 2024c; Feng et al., 2024). Other methods train LLMs to generate alternative responses, such as "I don't know" (Maini et al., 2024), random labels (Yao et al., 2024), or generic terms (Eldan and Russinovich, 2023). Recent studies have also explored task arithmetic (Ilharco et al., 2023; Bărbulescu and Triantafillou, 2024) and training-free methods that simulate unlearning through specific instructions (Thaker et al., 2024) or in-context examples (Pawelczyk et al., 2024). This work focuses on multi-hop knowledge unlearning, a novel challenge that introduces new complexities and opportunities in the field.

5.2 Multi-Hop Reasoning

Multi-hop reasoning involves connecting multiple pieces of evidence across contexts to derive information (Huang and Chang, 2023). Editing multi-hop knowledge in LLMs is challenging, as it requires consistent propagation of updates across interconnected facts (Valmeekam et al., 2022; Press et al., 2023; Dziri et al., 2023; Petty et al., 2024). Existing knowledge editing methods primarily modify individual facts (De Cao et al., 2021; Meng et al., 2022; Zhang et al., 2024b) but often struggle to update related knowledge (Onoe et al., 2023; Zhong et al., 2023; Cohen et al., 2024). Recent approaches address this by injecting information at inference time (Sakarvadia et al., 2023), removing shortcut-inducing neurons (Ju et al., 2024), or adjusting model representations to fix multi-hop reasoning errors (Ghandeharioun et al., 2024). In contrast, our work examines whether *removing* specific information from models can be generalized effectively in multi-hop scenarios.

6 Conclusion

This study explores the effectiveness of existing unlearning methods in eliminating multi-hop knowledge. Our results indicate that they struggle when an intermediate hop is unlearned. To overcome this issue, we propose MEMMUL, a simple yet effective memory-based approach that dissects multi-hop questions into subquestions and utilizes forgetting scores relative to stored facts to determine when to issue a rejective response. MEMMUL serves as a strong baseline for multi-hop knowledge unlearning, offering a highly efficient and practical solution for unlearning in LLMs.

Limitations

Building on previous training-free unlearning methods (Thaker et al., 2024; Pawelczyk et al., 2024) and memory-based multi-hop knowledge editing frameworks (Zhong et al., 2023), MEMMUL is a training-free approach that selectively refuses to answer multi-hop questions based on their relevance to forgotten facts stored in memory. Therefore, it does not essentially erase multi-hop knowledge from model parameters, meaning this information could still be extracted through advanced adversarial techniques. We emphasize that MEMMUL only serves as a baseline for multi-hop knowledge unlearning, demonstrating that effective performance can be achieved without additional training. We urge researchers and practitioners to exercise caution when attempting to fully remove multi-hop knowledge through training, as doing so may inadvertently erase broader knowledge interconnected through hops. Furthermore, we acknowledge the need for more rigorous evaluation metrics to better defend against state-of-the-art jailbreaking attacks. We hope this work stimulates further research and discussions on creating more robust frameworks for knowledge unlearning.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

George-Octavian Bărbulescu and Peter Triantafillou. 2024. To each (textual sequence) its own: Improving memorized-data unlearning in large language models. In *Forty-first International Conference on Machine Learning*.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

Lucas Bourtoule, 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, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association.

Jiaao Chen and Diyi Yang. 2023. Unlearn what you want to forget: Efficient unlearning for llms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12041–12052.

Minseok Choi, Kyunghyun Min, and Jaegul Choo. 2024. Cross-lingual unlearning of selective knowledge in multilingual language models. *arXiv preprint arXiv:2406.12354*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2024. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 12:283–298.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. 2023. Faith and fate: Limits of transformers on compositionality. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Ronen Eldan and Mark Russinovich. 2023. Who's harry potter? approximate unlearning in llms. *arXiv preprint arXiv:2310.02238*.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. T-REx: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

XiaoHua Feng, Chaochao Chen, Yuyuan Li, and Zibin Lin. 2024. Fine-grained pluggable gradient ascent for knowledge unlearning in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10141–10155, Miami, Florida, USA. Association for Computational Linguistics.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. 2024. Patchscopes: A unifying framework for inspecting hidden representations of language models. In *Forty-first International Conference on Machine Learning*.

Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. 2019. Making ai forget you: Data deletion in machine learning. *Advances in neural information processing systems*, 32.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yihuai Hong, Yuelin Zou, Lijie Hu, Ziqian Zeng, Di Wang, and Haiqin Yang. 2024. Dissecting fine-tuning unlearning in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3933–3941, Miami, Florida, USA. Association for Computational Linguistics.

Chris Jay Hoofnagle, Bart Van Der Sloot, and Frederik Zuiderveen Borgesius. 2019. The european union general data protection regulation: what it is and what it means. *Information & Communications Technology Law*, 28(1):65–98.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

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, Toronto, Canada. Association for Computational Linguistics.

Jiabao Ji, Yujian Liu, Yang Zhang, Gaowen Liu, Ramana Rao Kompella, Sijia Liu, and Shiyu Chang. 2024. Reversing the forget-retain objectives: An efficient LLM unlearning framework from logit difference. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Jinghan Jia, Yihua Zhang, Yimeng Zhang, Jiancheng Liu, Bharat Runwal, James Diffenderfer, Bhavya Kailkhura, and Sijia Liu. 2024. SOUL: Unlocking the power of second-order optimization for LLM unlearning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4276–4292, Miami, Florida, USA. Association for Computational Linguistics.

Tianjie Ju, Yijin Chen, Xinwei Yuan, Zhuosheng Zhang, Wei Du, Yubin Zheng, and Gongshen Liu. 2024. Investigating multi-hop factual shortcuts in knowledge editing of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8987–9001, Bangkok, Thailand. Association for Computational Linguistics.

Aly Kassem, Omar Mahmoud, and Sherif Saad. 2023. Preserving privacy through dememorization: An unlearning technique for mitigating memorization risks in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4360–4379, Singapore. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Dohyun Lee, Daniel Rim, Minseok Choi, and Jaegul Choo. 2024. Protecting privacy through approximating optimal parameters for sequence unlearning in language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15820–15839, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

LG AI Research, Soyoung An, Kyunghoon Bae, Eunbi Choi, Kibong Choi, Stanley Jungkyu Choi, Seokhee Hong, Junwon Hwang, Hyojin Jeon, Gerrard Jeongwon Jo, et al. 2024. Exaone 3.5: Series of large language models for real-world use cases. *arXiv preprint arXiv:2412.04862*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yujian Liu, Yang Zhang, Tommi Jaakkola, and Shiyu Chang. 2024a. Revisiting who's harry potter: Towards targeted unlearning from a causal intervention perspective. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8708–8731, Miami, Florida, USA. Association for Computational Linguistics.

Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. 2024b. Towards safer large language models through machine unlearning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1817–1829, Bangkok, Thailand. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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*.

Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Mistral AI. 2024. Un ministral, des ministraux. `https://mistral.ai/en/news/ministraux`.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2024. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*.

Yasumasa Onoe, Michael Zhang, Shankar Padmanabhan, Greg Durrett, and Eunsol Choi. 2023. Can LMs learn new entities from descriptions? challenges in propagating injected knowledge. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5469–5485, Toronto, Canada. Association for Computational Linguistics.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, Berlin, Germany. Association for Computational Linguistics.

Stuart L Pardau. 2018. The california consumer privacy act: Towards a european-style privacy regime in the united states. *J. Tech. L. & Pol'y*, 23:68.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. 2024. In-context unlearning: Language models as few-shot unlearners. In *Forty-first International Conference on Machine Learning*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Jackson Petty, Sjoerd Steenkiste, Ishita Dasgupta, Fei Sha, Dan Garrette, and Tal Linzen. 2024. The impact of depth on compositional generalization in transformer language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7239–7252, Mexico City, Mexico. Association for Computational Linguistics.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, Singapore. Association for Computational Linguistics.

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. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Jeffrey Rosen. 2011. The right to be forgotten. *Stan. L. Rev. Online*, 64:88.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Mansi Sakarvadia, Aswathy Ajith, Arham Khan, Daniel Grzenda, Nathaniel Hudson, André Bauer, Kyle Chard, and Ian Foster. 2023. Memory injections: Correcting multi-hop reasoning failures during inference in transformer-based language models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 342–356, Singapore. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Pratiksha Thaker, Yash Maurya, and Virginia Smith. 2024. Guardrail baselines for unlearning in LLMs. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*.

