The Twin Pillars of LLM Unlearning Audit: Ensuring Adequacy and Non-Redundancy through Knowledge Graphs

Anonymous ACL submission

Abstract

In recent years, LLMs have faced increasing demands to selectively remove sensitive information, protect privacy, and comply with copyright regulations through Machine Unlearning. While evaluating unlearning effectiveness is crucial, existing benchmarks are limited. We identify two critical challenges in generating holistic audit datasets: ensuring audit adequacy and handling knowledge redundancy between forget and retain dataset. To address these challenges, we propose HANKER, an automated framework for holistic audit dataset generation leveraging knowledge graphs to achieve fine-grained coverage and eliminate redundant knowledge. Applying HANKER to the popular MUSE benchmark, we successfully generated over 69,000 and 111,000 audit cases for the News and Books datasets respectively, identifying thousands of knowledge memorization instances that the previous benchmark failed to detect. Our empirical analysis uncovers how knowledge redundancy significantly skews unlearning effectiveness metrics, with redundant instances artificially inflating the observed memorization measurements ROUGE from 19.7% to 26.1% and Entailment Scores from 32.4% to 35.2%, highlighting the necessity of systematic deduplication for accurate assessment.

1 Introduction

In recent years, Large Language Models (LLMs) have undergone rapid development, demonstrating impressive capabilities across a wide range of applications, from natural language processing to code generation and complex problem-solving (Liu et al., 2023; Satpute et al., 2024). However, these advances have raised concerns about potential risks associated with the vast knowledge stored in these models, e.g., the inadvertent retention of personally identifiable information (PII) (Jang et al., 2022), the propagation of unsafe or biased behaviors (Liu et al., 2024e), and the unauthorized use of copyrighted content (Eldan and Russinovich, 2023). Furthermore, there is an increasing imperative for LLMs to comply with regulatory standards such as the General Data Protection Regulation (GDPR) (Hoofnagle et al., 2019), which enforces the "Right to be Forgotten" (Dang, 2021). To address these concerns, researchers are investigating various unlearning techniques (Jia et al., 2024a) to selectively remove specific knowledge from pretrained LLMs while preserving their general language modeling capabilities, thereby avoiding the substantial computational costs associated with building new models from scratch.

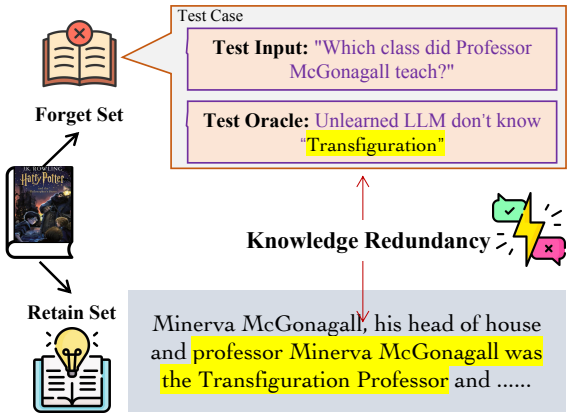


Figure 1: An illustrative example from MUSE demonstrating where knowledge targeted for forgetting also appears in the Retain Dataset, highlighting the challenge of knowledge redundancy in unlearning evaluation.

The growing significance of LLM unlearning has hignlighted the importance of rigorous evaluation or audit of unlearing performance. Recent benchmarks like MUSE (Shi et al., 2024) and TOFU (Maini et al., 2024) assess unlearning efficacy across multiple dimensions, ranging from verbatim text retention to embedded knowledge preservation. These pioneering frameworks have advanced the field by establishing standard-

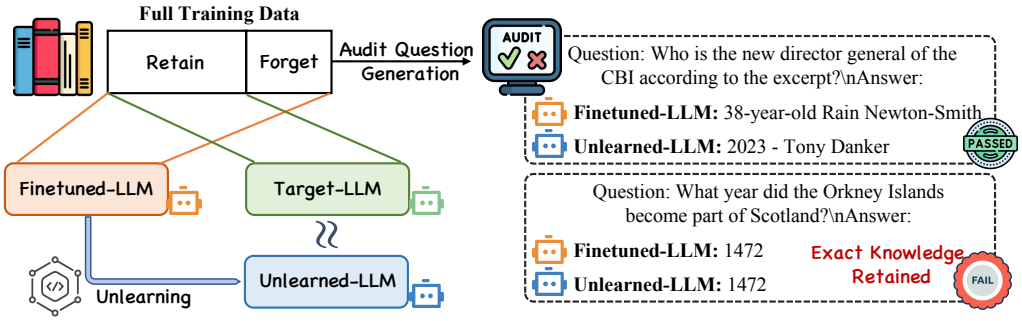


Figure 2: Illustration of the basic pipeline for LLM knowledge unlearning and its audit.

ized datasets, providing pre-trained target models, and introducing multifaceted evaluation metrics. However, their audit suites remain constrained in scope—for instance, MUSE employs only 100 test questions to evaluate 0.8M corpora. From an auditing perspective, such limited test coverage may inadequately assess the targeted knowledge removal, potentially compromising the comprehensive evaluation of unlearning effectiveness.

Our investigation reveals two fundamental challenges in holistic audit dataset synthesis. The primary concern about ***audit adequacy*** stems from simply relying on GPT-4 for automated QA generation from forget corpora. While this approach can generate multiple question-answer pairs for each target text, it introduces significant uncertainty in whether the generated questions comprehensively cover all the critical information contained within the source text. The second challenge involves ***knowledge redundancy*** between forget and retain corpora. As illustrated in Figure 2, shared knowledge should be preserved during an ideal exact unlearning process. However, current evaluation methods fail to account for test cases where the information targeted also appears in the retain dataset, as demonstrated in Figure 1.

In this paper, we propose HANKER, a novel automated framework for holistic audit dataset generation that leverages knowledge graphs (KGs) to address the aforementioned limitations. Benefiting from advances in named entity recognition and information extraction, various tools now enable efficient conversion of unstructured text into structured entity-relation graphs. HANKER first converts both forget and retain corpora into structural knowledge graphs. By treating each KG edge (i.e., one fact) as a minimal unit, we can explicitly control the coverage of the audit process. Subsequently, by identifying and eliminating identical facts within the forget and retain KGs, we remove redundant knowledge from the forget KG, ensuring a well-defined audit scope. Finally, HANKER utilizes specific facts to guide LLMs in generating high-quality, targeted test questions, guaranteeing comprehensive and accurate auditing. Through this pipeline, HANKER automatically generates large-scale, comprehensive audit datasets for any given forget and retain corpora, thereby providing robust support for LLM unlearning evaluation.

