From Chaos to Clarity: A Knowledge Graph-Driven Audit Dataset Generation Framework for LLM Unlearning

Weipeng Jiang[1], Juan Zhai[2], Shiqing Ma[2], Ziyan Lei[3], Xiaofei Xie[4], Yige Wang[1], Chao Shen[1]*

[1]Xi'an Jiaotong University
[2]University of Massachusetts Amherst
[3]University of Bristol
[4]Singapore Management University
{lenijwp, jihejue039}@stu.xjtu.edu.cn, chaoshen@mail.xjtu.edu.cn,
{juanzhai, shiqingma}@umass.edu, dq25474@bristol.ac.uk, xfxie@smu.edu.sg

Abstract

Recently LLMs have faced increasing demands to selectively remove specific information through Machine Unlearning. While evaluating unlearning effectiveness is crucial, existing benchmarks suffer from fundamental limitations in audit dataset generation from unstructured corpora. We identify two critical challenges: **ensuring audit adequacy** and **handling knowledge redundancy** between forget and retain datasets. Current approaches rely on ad-hoc question generation from unstructured text, leading to unpredictable coverage gaps and evaluation blind spots. Knowledge redundancy between forget and retain corpora further obscures evaluation, making it difficult to distinguish genuine unlearning failures from legitimately retained knowledge. To bring clarity to this challenge, we propose **LUCID**, an automated framework that leverages knowledge graphs to achieve comprehensive audit dataset generation with fine-grained coverage and systematic redundancy elimination. By converting unstructured corpora into structured knowledge representations, it transforms the ad-hoc audit dataset generation process into a transparent and automated generation pipeline that ensures both adequacy and non-redundancy. Applying LUCID to the MUSE benchmark, we generated over 69,000 and 111,000 audit cases for News and Books datasets respectively, identifying thousands of previously undetected knowledge memorization instances. Our analysis reveals that knowledge redundancy significantly skews metrics, artificially inflating ROUGE from 19.7% to 26.1% and Entailment Scores from 32.4% to 35.2%, highlighting the necessity of deduplication for accurate assessment.

Repository — https://github.com/lenijwp/LUCID

1 Introduction

In recent years, Large Language Models (LLMs) have demonstrated remarkable capabilities across diverse applications (Liu et al. 2023; Satpute et al. 2024), while simultaneously raising concerns about their retention of sensitive information, including personally identifiable information (PII)(Jang et al. 2022), unsafe behaviors(Liu et al. 2024d),

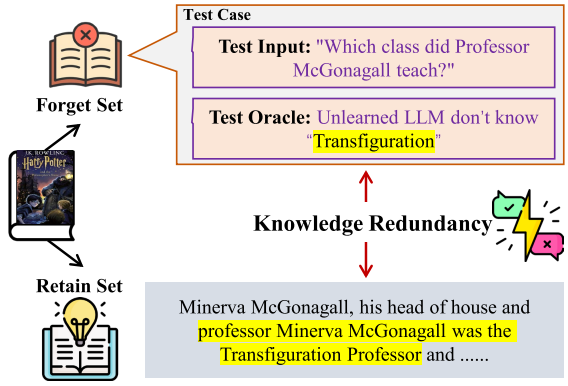


Figure 1: An illustrative example from MUSE demonstrating where knowledge targeted for forgetting also appears in the Retain Dataset, highlighting the challenge of knowledge redundancy in unlearning evaluation.

and 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, Van Der Sloot, and Borgesius 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 pre-trained LLMs while preserving their general language modeling capabilities, thereby avoiding the substantial computational costs associated with building new models from scratch.

The growing significance of LLM unlearning has highlighted the importance of rigorous evaluation frameworks for assessing unlearning performance. Recent benchmarks like MUSE (Shi et al. 2024) and TOFU (Maini et al. 2024) have advanced the field by establishing standardized datasets, providing pre-trained target models, and introducing multifaceted evaluation metrics that assess unlearning efficacy across multiple dimensions, from verbatim text retention to embedded knowledge preservation. However, these pioneering frameworks fundamentally rely on ad-hoc automated question generation from unstructured corpora using LLMs,

*Corresponding author.