Bozhong Tian, Xiaozhuan Liang, Siyuan Cheng, Qingbin Liu, Mengru Wang, Dianbo Sui, Xi Chen, Huajun Chen, and Ningyu Zhang. 2024. To forget or not? towards practical knowledge unlearning for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1524–1537, Miami, Florida, USA. Association for Computational Linguistics.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2022. Large language models still can't plan (a benchmark for LLMs on planning and reasoning about change). In *NeurIPS 2022 Foundation Models for Decision Making Workshop*.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Lingzhi Wang, Tong Chen, Wei Yuan, Xingshan Zeng, Kam-Fai Wong, and Hongzhi Yin. 2023. KGA: A general machine unlearning framework based on knowledge gap alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13264–13276, Toronto, Canada. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Jin Yao, Eli Chien, Minxin Du, Xinyao Niu, Tianhao Wang, Zezhou Cheng, and Xiang Yue. 2024. Machine unlearning of pre-trained large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8403–8419, Bangkok, Thailand. Association for Computational Linguistics.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2023. Large language model unlearning. In *Socially Responsible Language Modelling Research*.

Zheng Xin Yong, Cristina Menghini, and Stephen Bach. 2023. Low-resource languages jailbreak GPT-4. In *Socially Responsible Language Modelling Research*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024a. R-tuning: Instructing large language models to say 'I don't know'. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7113–7139, Mexico City, Mexico. Association for Computational Linguistics.

Ningyu Zhang, Yunzhi Yao, and Shumin Deng. 2024b. Knowledge editing for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): Tutorial Summaries*, pages 33–41, Torino, Italia. ELRA and ICCL.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. 2024c. Negative preference optimization: From catastrophic collapse to effective unlearning. In *First Conference on Language Modeling*.

Zexuan Zhong, Zhengxuan Wu, Christopher Manning, Christopher Potts, and Danqi Chen. 2023. MQuAKE: Assessing knowledge editing in language models via multi-hop questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15686–15702, Singapore. Association for Computational Linguistics.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

A Additional Details for Knowledge Unlearning Methods

A.1 GA

Gradient ascent (GA) (Jang et al., 2023) reverses the language modeling loss, which can be understood as equivalent to gradient descent on the negative next-token prediction loss:

$$\mathcal{L}_{\text{GA}} = -\mathbb{E}_{\mathcal{D}_f}[-\log(\pi_\theta(y|x))], \tag{1}$$

which serves to minimize the token probabilities of the specific token sequences in the forget set $\mathcal{D}_f$.

A.2 DPO

In direct preference optimization (DPO) (Rafailov et al., 2023), we are provided with a dataset of preference feedbacks $\mathcal{D}_{\text{paired}} = \{(x_i, y_{i,w}, y_{i,l})\}_{i=1}^N$, where "w" stands for "win" and "l" stands for "lose" for two responses y_w and y_l. The goal is to train the model π_θ to align more closely with human preferences. In this work, the winning responses are rejections (e.g., "I don't know."), randomly sampled from 100 candidates used by Maini et al. (2024). Formally, DPO minimizes

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{\mathcal{D}_{\text{paired}}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right], \tag{2}$$

where $\sigma(t) = 1/(1+e^{-t})$ represents the sigmoid function, $\beta > 0$ is the inverse temperature, and ϕ_{ref} is a reference model.

A.3 NPO

Negative preference optimization (NPO) (Zhang et al., 2024c) ignores the y_w term in DPO in Equation 2 and aligns the language model with negative responses exclusively:

$$\mathcal{L}_{\text{NPO}} = -\mathbb{E}_{\mathcal{D}_f}\left[\log\sigma\left(-\beta\log\frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}\right)\right]. \tag{3}$$

Minimizing $\mathcal{L}_{\text{NPO}}$ drives the prediction probability $\pi_\theta(y|x)$ on the forget set to be as low as possible, effectively achieving the goal of unlearning the forget set.

A.4 +RT

The explicit retention finetuning is achieved through standard language modeling on the retain set, which serves as the positive counterpart to Equation 1:

$$\mathcal{L}_r = -\mathbb{E}_{\mathcal{D}_r}[\log(\pi_\theta(y|x))]. \tag{4}$$

Finally, the overall training objective is minimizing the following loss:

$$\mathcal{L} = \alpha\cdot\mathcal{L}_f + (1-\alpha)\cdot\mathcal{L}_r, \tag{5}$$

where $\mathcal{L}_f$ is one of the unlearning losses $\mathcal{L}_{\text{GA}}$, $\mathcal{L}_{\text{DPO}}$, or $\mathcal{L}_{\text{NPO}}$, and α is a loss scaling hyperparameter balancing the forgetting and retaining losses.