In summary, our contributions are as follows:

- We introduce HANKER[1], a novel and automated framework for generating holistic audit datasets for LLM knowledge unlearning, which addresses the challenge of audit adequacy and knowledge redundancy.
- We apply HANKER to popular benchmark MUSE, significantly expanding the dataset scale and identifying knowledge memorization cases in unlearned LLMs that exceeded previous findings by three orders of magnitude ($10^3\times$).
- Our experimental results reveal that knowledge redundancy has a substantial impact on the assessment of unlearning effectiveness.

2 Preliminaries and Motivation

2.1 LLM Unlearning

LLM unlearning refers to techniques that selectively remove specific behaviors or knowledge from a pre-trained language model while maintaining its overall functionality (Yao et al., 2023). With the proliferation of LLMs, unlearning has gained significant attention due to its broad applications

[1] https://anonymous.4open.science/r/HANKER-FB86

in safety alignment, privacy protection, and copyright compliance (Eldan and Russinovich, 2023; Liu et al., 2024c; Jia et al., 2024b). The evaluation and auditing of LLM unlearning spans from basic verbatim memorization to deeper knowledge memorization (Shi et al., 2024), with this work focusing on the latter. As depicted in Figure 2, LLM unlearning operates as a targeted intervention within the model's knowledge representation framework. Its core objective is the selective removal of specific information while preserving the model's broader knowledge base (e.g, on retain set). This study focuses on the knowledge unlearning auditing that assesses unlearned models' behaviors through comprehensive audit cases. Given access to both forget and retain corpora, we generate a holistic set of test questions with reference answers to thoroughly evaluate whether an unlearned model exhibits any residual knowledge memorization.

2.2 Knowledge Graph

A knowledge graph (KG) is a structured multi-relational graph (Bordes et al., 2013), usually representing a collection of facts as a network of entities and the relationships between entities. Formally, a KG $\mathcal{G} = \langle \mathcal{E}, \mathcal{R}, \mathcal{F} \rangle$ could be considered a directed edge-labeled graph (Ji et al., 2021), which comprises a set $\mathcal{E}$ of entities (e.g., *Harry Potter*, *Hogwarts School*), a set $\mathcal{R}$ of relations (e.g., *attends*), and a set $\mathcal{F}$ of facts. A fact is a triple containing the head entity $e_1 \in \mathcal{E}$, the relation $r \in \mathcal{R}$, and the tail entity $e_2 \in \mathcal{E}$ to show that there exists the relation from the tail entity to the head entity, denoted as $(e_1, r, e_2) \in \mathcal{F}$ (Hogan et al., 2021). To illustrate, the fact (*Harry Potter*, *attends*, *Hogwarts School*) shows that there exists the *attends* relation between *Harry Potter* and *Hogwarts School*, which indicates"Harry Potter attends Hogwarts School".

2.3 Motivation

This section aims to illustrate why and how we consider employing KG to facilitate the holistic LLM unlearning audit. Two critical factors underpin this task. ❶**Audit Adequacy**: The Forget Dataset is an extensive, unstructured corpus. Existing benchmarks typically rely on the LLM's prior knowledge to directly generate QA pairs or segment the corpus and feed these segments to ChatGPT for automated QA pair generation. Such works like MUSE often fail to intuitively reflect and guarantee the sufficiency, as shown in § A.4. ❷**Knowledge Redundancy**: A more subtle and easily overlooked issue is that the Retain Dataset and Forget Dataset may contain overlapping knowledge. As illustrated in Figure 2, this overlapping knowledge should be retained by the unlearned model and, therefore not be treated as candidates for the unlearning efficacy audit. Existing evaluation benchmarks like MUSE often neglect this aspect, as evidenced by Figure 1.

A KG can offer an effective solution to address these two challenges. First, the KG inherently captures the knowledge facts within the Forget Dataset at a fine-grained level, with each edge representing a minimal testable unit. By ensuring coverage of every edge in the KG, one can achieve a more intuitive and relatively comprehensive audit. Moreover, the structured data provided by the KG can facilitate the identification of identical knowledge facts present in both the Retain and Forget Datasets. This capability allows for refinement of the initial forget knowledge graph by removing potentially retained information. Finally, owing to recent advances in KG extraction technology, numerous automated extraction models and pipelines are available to support the automated construction of an audit dataset.

3 Proposed Method

The core idea behind HANKER is to leverage knowledge graphs to achieve fine-grained and comprehensive test coverage, while rigorously eliminating redundancy between the forgetting and retain objectives. As illustrated in Figure 3, HANKER comprises three sequential stages. During the *Knowledge Graph Construction* stage, unstructured textual data is systematically transformed into structured knowledge representations. This enables the explicit modeling of atomic knowledge units and their semantic interconnections. Subsequently, the *Redundancy Removal* stage meticulously identifies and eliminates knowledge facts that are simultaneously present in both forget and retain datasets. This process helps prevent inaccurate assessments by ensuring the audit doesn't mistakenly flag knowledge meant for retain as candidates for removal. Finally, in the *Question Synthesis* stage, HANKER employs LLMs to generate targeted questions and corresponding reference answers, guided by specific knowledge facts from the pruned knowledge graph. This approach provides an automated and holistic evaluation framework for assessing LLM knowledge unlearning efficacy.

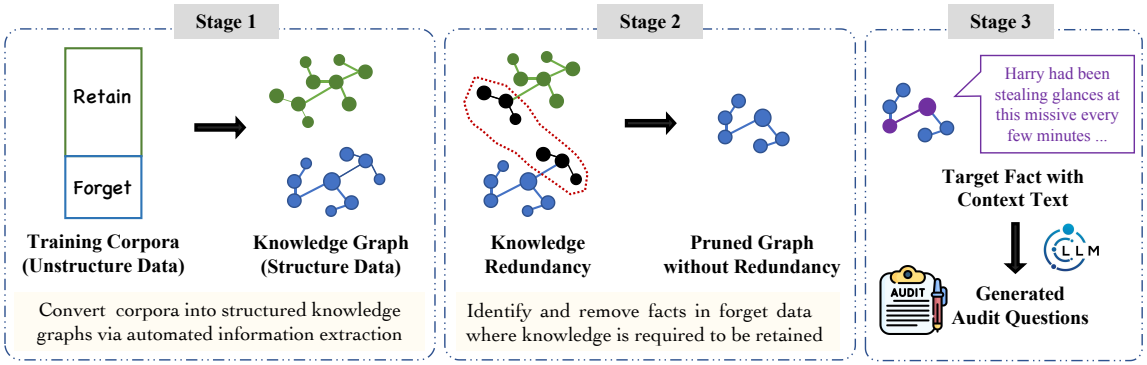


Figure 3: Overview of the proposed HANKER. The framework consists of three stages: (1) **Knowledge Graph Construction** that extracts structured knowledge from forget and retain data, (2) **Redundancy Removal** that identifies and removes redundant knowledge from the constructed knowledge graphs, and (3) **Question Synthesis** that generates QA pairs with the guidance of specific facts with LLMs automatically.