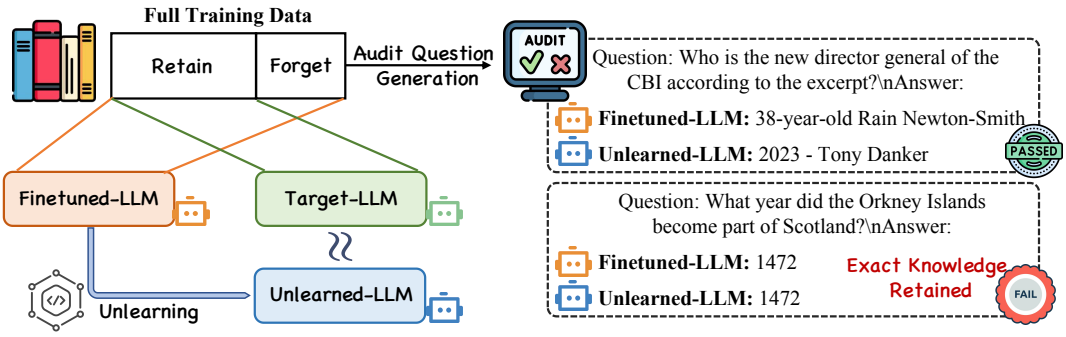


Figure 2: Illustration of the basic pipeline for LLM knowledge unlearning and its audit.

creating an inherently unpredictable and opaque audit dataset construction process. This unstructured approach to evaluation raises significant concerns about whether the resulting audit suites can reliably and comprehensively assess targeted knowledge removal, potentially undermining confidence in unlearning effectiveness claims.

Our systematic investigation reveals that these reliability concerns stem from two fundamental challenges in audit dataset generation from unstructured corpora. First, the ***audit adequacy***: while automated QA generation can produce multiple question-answer pairs for each target text, it introduces significant uncertainty about comprehensive coverage of critical knowledge within the source material. Our analysis demonstrates the severity of this issue—MUSE employs only 100 test questions to evaluate 0.8M corpora, capturing merely 0.8% of knowledge in the News dataset and 7.1% in the Books dataset, as shown by our analysis in Section 4.5. Second, the ***knowledge redundancy***: current evaluation methods fail to account for overlapping knowledge between forget and retain corpora. As shown in Figure 2, shared knowledge should be preserved during exact unlearning. However, existing frameworks ignore this overlap (Figure 1), leading to misleading evaluations that penalize models for legitimately retained knowledge.

In this paper, we propose LUCID, 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. LUCID 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, LUCID utilizes specific facts to guide LLMs in generating high-quality, targeted test questions, guaranteeing comprehensive and accurate auditing. Through this pipeline, LUCID 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 LUCID, 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 LUCID 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, Xu, and Liu 2023). With the proliferation of LLMs, unlearning has gained significant attention due to its broad applications 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. (I) **Audit Adequacy**: Existing benchmarks rely on LLM prior knowledge to generate QA pairs or segment corpora for ChatGPT-based automated QA generation. However, this approach suffers from significant coverage gaps. Taking MUSE as an example, our analysis in Section 4.5 demonstrates that its evaluation datasets capture only 7.1% of knowledge in Books sub-task and a mere 0.8% in News sub-task, indicating substantial inadequacy in current audit scope. Such limited coverage fails to provide a comprehensive knowledge assessment necessary for reliable unlearning evaluation. (II) **Knowledge Redundancy**: A more subtle and easily overlooked issue is that the Retain Dataset and Forget Dataset may contain overlapping knowledge. As illustrated in Table 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 Table 1.

Knowledge graphs(KGs) provide a promising solution to these challenges. First, KGs inherently capture 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, we can achieve more comprehensive and systematic auditing. Second, the structured nature of KGs facilitates the identification of identical knowledge facts present in both Retain and Forget Datasets. This capability enables refinement of the initial forget knowledge graph by removing potentially retained information. Finally, recent advances in KG extraction technology provide numerous automated models and pipelines that support scalable audit dataset construction.

3 Proposed Method

The core idea behind LUCID 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 Fig 3, LUCID comprises three sequential stages. During the (1) **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 (2) **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 (3) **Question Synthesis** stage, LUCID 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.