B Dataset Details

We describe further details in datasets and their preprocessing processes for our experiments.

MQuAKE (Zhong et al., 2023) considers whether models can adapt to updates in factual knowledge by modifying their responses to multi-hop queries when individual facts are altered. The benchmark comprises two datasets: MQuAKE-CF, which focuses on counterfactual scenarios, and MQuAKE-T, which addresses temporal knowledge updates by replacing outdated facts with current information. Both datasets are based on Wikidata and consist of knowledge triplets for single-hop reasoning, as well as multi-hop chains derived from these triplets. Each instance in the benchmark includes: (1) an edit set of single-hop knowledge triplets $(s, r, o \rightarrow o^*)$, where o^* represents the updated object; (2) a chain of facts $\mathcal{C}$ and its updated version $\mathcal{C}^*$ after knowledge editing; and (3) questions about both the single-hop knowledge and multi-hop chains before and after knowledge updates. We used the merged set of both MQuAKE-CF and MQuAKE-T for our experiments and only used the single-hop and multi-hop triplets before the knowledge update.

MuSiQue (Trivedi et al., 2022) is a benchmark designed to ensure that multi-hop QA models genuinely perform multi-hop reasoning rather than relying on single-hop shortcuts. It constructs multi-hop questions by carefully composing independent single-hop questions, ensuring that models must retrieve and integrate information from multiple documents. Built on Wikipedia, MuSiQue provides a challenging evaluation for multi-hop QA models. We used the answerable subset of MuSiQue, referred to as MuSiQue-Ans, for our experiments. We noticed that some single-hop triplets in MuSiQue lack a natural language question and are instead represented only by a subject-relation (s, r) pair. To address this, we used `gpt-4o-mini` with few-shot demonstrations to generate a corresponding question from each (s, r) pair. For example, the pair *(Just Ask Your Heart, performer)* is transformed into the question: *"Who performed the song 'Just Ask Your Heart'?"*.

To adapt both datasets for multi-hop knowledge unlearning, we preprocessed the data by: (1) splitting the single-hop triplets into a forget set and a retain set at a predefined ratio (1:9), with the retain set further divided into training, validation, and test splits and (2) We also ensured that multi-hop questions are linked to the corresponding single-hop triplets from both the forget and retain sets. If a multi-hop question contains both a forget and retain triplet, it was assigned exclusively to the forget set, ensuring that the final multi-hop forget and retain sets remain mutually exclusive. For MQuAKE dataset, as the frequency of single-hop triplets is imbalanced, we ensured that any triplet involved in more than two multi-hop questions is assigned to the retain set's training split.

C Full Evaluation Results

We report additional experimental results with different open-source LLMs including Phi-3.5-Mini-Instruct (Abdin et al., 2024), EXAONE-3.5-7.8B-Instruct (LG AI Research et al., 2024), Llama-3.1-8B-Instruct (Dubey et al., 2024), and Ministral-8B-Instruct (Mistral AI, 2024) in Tables 7 and 8. Furthermore, we display the utility performance of the model according to individual LLM benchmarks in Tables 9 and 10.

D Licenses and Terms of Use for Artifacts

We utilized multiple datasets and open-source LLMs, each governed by specific licensing terms. The MQuAKE repository is available under the MIT License, while MuSiQue is licensed under CC BY 4.0, permitting academic and research use. All open-source LLMs referenced in this paper, including OLMo, Qwen, Phi, EXAONE, Llama, and Ministral, are freely licensed for research purposes.

E Use of AI Assistants

We leveraged AI assistants, including ChatGPT and Copilot, to enhance research, writing, and coding. ChatGPT played a key role in refining the paper's narrative and ensuring clarity, while Copilot accelerated the coding process. All AI tools were used responsibly, with careful oversight to uphold the integrity and originality of the research. Their usage has been documented to ensure transparency and proper acknowledgment of AI contributions.