Algorithm 1 HANKER

Input: Forget dataset D_{fgt}, Retain dataset D_{ret}
Output: Audit suite S

function GENERATION(D_{fgt}, D_{ret})
  ▷ Knowledge Graph Construction
  $G_{\text{fgt}} \leftarrow$ KGExtraction(D_{fgt})
  $G_{\text{ret}} \leftarrow$ KGExtraction(D_{ret})
  ▷ Redundancy Removal
  $G_{\text{test}} \leftarrow \emptyset$
  for all $e \in G_{\text{fgt}}$ **do**
    if $e \notin G_{\text{ret}}$ **then**
      $G_{\text{test}} \leftarrow G_{\text{test}} \cup \{e\}$
  ▷ Question Synthesis
  $S \leftarrow \emptyset$
  for all $e \in G_{\text{test}}$ **do**
    $ctx \leftarrow$ RetrieveContext(e)
    $prompt \leftarrow$ ComposePrompt(e, ctx)
    $qa \leftarrow$ LLM($prompt$)
    $S \leftarrow S \cup \{qa\}$
  return S

3.1 Stage 1: Knowledge Graph Construction

Our framework transforms unstructured text corpora into structured knowledge graphs to enable fine-grained knowledge evaluation. This transformation is crucial for capturing semantic relationships and facilitating precise knowledge auditing. Specifically, we construct two distinct knowledge graphs from the forget and retain datasets: $\mathcal{G}_{\text{fgt}}$ and $\mathcal{G}_{\text{ret}}$, respectively. Each knowledge graph represents a structured network of entities and their relationships, allowing for systematic analysis of knowledge units. For implementation, following standard practices, we first segment the input text and perform coreference resolution preprocessing (Lee et al., 2017), to ensure accurate entity identification and relationship mapping. We then employ the REBEL-large model (Huguet Cabot and Navigli, 2021), which has been specifically fine-tuned for entity and relation extraction. This model demonstrates robust performance in extracting structured knowledge from natural language text, making it particularly suitable for our knowledge graph construction pipeline.

3.2 Stage 2: Redundancy Removal

The intricate entanglement of information across retain and forget datasets complicates the identification of specific elements requiring audit. To address this challenge, we implement a graph alignment strategy to detect shared information between $\mathcal{G}_{\text{fgt}}$ and $\mathcal{G}_{\text{ret}}$. We identify redundancy through triples that match exactly across both graphs. Concretely, each directed edge is represented as a triple (e_1, r, e_2), and we mark an edge as redundant if the same entity pair and relation appear in both $\mathcal{G}_{\text{fgt}}$ and $\mathcal{G}_{\text{ret}}$. Our method examines each triple $(e_1, r, e_2) \in \mathcal{G}_{\text{fgt}}$ to locate its potential counterpart in $\mathcal{G}_{\text{ret}}$. We express the overlapping edges mathematically as:

$$E_{\text{conf}} = E(\mathcal{G}_{\text{fgt}}) \cap E(\mathcal{G}_{\text{ret}}). \quad (1)$$

The refined test graph is then constructed by removing these intersecting elements:

$$\mathcal{G}_{\text{test}} = \mathcal{G}_{\text{fgt}} \setminus E_{\text{conf}}. \quad (2)$$

This process yields $\mathcal{G}_{\text{test}}$, which maintains the fundamental structure of $\mathcal{G}_{\text{fgt}}$ but excludes direct knowledge overlap with $\mathcal{G}_{\text{ret}}$. The resulting graph provides a clean foundation for assessing selective forgetting performance, preserving crucial network relationships while eliminating redundant elements. It is important to note that this step provides an approximation rather than a perfectly precise identification of redundant knowledge. Even if two facts

appear to be identical, their meanings may vary depending on the surrounding context, making exact equivalence challenging to determine. Nevertheless, the distant supervision strategy employed here has been shown to effectively capture the majority of overlapping knowledge (Mintz et al., 2009).

3.3 Stage 3: Question Synthesis

Previous benchmarks generate QA pairs by directly feeding entire text segments to LLMs, making it difficult to ensure comprehensive coverage and quality control of the resulting questions. To address this limitation, we adopt a fine-grained, dual-input prompting strategy. Specifically, for each knowledge triple in $\mathcal{G}_{\text{test}}$, we leverage an LLM to automatically generate targeted test questions. Our dual-input prompting strategy equips LLMs with two complementary information sources: structured knowledge triples and their corresponding source text passages. This approach guides the model to generate fact-anchoring questions while maintaining fidelity to the original context. By anchoring question generation in both structured knowledge and source text, we ensure the generated questions accurately reflect the intended specific facts while preserving contextual relevance. By enumerating each edge in $\mathcal{G}_{\text{test}}$ and instructing the LLM to generate corresponding QA questions, we can guarantee at least a lower bound on the audit adequacy.

Our prompt design is based on several key principles. First, we explicitly define the LLM's role as an expert quiz question generator to set clear expectations. Second, by providing structured inputs consisting of both the knowledge triple and its original context, we ensure that the generated questions are firmly grounded in the relevant information. Third, we impose strict criteria on the generated questions: each must be answerable solely from the provided context, specific enough to yield a unique answer, and directly assess the semantic relationship between target entities. To facilitate automated evaluation, we require that each question-answer pair be output in a structured JSON format.

Furthermore, we adopt the one-shot learning by incorporating carefully selected example question-answer pairs into the prompt. These examples illustrate the desired question format and level of specificity, guiding the LLM toward generating high-quality, targeted questions. This comprehensive prompting strategy ensures that the synthesized questions effectively evaluate selective forgetting while maintaining human interpretability. The specific prompt employed in our experiments is provided in § A.1.