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:

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

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

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

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

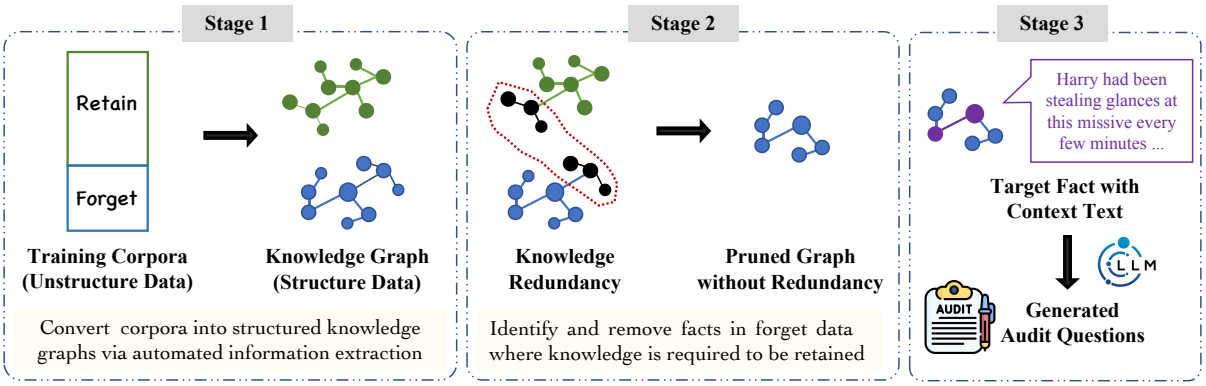


Figure 3: Overview of the proposed LUCID. The framework consists of three stages: (1) **Knowledge Graph Construction** that extracts structured knowledge from forget and retain data, (2) **Redundancy Removal** that identifies and removes redundant knowledge from the constructed knowledge graphs, and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

Algorithm 1: LUCID

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

function GENERATION(D_{fgt}, D_{ret})
  ▷ Knowledge Graph Construction
  $G_{\text{fgt}} \leftarrow \text{KGExtraction}(D_{\text{fgt}})$
  $G_{\text{ret}} \leftarrow \text{KGExtraction}(D_{\text{ret}})$
  ▷ 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**
    $ctx \leftarrow \text{RetrieveContext}(e)$
    $prompt \leftarrow \text{ComposePrompt}(e, ctx)$
    $qa \leftarrow \text{LLM}(prompt)$
    $S \leftarrow S \cup \{qa\}$
  return S

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 can be found in our repository.

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 LUCID to enhance its knowledge unlearning evaluation. For question generation, we leverage the DeepSeek-V3 model (Liu et al. 2024a), which has demonstrated superior performance re-

cently. 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 approach (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 LUCID to two MUSE corpora: the News and Books datasets. Dataset details are presented in Table 2, with knowledge graph statistics shown in Table 1. During **Knowledge Graph Construction**, LUCID successfully extracted 24,763 facts from the News forget dataset and 41,123 facts from the Books forget dataset. Our extraction process effectively captures multiple knowledge triples from individual passages, averaging 1.74 triples per News passage and 2.11 triples per Books passage. This discrepancy reflects the inherent characteristics of these text types: narrative texts contain richer relational information compared to concise news articles. With **Redundancy Removal**, the News KG was reduced to 16,912 facts (a 31.7% reduction), while the Books KG was reduced to 27,254 facts (a 33.7% reduction). This substantial overlap highlights a critical limitation in existing benchmarks—without proper redundancy handling, audit results may erroneously penalize models for retaining knowledge they should legitimately preserve. Finally, by **Question Synthesis**, based on the refined KGs, LUCID generated 69,609 QA pairs for News (4.11 per fact) and 111,855 QA pairs for Books (4.10 per fact), demonstrating the scalability of our approach with comprehensive coverage.

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 1: Knowledge Graph Statistics Before and After Redundancy Removal