	Forget Set (Single-Hop)			Forget Set (Multi-Hop)			Retain Set (Single-Hop)			Retain Set (Multi-Hop)			Utility Set
	PA(↓)	R-L(↓)	LM(↑)	PA(↓)	R-L(↓)	LM(↑)	PA(↑)	R-L(↑)	LM(↓)	PA(↑)	R-L(↑)	LM(↓)	Avg.(↑)
Phi-3.5-Mini-Instruct													
Original	82.7	83.3	3.8	79.8	56.2	3.3	87.7	80.8	3.7	81.1	51.6	3.4	65.0
GA	24.0	15.6	24.4	36.5	7.4	20.4	28.4	19.2	22.9	32.0	3.2	20.5	62.6
DPO	31.7	0.7	12.0	43.3	0.2	8.3	38.1	0.9	11.4	41.8	0.4	8.2	64.0
NPO	22.1	19.5	24.4	33.7	6.2	20.3	26.9	18.4	22.9	32.5	4.1	20.3	62.7
GA+RT	39.7	41.1	7.8	87.8	31.8	3.1	78.6	61.7	3.0	**88.9**	36.6	2.8	**64.7**
DPO+RT	44.9	**7.8**	6.7	**77.2**	**4.6**	3.2	**80.4**	30.0	**2.7**	80.6	6.6	**2.7**	64.6
NPO+RT	**39.1**	44.9	**8.5**	87.2	33.5	**3.5**	78.2	**65.5**	3.2	87.7	**38.7**	3.2	64.6
EXAONE-3.5-7.8B-Instruct													
Original	92.3	82.4	3.6	97.1	49.1	3.3	90.2	76.6	3.5	95.5	46.1	3.4	62.6
GA	22.1	0.0	87.4	21.2	0.0	88.4	24.9	0.0	87.4	18.4	0.0	88.4	58.7
DPO	18.3	0.6	53.2	17.3	0.9	52.8	20.3	1.0	52.8	14.3	0.7	53.0	60.4
NPO	21.2	0.0	86.9	19.2	0.0	87.9	25.1	0.0	86.8	18.7	0.0	87.9	58.7
GA+RT	35.3	44.2	**12.3**	86.2	40.8	**3.9**	79.2	66.4	4.2	91.8	38.2	3.2	**62.1**
DPO+RT	**30.4**	**3.6**	9.2	**77.9**	**1.8**	3.3	**80.1**	23.7	**2.4**	89.5	4.8	**2.5**	**62.1**
NPO+RT	37.2	42.2	12.3	85.3	41.0	3.9	79.8	**67.6**	4.0	**92.1**	**38.8**	3.2	**62.1**
Llama-3.1-8B-Instruct													
Original	99.0	60.8	0.6	99.0	31.2	1.0	98.9	60.1	0.6	98.4	28.6	1.0	64.7
GA	33.7	0.0	88.7	28.8	0.0	87.0	34.3	0.0	88.3	24.5	0.0	87.1	62.9
DPO	10.6	2.6	39.8	11.5	2.4	39.0	15.5	2.2	39.5	12.5	1.4	39.0	61.8
NPO	30.8	0.0	89.1	30.8	0.0	87.7	33.7	0.0	88.9	25.8	0.0	87.7	62.8
GA+RT	56.4	49.4	**13.8**	87.2	28.7	5.8	86.6	**79.3**	5.7	**91.2**	**29.9**	**4.9**	**64.7**
DPO+RT	**49.4**	**9.3**	13.3	**86.9**	**9.8**	**6.1**	**88.3**	44.7	5.1	90.7	13.2	5.0	63.7
NPO+RT	56.7	48.0	12.5	88.1	26.3	5.7	87.5	78.5	**4.9**	90.1	29.6	4.9	**64.7**
Ministral-8B-Instruct													
Original	98.1	85.3	0.9	99.0	46.9	0.8	97.5	84.8	0.8	98.2	41.9	0.8	64.5
GA	23.1	0.0	69.6	28.8	0.0	70.6	20.5	0.0	69.7	23.6	0.0	70.6	41.9
DPO	15.4	0.0	38.7	21.2	0.9	38.1	12.3	1.6	38.6	14.0	0.7	38.2	44.0
NPO	22.1	0.0	69.8	30.8	0.0	70.8	21.0	0.0	69.8	23.5	0.0	70.7	42.0
GA+RT	**34.0**	17.1	**18.4**	77.9	9.8	**10.5**	76.9	41.3	6.9	87.8	**13.9**	6.0	**59.3**
DPO+RT	36.9	**1.2**	14.3	**75.3**	**1.4**	7.6	**82.8**	5.1	5.1	88.4	2.4	4.9	56.9
NPO+RT	36.2	19.6	14.8	81.1	11.6	7.4	80.8	**45.6**	**4.6**	**90.2**	13.6	**4.2**	59.2

Table 7: Performance comparison of different knowledge unlearning methods after erasing single-hop facts from the forget set in MQuAKE-NoEdit. The best results amongst +RT models are highlighted in **bold**.