4 Experiments

4.1 Experimental Setup

Building upon MUSE, a comprehensive benchmark for LLM unlearning that provides extensive datasets and evaluation frameworks (Shi et al., 2024), we integrate HANKER to enhance its knowledge unlearning evaluation. For question generation, we leverage the DeepSeek-V3 model (Liu et al., 2024a), which has demonstrated superior performance recently. The MUSE framework incorporates two primary dataset-News and Books. For fairness and methodological rigor, we utilize MUSE's fine-tuned LLaMA2-7B model as our initial LLM, along with their default unlearning algorithm implementations and parameter configurations.

Unlearning Methods. We evaluate three representative unlearning methods from MUSE. Gradient Ascent (GA) inverts the training objective by maximizing loss on forgotten data to discourage memorized content generation. Negative Preference Optimization (NPO) treats forgotten knowledge as negative examples within a preference optimization framework. Task Vectors (TV) employs weight arithmetic by first training a model on forgotten content, deriving a memorization vector, then subtracting it from the original weights. Both GA and NPO can be enhanced with Gradient Descent on Retain set (GDR) or KL Divergence Regularization (KLR) for utility preservation.

Metrics. We evaluate the effectiveness of unlearning through our generated audit suite by quantifying the number of *knowledge memorization cases (KMCs)* in the unlearned model. While we maintain compatibility with existing approaches by using the same metrics as MUSE for overall assessment (i.e., ROUGE), we extend beyond aggregate similarity-based evaluation to identify specific failure instances. Our method applies software testing principles to pinpoint specific failure-revealing test cases—scenarios in which an LLM provider might be liable for disclosing sensitive information. The identification process employs two complementary criteria for judgment. The first criteria uses ROUGE Recall to measure surface-level similarity, requiring model outputs to exceed a strict threshold (Recall=1) compared to reference answers. The second metric leverages an entailment-based ap-

Table 1: Statistics of Knowledge Extraction and QA Dataset

Dataset	Initial Facts	Final Facts	QA Pairs	Average
News	24,763	16,912	69,609	4.11
Books	41,123	27,254	111,855	4.10

Table 2: Quality assessment of generated knowledge graphs and QA pairs based on the following metrics: Knowledge Fact Accuracy (AK), Question–Fact Relevance (QR), Question Clarity (QC), and Answer–Context Consistency (AC).

	AK	QR	QC	AC
News	0.76	0.91	0.99	0.91
Books	0.61	0.84	0.99	0.84

Table 3: Numbers of Knowledge Memorization Cases on News.

Method	MUSE		HANKER	
	ROUGE	Entail.	ROUGE	Entail.
w/o unlearn	33	19	4688	23605
GA_{KLR}	18	3	3702	21650
NPO_{GDR}	27	13	4454	23474
NPO_{KLR}	19	6	3780	21571
Task Vector	33	10	4853	23808

Table 4: Numbers of Knowledge Memorization Cases on Books.

Method	MUSE		HANKER	
	ROUGE	Entail.	ROUGE	Entail.
w/o unlearn	25	15	4729	38388
GA_{KLR}	6	7	3490	32365
NPO_{GDR}	0	34	1435	18094
NPO_{KLR}	4	8	3447	32332
Task Vector	25	15	4700	38210

proach (Yuan et al., 2024), utilizing a pre-trained NLI model as described in (Sileo, 2024) to verify semantic equivalence between generated and reference answers without logical inconsistencies. A higher frequency of detected memorization cases indicates less successful unlearning, while simultaneously demonstrating the comprehensiveness of our testing methodology.

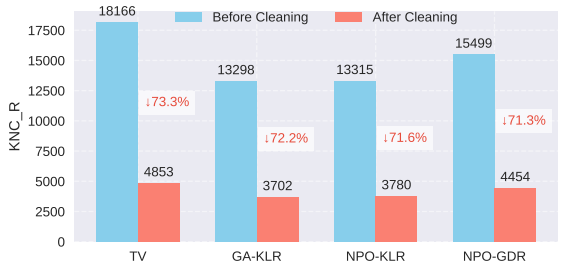
4.2 Details of Generated Audit Suite

We applied HANKER to two corpora provided by MUSE, namely the News and Books datasets. The details are summarized in Table 1, and the specific information about our constructed knowledge graphs can be found in § A.3. For the News dataset, HANKER extracted a knowledge graph (KG) from the forget dataset comprising 24,763 facts. After removing redundant knowledge, a final KG containing 16912 facts was obtained, from which 69,609 QA pairs were generated (On average, one fact corresponds to the generation of 4.11 QA pairs). Similarly, for the Books dataset, HANKER extracted a KG with 41,123 facts from the forget dataset. Following the elimination of redundant knowledge, a final KG comprising 27,254 facts was produced, and 111,855 QA pairs were generated from this KG (on average, one fact corresponds to the generation of 4.10 QA pairs). These results demonstrate the capability of HANKER to automatically extract fine-grained knowledge graphs and generate large-scale audit suites.

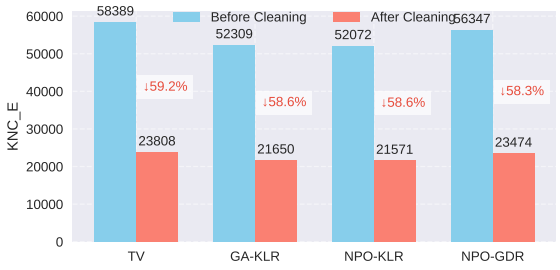
Mannual Assessment of the Generated Data. To rigorously assess the quality of HANKER's generated audit dataset, we conducted a detailed manual evaluation on randomly sampled 100 text chunks from each of the News and Books datasets. Our assessment focused on both the accuracy of extracted knowledge triples and the quality of generated QA pairs through four key metrics. Accuracy of Knowledge Fact (AK) measures the precision of knowledge triple extraction from the source text (whether entities and relations accurately represent the original text), achieving scores of 0.76 and 0.61 for News and Books respectively. The relatively lower score on Books reflects the inherent challenges in extracting structured knowledge from narrative text compared to more factual News articles. Question-Fact Relevance (QR) evaluates how well generated questions align with both the context and extracted facts. High scores of 0.91 (News) and 0.84 (Books) indicate that our framework effectively translates extracted knowledge into contextually appropriate questions. Question Clarity (QC) assesses the linguistic quality and specificity of generated questions. Near-perfect scores of 0.99 across both domains demonstrate our system's exceptional ability to generate clear, unambiguous, and well-formed questions regardless of source material complexity. Answer-Context Consistency (AC) gauges whether generated reference answers accurately reflect the source context. Strong performance of 0.91 (News) and 0.84 (Books) suggests reliable answer generation that maintains fidelity to the original text. These results demonstrate HANKER's capability in generating high-quality audit datasets.