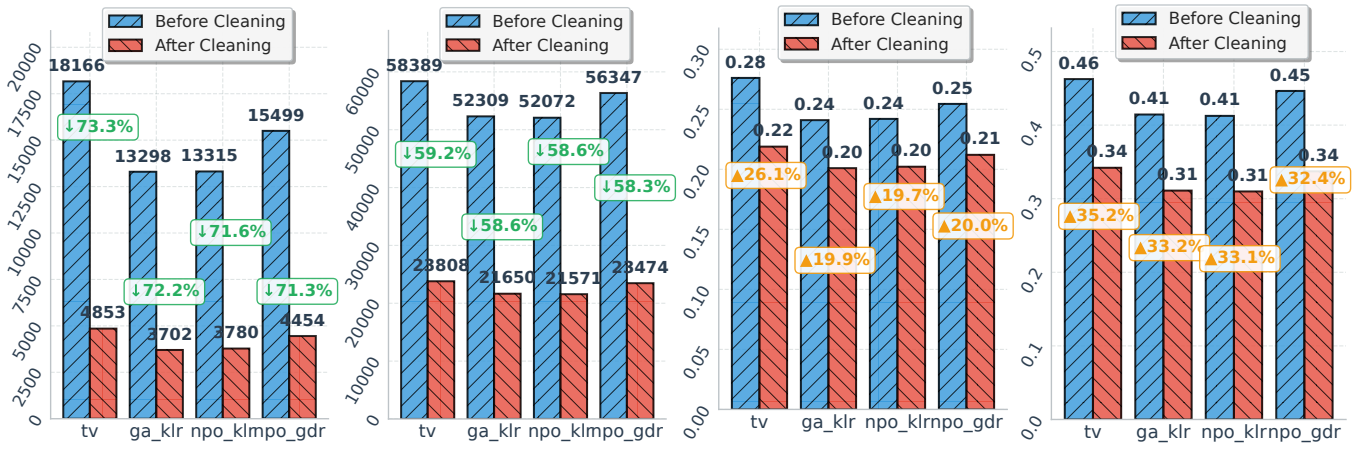
Dataset	Initial Facts	Final Facts	QA Pairs	Average
News	24,763	16,912	69,609	4.11
Books	41,123	27,254	111,855	4.10

Table 2: Statistics of Knowledge Extraction and QA Dataset

Manual Assessment of the Generated Data. To rigorously assess the quality of LUCID'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 context and extracted facts. High scores of 0.91 (News) and 0.84 (Books) indicate effective translation of 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 exceptional ability to generate clear, unambiguous, well-formed questions regardless of source complexity. Answer-Context Consistency (AC) measures whether generated answers accurately reflect the source context. Strong performance of 0.91 (News) and 0.84 (Books) suggests reliable answer generation with fidelity to the original text. These results demonstrate LUCID'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 LUCID's comprehensive audit suite and MUSE's baseline approach, as shown in Table 4 and Table 5. 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, LUCID's detec-



(a) KMCs by ROUGE (b) KMCs by Entailment (c) ROUGE Score (d) Entailment Score

Figure 4: Impact of Redundancy on Knowledge Memorization Cases Detection and Evaluation.

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

Table 3: 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).

Method	MUSE		LUCID	
	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: Number of KMCs on News.

Method	MUSE		LUCID	
	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

Table 5: Number of KMCs on Books.

tion 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 LUCID 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. LUCID'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}, LUCID 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 striking 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

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%

Table 6: Coverage of MUSE Audit Questions.

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.

These findings provide evidences 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 further validate these insights, confirming that redundancy interference in forgetting evaluation is a generalizable phenomenon. Details are provided in the appendix of our repository.

4.5 Coverage Analysis of MUSE

A critical question in MUSE is whether existing test sets adequately represent the knowledge in the forget corpus. To rigorously assess this beyond just the number of QA pairs, we conducted a detailed coverage analysis of MUSE's audit dataset on our extracted knowledge graph. We define coverage as the percentage of knowledge graph edges whose endpoint entities match those in MUSE QA questions. While approximate, as matched entity pairs may not imply captured relations, this metric offers a reasonable semantic-level assessment. Notably, even under this lenient measure, Table 6 reveals substantial coverage gaps in MUSE. Since our metric likely overestimates true coverage, the actual semantic limitations of the benchmark are presumably more severe than reported.

Our analysis reveals two critical insights: (1) 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. (2) 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 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 misleading test cases that evaluate knowledge preservation rather than forgetting.