	Forget Set (Single-Hop)			Forget Set (Multi-Hop)			Retain Set (Single-Hop)			Retain Set (Multi-Hop)			Utility Set
	PA(↓)	**R-L**(↓)	**LM**(↑)	**PA**(↓)	**R-L**(↓)	**LM**(↑)	**PA**(↑)	**R-L**(↑)	**LM**(↓)	**PA**(↑)	**R-L**(↑)	**LM**(↓)	**Avg.**(↑)
Phi-3.5-Mini-Instruct													
Original	67.6	34.5	4.1	67.7	23.0	3.4	71.0	34.0	4.1	69.9	22.7	3.4	65.0
GA	21.4	7.1	61.8	19.0	7.9	52.6	19.4	5.4	61.3	19.5	6.7	53.5	64.1
DPO	11.6	0.1	52.0	11.7	0.5	43.3	14.6	0.2	51.4	13.0	0.5	44.2	64.5
NPO	20.2	7.3	62.9	20.2	8.5	53.7	19.4	5.0	62.4	19.3	6.5	54.6	64.0
GA+RT	31.2	17.4	11.6	77.2	16.2	4.8	**82.5**	28.0	**3.3**	**83.9**	18.0	**4.2**	64.1
DPO+RT	**22.4**	**2.2**	**14.5**	**63.5**	**1.7**	**6.1**	73.4	12.2	5.7	70.6	2.3	5.7	**64.5**
NPO+RT	30.8	19.1	13.1	75.8	18.2	5.6	81.0	**29.4**	3.8	82.2	**20.0**	4.9	64.0
EXAONE-3.5-7.8B-Instruct													
Original	76.3	34.6	4.0	90.1	18.9	3.5	77.0	33.2	4.0	89.4	21.1	3.6	62.6
GA	24.3	0.0	105.4	14.4	0.0	107.6	27.9	0.0	105.4	16.1	0.0	107.4	53.9
DPO	16.8	1.4	79.1	12.1	3.0	82.1	19.9	1.0	78.7	13.0	2.1	82.0	57.8
NPO	24.9	0.0	99.0	13.8	0.0	100.7	28.1	0.0	99.0	15.7	0.0	100.6	54.0
GA+RT	23.5	13.7	**20.7**	**70.6**	14.5	**9.7**	80.1	23.0	5.9	84.0	16.5	7.2	59.5
DPO+RT	**11.6**	**1.6**	12.1	72.7	**3.4**	4.0	80.8	9.1	**2.6**	**88.1**	4.4	**2.6**	**61.7**
NPO+RT	24.1	15.2	19.5	72.1	16.4	8.9	**81.5**	**26.7**	5.1	85.6	**18.8**	6.4	59.8
Llama-3.1-8B-Instruct													
Original	90.2	35.9	1.1	96.5	16.5	1.3	90.9	32.8	1.1	96.2	16.1	1.3	64.7
GA	30.1	0.0	103.0	30.0	0.0	102.0	29.4	0.0	103.0	33.4	0.0	102.1	60.5
DPO	28.3	1.4	73.7	31.4	1.1	70.7	28.9	0.6	73.6	31.0	0.9	71.1	59.9
NPO	25.4	0.0	99.1	23.8	0.0	98.3	26.0	0.0	99.1	25.5	0.0	98.4	60.5
GA+RT	**32.2**	12.0	**28.6**	**77.1**	11.8	**14.4**	80.7	23.7	9.5	86.4	13.3	10.4	61.7
DPO+RT	34.3	**5.2**	24.3	79.1	**4.8**	11.9	**85.0**	18.7	6.6	90.1	6.3	8.3	**63.5**
NPO+RT	37.4	14.9	23.0	84.4	14.0	10.3	**85.0**	**31.7**	**6.4**	**91.4**	**18.3**	**7.1**	**63.5**
Ministral-8B-Instruct													
Original	88.4	40.8	1.1	96.3	23.4	1.0	87.2	38.7	1.1	96.2	23.1	1.0	64.5
GA	22.5	0.1	88.8	25.7	0.2	89.5	32.5	0.0	88.8	30.4	0.3	89.4	30.7
DPO	35.5	0.6	58.0	28.8	1.0	60.9	34.4	0.5	57.7	28.9	0.5	60.6	36.2
NPO	23.1	0.1	87.1	24.8	0.2	87.6	30.4	0.0	87.0	29.2	0.3	87.5	31.1
GA+RT	28.9	7.9	**21.8**	78.9	10.2	**11.7**	78.4	16.1	7.8	87.1	10.3	8.1	**55.4**
DPO+RT	**24.3**	**3.7**	15.7	**79.4**	**3.6**	7.6	82.5	9.9	**4.5**	89.5	4.4	**5.1**	55.3
NPO+RT	28.1	10.4	17.5	81.8	12.3	7.8	**82.9**	**19.8**	4.9	**91.2**	**12.2**	5.1	**55.4**