4.3 Evaluation on Unlearning Methods

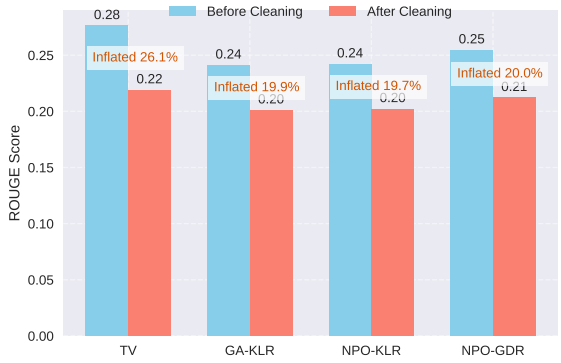
Our result reveals a striking disparity in the ability to detect knowledge memorization cases between HANKER's comprehensive audit suite and



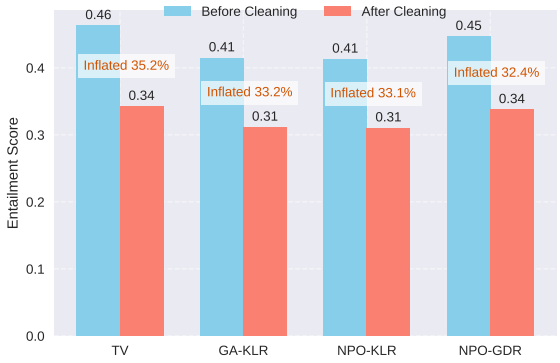
(a) Number of KMCs (by Rouge)



(b) Number of KMCs (by Entailment)



(c) ROUGE Score



(d) Entailment Score

Figure 4: Impact of Redundancy on Knowledge Memorization Cases.

MUSE's baseline approach. The results paint a concerning picture about the extent of retained knowledge in supposedly unlearned models that were previously undetectable with limited audit sets. On the News dataset, HANKER's detection capability proves remarkably more sensitive: using the ROUGE metric, it identifies over 4,600 memorization cases in the unmodified model, compared to just 33 cases detected by MUSE - a 142-fold increase in detection power. This gap widens even further when examining semantic understanding through the Entailment metric, where HANKER detects more than 23,600 cases versus MUSE's 19 cases, representing a dramatic 1,242-fold improvement in identifying retained knowledge. The Books dataset tells an equally compelling story. HANKER's comprehensive evaluation uncovers more than 4,700 memorization cases using ROUGE (compared to MUSE's 25 cases), and a remarkable 38,388 cases using Entailment (versus MUSE's 15 cases). These findings represent average improvements of 188× and 1,125× respectively in detection capability.

Particularly noteworthy is how these results persist across different unlearning methods. Even with state-of-the-art approaches like GA_{KLR} and NPO_{KLR}, HANKER consistently reveals significantly more cases where knowledge removal was incomplete. This suggests that current unlearning methods may be less effective than previously thought, with their apparent success potentially being an artifact of insufficient testing rather than genuine knowledge removal. These findings underscore the critical importance of comprehensive testing in evaluating unlearning effectiveness, revealing that the challenge of selective knowledge removal may be substantially more complex than indicated by previous benchmarks.

4.4 Impact of Knowledge Redundancy on Unlearning Effectiveness Audits

To validate the necessity of knowledge redundancy detection and elimination, we conducted a comprehensive experiment to assess its impact on unlearning evaluation effectiveness. Using the News dataset as our testbed, we compared evaluation outcomes between two scenarios: one using the full dataset (126,224 test cases) and another using our deduplicated dataset (69,609 test cases). Our analysis considered both the number of identified knowledge memorization cases and standard dataset-level metrics (ROUGE and Entailment scores) used in existing evaluations. The results reveal a strik-

ing impact of knowledge redundancy on evaluation outcomes. When using our deduplicated audit set, the number of identified knowledge memorization cases decreased substantially: detection rates dropped by 71.3-73.3% under the ROUGE criterion and by 58.3-59.2% under the Entailment criterion. This significant reduction suggests that knowledge redundancy leads to substantial false positives, where retained knowledge is incorrectly flagged as forgetting failures. Furthermore, our analysis of quantitative metrics demonstrates that knowledge redundancy artificially inflates unlearning effectiveness measures. Without deduplication, ROUGE scores showed artificial inflation ranging from 19.7% to 26.1%, while Entailment scores were inflated by 32.4% to 35.2%. These inflated metrics indicate that traditional evaluation approaches may significantly overestimate unlearning effectiveness when redundant knowledge is not properly controlled for.

These findings provide compelling evidence for the critical importance of knowledge redundancy elimination in unlearning evaluation. The substantial reductions in false positives and metric inflation demonstrate that rigorous knowledge deduplication is essential for accurate assessment of unlearning effectiveness. Extended observations on Llama3-8B in § A.5 further corroborate these insights.

5 Related Work and Discussion

Machine Unlearning for LLMs. Machine unlearning has progressively evolved toward applications in large language models from classification tasks. Contemporary research predominantly explores parameter optimization methodologies, achieved through targeted fine-tuning procedures (Yao et al., 2023; Jang et al., 2022; Wang et al., 2024c; Yao et al., 2024; Tian et al., 2024; Liu et al., 2024d; Gu et al., 2024; Jia et al., 2024a) The transparent nature of modifying neural architectures engenders enhanced user trust, despite potential compromises to overall model performance. Beyond parameter-based approaches, researchers have pioneered diverse methodologies including advanced contrastive decoding frameworks (Eldan and Russinovich, 2023; Wang et al., 2024a; Ji et al., 2024; Huang et al., 2024), task-specific vector implementations (Liu et al., 2024e; Dou et al., 2025), contextual learning strategies (Pawelczyk et al., 2024; Muresanu et al., 2024), and sophisticated input processing mechanisms (Gao et al., 2024; Liu et al., 2024b).