5 Related Work and Discussion

Machine Unlearning for LLMs. Machine unlearning has evolved from classification tasks to applications in large language models. Contemporary research predominantly explores parameter optimization through targeted fine-tuning (Yao, Xu, and Liu 2023; Jang et al. 2022; Wang et al. 2024c; Yao et al. 2024; Tian et al. 2024). The transparency of modifying neural architectures enhances user trust, despite potential performance trade-offs. Beyond parameter-based approaches, researchers have developed diverse methodologies including advanced contrastive decoding frameworks (Eldan and Russinovich 2023; Wang et al. 2024a; Ji et al. 2024), task-specific vector implementations (Liu et al. 2024d; Dou et al. 2025), contextual learning strategies (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$\rightarrow$b and b$\rightarrow$c, then testing whether the model infers a$\rightarrow$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 knowledge memorization. We utilize default MUSE configurations rather than optimizing each algorithm, as our primary contribution is developing the audit dataset rather than benchmarking unlearning methods. Extending LUCID to evaluate utility on the retain KG is our future work.

6 Conclusion

This paper introduces LUCID, an automated framework for generating audit datasets to evaluate LLM unlearning. Leveraging knowledge graphs, LUCID addresses two critical challenges: ensuring audit adequacy and eliminating knowledge redundancy between forget and retain datasets. Our empirical analysis on MUSE demonstrates that LUCID significantly expands audit coverage and reveals how knowledge redundancy substantially skews unlearning effectiveness metrics.

Acknowledgments

This research is supported by the National Key Research and Development Program of China (2023YFB3107400), the National Natural Science Foundation of China (62521002, U24B20185, T2442014, 62161160337, 62132011, U21B2018), the Shaanxi Province Key Industry Innovation Program (2023-ZDLGY-38, 2021ZDLGY01-02). It is also supported by the National Research Foundation, Singapore, and the Cyber Security Agency under its National Cybersecurity R&D Programme (NCRP25-P04-TAICeN). Thanks to the New Cornerstone Science Foundation and the Xplorer Prize. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the funding agencies.

References

Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; and Yakhnenko, O. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.

Chen, J.; and Yang, D. 2023. Unlearn What You Want to Forget: Efficient Unlearning for LLMs. arXiv:2310.20150.

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

Dou, G.; Liu, Z.; Lyu, Q.; Ding, K.; and Wong, E. 2025. Avoiding Copyright Infringement via Large Language Model Unlearning. arXiv:2406.10952.

Eldan, R.; and Russinovich, M. 2023. Who's Harry Potter? Approximate Unlearning in LLMs. *arXiv preprint arXiv:2310.02238*.

Gao, C.; Wang, L.; Weng, C.; Wang, X.; and Zhu, Q. 2024. Practical Unlearning for Large Language Models. arXiv:2407.10223.

Hogan, A.; Blomqvist, E.; Cochez, M.; d'Amato, C.; Melo, G. D.; Gutierrez, C.; Kirrane, S.; Gayo, J. E. L.; Navigli, R.; Neumaier, S.; et al. 2021. Knowledge graphs. *ACM Computing Surveys (Csur)*, 54(4): 1–37.

Hoofnagle, C. J.; Van Der Sloot, B.; and Borgesius, F. Z. 2019. The European Union general data protection regulation: what it is and what it means. *Information & Communications Technology Law*, 28(1): 65–98.

Huguet Cabot, P.-L.; and Navigli, R. 2021. REBEL: Relation Extraction By End-to-end Language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2370–2381. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Jang, J.; Yoon, D.; Yang, S.; Cha, S.; Lee, M.; Logeswaran, L.; and Seo, M. 2022. Knowledge unlearning for mitigating privacy risks in language models. *arXiv preprint arXiv:2210.01504*.

Ji, J.; Liu, Y.; Zhang, Y.; Liu, G.; Kompella, R. R.; Liu, S.; and Chang, S. 2024. Reversing the Forget-Retain Objectives: An Efficient LLM Unlearning Framework from Logit Difference. *arXiv preprint arXiv:2406.08607*.

Ji, S.; Pan, S.; Cambria, E.; Marttinen, P.; and Philip, S. Y. 2021. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE transactions on neural networks and learning systems*, 33(2): 494–514.

Jia, J.; Zhang, Y.; Zhang, Y.; Liu, J.; Runwal, B.; Diffenderfer, J.; Kailkhura, B.; and Liu, S. 2024a. Soul: Unlocking the power of second-order optimization for llm unlearning. *arXiv preprint arXiv:2404.18239*.

Jia, J.; Zhang, Y.; Zhang, Y.; Liu, J.; Runwal, B.; Diffenderfer, J.; Kailkhura, B.; and Liu, S. 2024b. SOUL: Unlocking the Power of Second-Order Optimization for LLM Unlearning. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 4276–4292. Miami, Florida, USA: Association for Computational Linguistics.