Table 8: Performance comparison of different knowledge unlearning methods after erasing single-hop facts from the forget set in MuSiQue-Ans. The best results amongst +RT models are highlighted in **bold**.

	ARC-C	CSQA	Hella.	Lamba.	MMLU	OBQA	PIQA	Wino.	Avg.
OLMo-2-7B-Instruct									
Original	54.6	72.6	65.7	68.6	59.2	39.4	80.5	71.6	64.0
GA+RT	50.8	71.3	64.4	69.5	59.2	39.1	75.7	70.3	62.6
DPO+RT	50.9	70.7	64.5	66.4	58.9	39.2	77.5	71.4	62.4
NPO+RT	50.9	71.3	64.5	69.7	59.2	38.9	75.8	70.3	62.6
Qwen-2.5-7B-Instruct									
Original	52.6	82.6	62.1	69.4	71.8	34.8	79.3	71.5	65.5
GA+RT	56.3	83.4	62.1	71.8	71.5	37.7	79.0	70.4	66.5
DPO+RT	53.4	81.2	60.6	67.9	71.2	35.8	79.5	71.5	65.1
NPO+RT	56.1	83.2	62.1	71.9	71.5	37.3	79.1	70.7	66.5
Phi-3.5-Mini-Instruct									
Original	59.5	75.3	58.8	65.1	68.7	37.6	80.0	74.6	65.0
GA+RT	58.9	75.3	59.5	63.3	68.6	39.1	79.0	73.6	64.7
DPO+RT	59.8	74.4	58.2	60.4	68.5	38.4	80.3	76.4	64.6
NPO+RT	58.8	75.0	59.6	63.5	68.6	39.2	78.9	73.4	64.6
EXAONE-3.5-7.8B-Instruct									
Original	56.9	75.4	60.2	62.4	65.2	35.6	76.9	68.0	62.6
GA+RT	56.0	75.0	59.7	61.5	64.6	36.9	76.3	67.3	62.1
DPO+RT	57.6	75.2	59.6	55.8	64.4	36.7	77.7	69.7	62.1
NPO+RT	55.7	75.2	59.7	61.3	64.7	36.7	76.1	67.3	62.1
Llama-3.1-8B-Instruct									
Original	51.8	77.1	59.2	73.2	68.1	33.8	80.2	74.1	64.7
GA+RT	53.2	75.8	59.2	74.7	67.2	35.9	79.1	72.7	64.7
DPO+RT	51.1	74.7	59.2	70.1	66.3	34.9	80.2	73.2	63.7
NPO+RT	53.5	75.5	59.3	74.7	67.3	35.7	79.2	72.8	64.7
Ministral-8B-Instruct									
Original	54.6	72.5	59.6	74.6	64.1	36.2	81.0	73.6	64.5
GA+RT	53.0	38.6	58.4	77.6	55.5	39.2	80.6	71.3	59.3
DPO+RT	51.8	46.4	57.8	49.7	59.2	38.4	80.7	71.3	56.9
NPO+RT	52.8	37.1	58.5	77.5	55.9	39.5	80.8	71.4	59.2

Table 9: Model utility performance per task after erasing single-hop facts from the forget set in MQuAKE-NoEdit.