Evaluation of LLM Unlearning. The evaluation of LLM unlearning effectiveness encompasses diverse task scenarios. Early research focused on traditional NLP classification tasks to examine models' prediction (Chen and Yang, 2023). Subsequently, researchers developed specialized datasets to provide standardized evaluation platforms (Eldan and Russinovich, 2023; Shi et al., 2024; Maini et al., 2024). Besides, some work has been devoted to focusing on the robustness of unlearning, i.e., adding perturbations to the same problem to activate model memory (Joshi et al., 2024).

Knowledge Graphs for Evaluation. Knowledge graphs offer distinct advantages beyond the completeness and identifiability properties utilized in this study. They serve as effective tools for evaluating both QA systems (Wang et al., 2024b) and LLM unlearning (Wu et al., 2024). Notably, knowledge graphs enable the assessment of multi-hop reasoning through transitive relationships (if a→b and b→c, then testing whether the model infers a→c). The framework we propose in this paper conveniently integrates with these techniques.

Discussion. While unlearning evaluation encompasses multiple dimensions, our work focuses specifically on unlearned knowledge audit data generation. We utilize default MUSE configurations rather than optimizing each unlearning algorithm, as our primary contribution is the development of a robust audit framework rather than establishing state-of-the-art unlearning performance benchmarks. Although HANKER could be readily extended to evaluate normal utility based the retain KG, this extension falls outside our current scope and represents our future work.

6 Conclusion

In this paper, we introduce HANKER, an automated framework for generating holistic audit datasets to evaluate the effectiveness of LLM unlearning. By leveraging knowledge graphs, HANKER addresses two critical challenges: ensuring audit adequacy and eliminating knowledge redundancy between forget and retain datasets. Our empirical analysis on the MUSE benchmark demonstrates that HANKER significantly expands audit coverage, identifying thousands of previously undetected knowledge memorization cases and revealing how knowledge redundancy substantially skews unlearning effectiveness metrics.

Limitations and Ethical Considerations

Limitations. The primary limitation of our work is that it extends only the dataset provided by MUSE and employs DeepSeek-V3 for question generation. Additionally, we acknowledge that the current knowledge extraction methods may impact overall effectiveness. To mitigate these limitations, we have released our code and the generated audit suite, allowing researchers to utilize our framework with their preferred extraction models. Our framework is designed to be modular, enabling future improvements through integration of more advanced extraction techniques. Meanwhile, extending our framework to other benchmarks remains an important direction for our future work.

Ethical Considerations. Machine unlearning can be employed to mitigate risks associated with LLMs in terms of privacy, security, bias, and copyright. Our work is dedicated to providing a comprehensive evaluation framework to help researchers better understand the unlearning effectiveness of LLMs, which we believe will have a positive impact on society.

References

AI@Meta. 2024. Llama 3 model card.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.

Jiaao Chen and Diyi Yang. 2023. Unlearn what you want to forget: Efficient unlearning for llms. *Preprint*, arXiv:2310.20150.

Quang-Vinh Dang. 2021. Right to be forgotten in the age of machine learning. In *Advances in Digital Science: ICADS 2021*, pages 403–411. Springer.

Guangyao Dou, Zheyuan Liu, Qing Lyu, Kaize Ding, and Eric Wong. 2025. Avoiding copyright infringement via large language model unlearning. *Preprint*, arXiv:2406.10952.

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

Chongyang Gao, Lixu Wang, Chenkai Weng, Xiao Wang, and Qi Zhu. 2024. Practical unlearning for large language models. *Preprint*, arXiv:2407.10223.

Kang Gu, Md Rafi Ur Rashid, Najrin Sultana, and Shagufta Mehnaz. 2024. Second-order information matters: Revisiting machine unlearning for large language models. *arXiv preprint arXiv:2403.10557*.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, et al. 2021. Knowledge graphs. *ACM Computing Surveys (Csur)*, 54(4):1–37.

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.

James Y. Huang, Wenxuan Zhou, Fei Wang, Fred Morstatter, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2024. Offset unlearning for large language models. *Preprint*, arXiv:2404.11045.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. REBEL: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 2022. Knowledge unlearning for mitigating privacy risks in language models. *arXiv preprint arXiv:2210.01504*.

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. *arXiv preprint arXiv:2406.08607*.

Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and S Yu Philip. 2021. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE transactions on neural networks and learning systems*, 33(2):494–514.

Jinghan Jia, Yihua Zhang, Yimeng Zhang, Jiancheng Liu, Bharat Runwal, James Diffenderfer, Bhavya Kailkhura, and Sijia Liu. 2024a. Soul: Unlocking the power of second-order optimization for llm unlearning. *arXiv preprint arXiv:2404.18239*.

Jinghan Jia, Yihua Zhang, Yimeng Zhang, Jiancheng Liu, Bharat Runwal, James Diffenderfer, Bhavya Kailkhura, and Sijia Liu. 2024b. 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.

Abhinav Joshi, Shaswati Saha, Divyaksh Shukla, Sriram Vema, Harsh Jhamtani, Manas Gaur, and Ashutosh Modi. 2024. Towards robust evaluation of unlearning in LLMs via data transformations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12100–12119, Miami, Florida, USA. Association for Computational Linguistics.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. *arXiv preprint arXiv:1707.07045*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Chris Yuhao Liu, Yaxuan Wang, Jeffrey Flanigan, and Yang Liu. 2024b. Large language model unlearning via embedding-corrupted prompts. *Preprint*, arXiv:2406.07933.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. 2023. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. *Preprint*, arXiv:2305.01210.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, et al. 2024c. Rethinking machine unlearning for large language models. *arXiv preprint arXiv:2402.08787*.

Zhenhua Liu, Tong Zhu, Chuanyuan Tan, and Wenliang Chen. 2024d. Learning to refuse: Towards mitigating privacy risks in llms. *arXiv preprint arXiv:2407.10058*.

Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. 2024e. Towards safer large language models through machine unlearning. *Preprint*, arXiv:2402.10058.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C Lipton, and J Zico Kolter. 2024. Tofu: A task of fictitious unlearning for llms. *arXiv preprint arXiv:2401.06121*.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011.

Andrei Muresanu, Anvith Thudi, Michael R. Zhang, and Nicolas Papernot. 2024. Unlearnable algorithms for in-context learning. *Preprint*, arXiv:2402.00751.

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. 2024. In-context unlearning: Language models as few shot unlearners. *Preprint*, arXiv:2310.07579.

Ankit Satpute, Noah Gießing, André Greiner-Petter, Moritz Schubotz, Olaf Teschke, Akiko Aizawa, and Bela Gipp. 2024. Can llms master math? investigating large language models on math stack exchange. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 2316–2320.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. 2024. Muse: Machine unlearning six-way evaluation for language models.