Joshi, A.; Saha, S.; Shukla, D.; Vema, S.; Jhamtani, H.; Gaur, M.; and Modi, A. 2024. Towards Robust Evaluation of Unlearning in LLMs via Data Transformations. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 12100–12119. Miami, Florida, USA: Association for Computational Linguistics.

Lee, K.; He, L.; Lewis, M.; and Zettlemoyer, L. 2017. End-to-end neural coreference resolution. *arXiv preprint arXiv:1707.07045*.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Liu, C. Y.; Wang, Y.; Flanigan, J.; and Liu, Y. 2024b. Large Language Model Unlearning via Embedding-Corrupted Prompts. arXiv:2406.07933.

Liu, J.; Xia, C. S.; Wang, Y.; and Zhang, L. 2023. Is Your Code Generated by ChatGPT Really Correct? Rigorous Evaluation of Large Language Models for Code Generation. arXiv:2305.01210.

Liu, S.; Yao, Y.; Jia, J.; Casper, S.; Baracaldo, N.; Hase, P.; Yao, Y.; Liu, C. Y.; Xu, X.; Li, H.; et al. 2024c. Rethinking machine unlearning for large language models. *arXiv preprint arXiv:2402.08787*.

Liu, Z.; Dou, G.; Tan, Z.; Tian, Y.; and Jiang, M. 2024d. Towards Safer Large Language Models through Machine Unlearning. arXiv:2402.10058.

Maini, P.; Feng, Z.; Schwarzschild, A.; Lipton, Z. C.; and Kolter, J. Z. 2024. Tofu: A task of fictitious unlearning for llms. *arXiv preprint arXiv:2401.06121*.

Mintz, M.; Bills, S.; Snow, R.; and Jurafsky, D. 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*, 1003–1011.

Muresanu, A.; Thudi, A.; Zhang, M. R.; and Papernot, N. 2024. Unlearnable Algorithms for In-context Learning. arXiv:2402.00751.

Satpute, A.; Gießing, N.; Greiner-Petter, A.; Schubotz, M.; Teschke, O.; Aizawa, A.; and Gipp, B. 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*, 2316–2320.

Shi, W.; Lee, J.; Huang, Y.; Malladi, S.; Zhao, J.; Holtzman, A.; Liu, D.; Zettlemoyer, L.; Smith, N. A.; and Zhang, C. 2024. MUSE: Machine Unlearning Six-Way Evaluation for Language Models.

Sileo, D. 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)*, 15655–15684.

Tian, B.; Liang, X.; Cheng, S.; Liu, Q.; Wang, M.; Sui, D.; Chen, X.; Chen, H.; and Zhang, N. 2024. To forget or not? towards practical knowledge unlearning for large language models. *arXiv preprint arXiv:2407.01920*.

Wang, B.; Zi, Y.; Sun, Y.; Zhao, Y.; and Qin, B. 2024a. RKLD: Reverse KL-Divergence-based Knowledge Distillation for Unlearning Personal Information in Large Language Models. *arXiv preprint arXiv:2406.01983*.

Wang, J.; Li, Y.; Chen, Z.; Chen, L.; Zhang, X.; and Zhou, Y. 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. ISBN 9798400702174.

Wang, L.; Zeng, X.; Guo, J.; Wong, K.-F.; and Gottlob, G. 2024c. Selective forgetting: Advancing machine unlearning techniques and evaluation in language models. *arXiv preprint arXiv:2402.05813*.

Wu, R.; Yadav, C.; Salakhutdinov, R.; and Chaudhuri, K. 2024. Evaluating Deep Unlearning in Large Language Models. arXiv:2410.15153.

Yao, J.; Chien, E.; Du, M.; Niu, X.; Wang, T.; Cheng, Z.; and Yue, X. 2024. Machine unlearning of pre-trained large language models. *arXiv preprint arXiv:2402.15159*.

Yao, Y.; Xu, X.; and Liu, Y. 2023. Large language model unlearning. *arXiv preprint arXiv:2310.10683*.

Yuan, X.; Pang, T.; Du, C.; Chen, K.; Zhang, W.; and Lin, M. 2024. A Closer Look at Machine Unlearning for Large Language Models. *arXiv preprint arXiv:2410.08109*.