	ARC-C	CSQA	Hella.	Lamba.	MMLU	OBQA	PIQA	Wino.	Avg.
OLMo-2-7B-Instruct									
Original	54.6	72.6	65.7	68.6	59.2	39.4	80.5	71.6	64.0
GA+RT	47.2	70.1	63.3	69.8	58.6	38.3	73.1	67.8	61.0
DPO+RT	46.8	69.2	63.2	68.1	58.6	37.4	73.1	68.3	60.6
NPO+RT	47.4	70.5	63.4	70.2	58.7	38.1	73.3	67.9	61.2
Qwen-2.5-7B-Instruct									
Original	52.6	82.6	62.1	69.4	71.8	34.8	79.3	71.5	65.5
GA+RT	51.9	83.3	62.5	71.6	71.5	36.6	76.4	68.5	65.3
DPO+RT	52.9	81.7	61.8	69.1	71.4	36.7	78.8	69.5	65.2
NPO+RT	52.0	83.3	62.6	71.5	71.5	36.5	76.5	68.2	65.3
Phi-3.5-Mini-Instruct									
Original	59.5	75.3	58.8	65.1	68.7	37.6	80.0	74.6	65.0
GA+RT	57.7	73.1	60.5	65.0	68.7	38.4	76.9	72.4	64.1
DPO+RT	60.5	74.5	58.7	61.8	68.6	38.2	79.8	74.0	64.5
NPO+RT	57.5	72.8	60.6	64.8	68.8	38.3	76.9	72.3	64.0
EXAONE-3.5-7.8B-Instruct									
Original	56.9	75.4	60.2	62.4	65.2	35.6	76.9	68.0	62.6
GA+RT	50.2	75.4	57.6	59.7	64.7	34.7	71.0	62.3	59.5
DPO+RT	56.5	75.0	59.6	57.0	64.5	36.7	76.9	67.0	61.7
NPO+RT	50.4	75.5	57.9	59.9	64.8	35.2	71.7	63.0	59.8
Llama-3.1-8B-Instruct									
Original	51.8	77.1	59.2	73.2	68.1	33.8	80.2	74.1	64.7
GA+RT	43.4	73.1	58.6	75.8	66.0	35.8	70.9	70.2	61.7
DPO+RT	48.1	74.6	59.0	75.1	65.5	36.3	78.1	71.5	63.5
NPO+RT	48.0	74.4	59.3	76.5	65.9	36.5	75.5	71.7	63.5
Ministral-8B-Instruct									
Original	54.6	72.5	59.6	74.6	64.1	36.2	81.0	73.6	64.5
GA+RT	43.6	30.5	60.3	77.0	55.0	34.6	73.9	68.1	55.4
DPO+RT	44.2	32.7	59.8	70.8	58.0	34.9	76.1	65.6	55.3
NPO+RT	43.6	29.4	60.3	77.5	54.2	35.3	75.0	68.2	55.4

Table 10: Model utility performance per task after erasing single-hop facts from the forget set in MuSiQue-Ans.

[System]
You are an AI assistant helping a user decompose a multi-hop question into a series of single-hop questions. The number of single-hop questions may vary. Each single-hop question should be answerable by a short word or phrase. Questions should be connected in a logical sequence, where the coreference resolution is maintained.

[User]
Who is the ex-wife of the founder of Microsoft?

[Assistant]
Single-Hop Question 1: Who is the founder of Microsoft?
Single-Hop Question 2: Who is the ex-wife of that person?

[User]
What is the continent of the country of origin of the sport Tottenham Hotspur F.C. plays?

[Assistant]
Single-Hop Question 1: What is the sport that Tottenham Hotspur F.C. plays?
Single-Hop Question 2: What is the country of origin of that sport?
Single-Hop Question 3: What is the continent of that country?

[User]
What is the capital of the country where the CEO of the developer of the game Fortnite was born?

[Assistant]
Single-Hop Question 1: Who is the developer of the game Fortnite?
Single-Hop Question 2: Who is the CEO of that company?
Single-Hop Question 3: What is the country where that person was born?
Single-Hop Question 4: What is the capital of that country?

[User]
{multi-hop question}

[Assistant]

Figure 5: Prompt used in MEMMUL to decompose a multi-hop question into a series of subquestions using GPT-4o. It consists of a system prompt followed by three fixed demonstration examples.