Damien Sileo. 2024. tasksource: A large collection of nlp tasks with a structured dataset preprocessing framework. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15655–15684.

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. *arXiv preprint arXiv:2407.01920*.

Bichen Wang, Yuzhe Zi, Yixin Sun, Yanyan Zhao, and Bing Qin. 2024a. Rkld: Reverse kl-divergence-based knowledge distillation for unlearning personal information in large language models. *arXiv preprint arXiv:2406.01983*.

Jun Wang, Yanhui Li, Zhifei Chen, Lin Chen, Xiaofang Zhang, and Yuming Zhou. 2024b. Knowledge graph driven inference testing for question answering software. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, ICSE '24, New York, NY, USA. Association for Computing Machinery.

Lingzhi Wang, Xingshan Zeng, Jinsong Guo, Kam-Fai Wong, and Georg Gottlob. 2024c. Selective forgetting: Advancing machine unlearning techniques and evaluation in language models. *arXiv preprint arXiv:2402.05813*.

Ruihan Wu, Chhavi Yadav, Russ Salakhutdinov, and Kamalika Chaudhuri. 2024. Evaluating deep unlearning in large language models. *Preprint*, arXiv:2410.15153.

Jin Yao, Eli Chien, Minxin Du, Xinyao Niu, Tianhao Wang, Zezhou Cheng, and Xiang Yue. 2024. Machine unlearning of pre-trained large language models. *arXiv preprint arXiv:2402.15159*.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2023. Large language model unlearning. *arXiv preprint arXiv:2310.10683*.

Xiaojian Yuan, Tianyu Pang, Chao Du, Kejiang Chen, Weiming Zhang, and Min Lin. 2024. A closer look at machine unlearning for large language models. *arXiv preprint arXiv:2410.08109*.

A Appendix

A.1 Prompts for Question Synthesis

Below, we present the specific prompts used with DeepSeek-V3 for generating audit questions, as discussed in § 3.3.

```
SYS_PROMPT = """You are an expert quiz generator. Given a text passage and a relationship triple, generate specific questions to test knowledge about this relationship based on the context provided.

Input Format:
- Text: A passage containing information about the relationship
- Relationship: A triple containing {'head': entity1, 'type': relation_type, 'tail': entity2}

Task:
Generate up to 5 focused questions that test understanding of the relationship between the head entity and tail entity, considering:
1. Questions should be answerable solely from the given context
2. Questions should be specific enough to have a unique correct answer
3. Questions can ask about the tail entity given the head entity and relationship type
4. Questions can ask about the relationship between the two entities
5. Questions can ask about specific details that establish this relationship

Requirements:
1. Each question must have a clear, unambiguous answer based on the context
2. Avoid overly broad or general questions
3. Focus on the specific relationship provided
4. Use the context to add specific details to questions
5. Ensure questions and answers are factually consistent with the provided text

Response Format:
The response must be a valid JSON object with the following structure:
{
    "1": {
        "question": "Your question text here",
        "reference_answer": "The correct answer based on context"
    },
    "2": {
        "question": "...",
        "reference_answer": "..."
    }
    // ... up to 5 questions
}

Example Input:
Text: "The Greek Orthodox Church observes Lent as a period of fasting and spiritual reflection that begins on Clean Monday and lasts for 40 days. During this time, adherents follow strict dietary restrictions and increase their prayer and attendance at special services."
Relationship: {'head': 'Lent', 'type': 'religion', 'tail': 'Greek Orthodox'}

Example Output:
{
    "1": {
        "question": "Which religious denomination observes Lent beginning on Clean Monday with a 40-day period of fasting and spiritual reflection?",
        "reference_answer": "Greek Orthodox"
    },
    "2": {
        "question": "In the Greek Orthodox tradition, what is the length of the Lent period?",
        "reference_answer": "40 days"
    }
}
"""

USER_PROMPT = """
Please generate questions based on the following input:

Text: {text}
Relationship: {relationship}
"""
```

Figure 5: Our prompt.

A.2 Computational Resource Requirements

Implementing HANKER at scale requires moderate computational resources, making it accessible for most research environments. The resource requirements can be divided into two main components: Step 1 (Knowledge Graph Construction) and Step 3 (Question Synthesis). The entity and relation extraction model used in Step 1 can be executed on consumer-grade hardware. Our implementation utilizes the REBEL-large model, which can operate efficiently on a single 12GB GPU. This is within the specifications of widely available personal computing setups, making the knowledge graph construction phase accessible without specialized infrastructure. For generating high-quality audit questions in Step 3, we leverage the DeepSeek-V3 model, which offers state-of-the-art performance in targeted question generation. While this process could potentially be resource-intensive, we utilized API access rather than local deployment. The cost efficiency of this model is notable—our complete implementation, including processing the entire MUSE benchmark corpus (generating over 180,000 audit questions), incurred API expenses of less than $20.

A.3 Knowledge Graph Statistics

In this section, we provide detailed statistics about the knowledge graphs generated during Step 1 (Knowledge Graph Construction) and Step 2 (Redundancy Removal). Table 5 shows the number of nodes, edges, and average node degree for both the News and Books datasets, before and after redundancy removal. It is worth noting that our knowledge extraction process is capable of identifying multiple triples from a single text passage. For the Books dataset, each text passage yields an average of 2.11 relation triples, reflecting the rich informational content of narrative text. In comparison, the News dataset yields an average of 1.74 triples per text passage, which aligns with the more concise nature of news articles. This extraction density demonstrates the effectiveness of our approach in capturing fine-grained knowledge units from unstructured text, enabling more comprehensive coverage in the audit process. The reduction in edges after redundancy removal (33.7% for Books and 31.7% for News) highlights the significant overlap between the forget and retain datasets, underscoring the importance of our redundancy removal step for accurate unlearning audit.

Table 5: Statistics of the constructed knowledge graphs before and after redundancy removal.

Dataset	Nodes	Edges	Degree
Books (w/o removal)	21,523	41,123	3.8213
Books (w removal)	21,474	27,254	2.5383
News (w/o removal)	21,058	24,763	2.3519
News (w removal)	20,079	16,912	1.6845

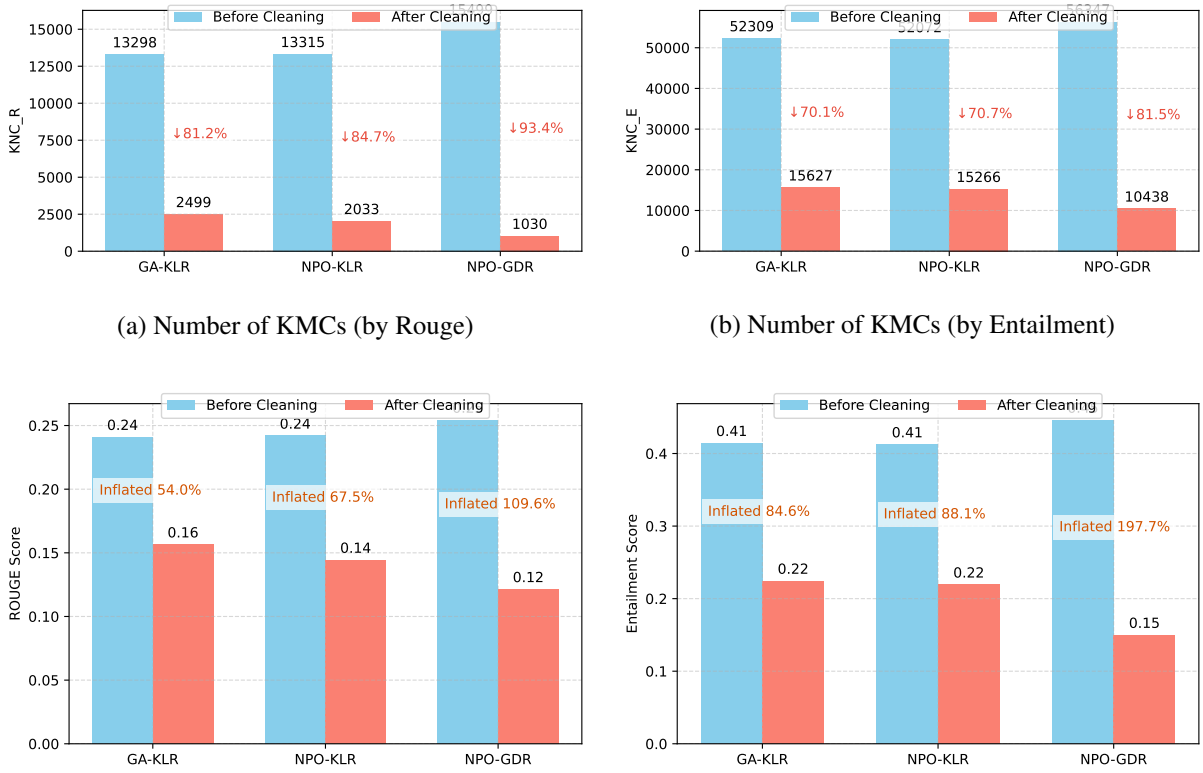
Table 6: Coverage analysis of MUSE QA pairs against knowledge graph edges.

KG	Total / Covered Edges	Coverage
Books (w/o removal)	41,123 / 2,922	7.11%
Books (w removal)	27,254 / 473	1.74%
News (w/o removal)	24,763 / 193	0.78%
News (w removal)	16,912 / 102	0.60%

A.4 Coverage Analysis of MUSE

A critical question in evaluating MUSE is whether existing test sets adequately represent the knowledge in the forget corpus. To rigorously assess this beyond just comparing the number of QA pairs, we conducted a detailed coverage analysis of MUSE's audit dataset against our extracted knowledge graph. We defined coverage as the percentage of knowledge graph edges where both endpoints (entities) match entity pairs extracted from MUSE QA questions. While this approach provides a reasonable approximation that evaluates representation at a semantic level rather than merely counting instances, it likely overestimates the actual coverage. This is because matching entity endpoints does not guarantee that the specific relationship between entities is correctly represented in the QA pair. Therefore, the true semantic coverage is likely even lower than our reported estimates. The results, presented in Table 6, reveal significant limitations in existing MUSE.

Our analysis reveals two critical insights: ❶ MUSE's current dataset coverage is extremely limited, representing only 7.11% of knowledge edges in the Books dataset and a mere 0.78% in the News dataset, highlighting the insufficient evaluation scope of existing benchmarks. ❷ More concerning is the significant drop in covered edges after redundancy removal—from 2,922 to just 473 edges (83.8% reduction) in the Books dataset and from 193 to 102 edges (47.2% reduction) in the News dataset. This dramatic reduction demonstrates that a substantial portion of MUSE's original test questions are actually evaluating knowledge



(a) Number of KMCs (by Rouge)

(b) Number of KMCs (by Entailment)

(c) ROUGE Score

(d) Entailment Score

Figure 6: Impact of Redundancy on Knowledge Memorization Cases on LLama3-8B.

that should be retained rather than forgotten, which could lead to misleading conclusions about unlearning effectiveness. These findings provide quantitative evidence supporting our observation in Figure 1, where we illustrated how knowledge targeted for forgetting also appears in the retain dataset. The substantial drop in coverage after redundancy removal confirms that existing benchmarks not only provide insufficient coverage but also contain a significant proportion of potentially misleading test cases that evaluate knowledge preservation rather than forgetting.

A.5 Observation on Llama3

To address the potential limitations of results derived from a single model, we extended our evaluation to the more recent Llama3-8B (AI@Meta, 2024) in addition to LLaMA2-7B. Figure 6 illustrates HANKER results on Llama3-8B, which demonstrate remarkable consistency with our LLaMA2 observations. When using our deduplicated audit set, the number of identified knowledge memorization cases decreased substantially: detection rates dropped by 70.1-81.5% under the Entailment criterion and by 81.2-93.4% under the ROUGE criterion, demonstrating our framework's ability to precisely identify retained knowledge. This significant reduction suggests that knowledge redundancy leads to substantial false positives, where retained knowledge is incorrectly flagged as forgetting failures. Furthermore, our analysis of quantitative metrics demonstrates that knowledge redundancy artificially inflates unlearning effectiveness measures. Without deduplication, ROUGE scores showed artificial inflation ranging from 54.0% to 109.6%, while Entailment scores were inflated by 84.6% to 197.7%. These inflated metrics indicate that traditional evaluation approaches may significantly overestimate unlearning effectiveness when redundant knowledge is not properly controlled for. The consistency of these patterns across different model architectures highlights that knowledge redundancy constitutes a fundamental challenge in unlearning evaluation rather than a model-specific phenomenon.