# UNLEARNBENCH: A CLEAN AND CONTROLLABLE BENCHMARK FOR FINE-GRAINED AND COMPOSITIONAL LLM UNLEARNING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

LLM unlearning is essential for ensuring technical and responsible AI use, especially in addressing privacy leakage and evolving regulatory requirements. Existing unlearning benchmarks primarily evaluate unlearning through shallow question answering suppression, while overlooking several fundamental challenges: Knowledge in LLMs is acquired through complex memorization processes. It is further organized in compositional, multi-stage, and structurally entangled forms. As a result, it remains unclear whether current unlearning methods can reliably remove knowledge learned through staged training pipelines, relational reasoning, and fine-grained attribute associations. In this work, we introduce **UnlearnBench**, a benchmark for studying unlearning across diverse settings including single-hop QA, two-stage memorization, multi-hop reasoning, and fine-grained attribute-level forgetting. Compared with existing benchmarks, UnlearnBench better captures realistic memorization and retrieval behaviors while enabling systematic analysis of unlearning performance under clean and controlled settings. We additionally provide clean pretrained models and exact retraining checkpoints as gold-standard references for future evaluation. Extensive experiments on popular unlearning methods uncover several consistent failure modes. First, entity-level unlearning often fails even in settings where retraining successfully removes the target knowledge. Second, reasoning abilities can recover supposedly forgotten information from retained representations, revealing strong interactions between memorization and compositional reasoning. Third, attribute-level knowledge is highly entangled across entities and relation types, making selective forgetting substantially more difficult than existing benchmarks suggest. Our results demonstrate that current evaluations significantly underestimate the challenges of LLM unlearning and establish UnlearnBench as a structured testbed for future unlearning research.

## 1 INTRODUCTION

Training Large Language Models (LLMs) often involves using a large amount of text data, which may inadvertently contain undesirable data (Yao et al., 2025; Li et al., 2024; Harandizadeh et al., 2024; Sandbrink, 2023), such as private information (Staab et al., 2023; Neel & Chang, 2023; Mireshghallah et al., 2023) or copyrighted content (Eldan & Russinovich, 2023; Shi et al., 2024). Such sensitive content may need to be removed upon request, as mandated by regulations such as General Data Protection Regulation. LLM unlearning (Eldan & Russinovich, 2023; Yao et al., 2025) has thus become an important research direction, which aims to transform a trained model into one that behaves as if it had never been trained on certain data. Exact unlearning requires the post-unlearning model to have the same distribution as a model trained on the remaining data alone, which is often computationally expensive to achieve. As a result, many approximate unlearning methods have been proposed, optimization-based approaches (Zhang et al., 2024a; Jia et al., 2024; Ji et al., 2024; Chen & Yang, 2023), localization and model-editing methods (Cai et al., 2026; Dong et al., 2024), and training-free methods (Pawelczyk et al., 2023; Muresanu et al., 2024; Liu et al., 2025). In order to measure the performance and progress of these emergent methods, several unlearning benchmarks (Shi et al., 2024; Ramakrishna et al., 2025; Hu et al., 2025b; Wei et al., 2025b) have recently been introduced. However, many benchmarks, such as TOFU (Maini et al., 2024) and WMDP (Li et al., 2024), rely



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

on uncontrolled corpora and training settings (typically supervised finetuning), making it difficult to precisely track knowledge provenance and memorization sources. Understanding how LLMs learn and memorize knowledge is crucial for studying unlearning (Wei et al., 2025a). How knowledge is acquired and organized directly affects how difficult it is to selectively remove. Therefore, clean and controllable benchmarks with explicit knowledge provenance are essential for accurately evaluating unlearning methods. Moreover, almost all evaluations (Cao et al., 2024; Scholten et al., 2024) focus on shallow single-hop question answering (QA) recall and suppression, while overlooking compositional reasoning and multi-stage memorization settings. Current benchmarks provide limited support for fine-grained analysis, such as attribute-level and multi-hop relation-level forgetting.

To address these limitations, we introduce **UnlearnBench**, a clean and controllable benchmark for evaluating unlearning under progressively harder memorization regimes. UnlearnBench uses synthetic pretraining and staged memorization pipelines to precisely control knowledge provenance, memorization stages, and forget-retain splits. We provide the pipeline of synthetic pretraining from scratch for verbatim continuation, QA extraction skills, and reasoning ability. The benchmark supports single-hop QA, multi-hop reasoning, entity-level, fine-grained attribute-level, and relation-level forgetting, enabling systematic evaluation of unlearning methods under progressively more challenging and controlled settings. We additionally provide a suite of target LLMs and exact unlearning checkpoints obtained by retraining models from scratch only on the remaining data, which can serve as gold-standard references for performance comparison and future unlearning evaluation design. The model choice and dataset sizes are manageable for academic study.

Using UnlearnBench, we conduct systematic study of LLM unlearning methods and uncover several important findings: 1) existing baseline methods consistently fail under entity-level unlearning, despite successful retraining; 2) multi-stage memorization fundamentally changes unlearning behavior: QA fine-tuning often learns retrieval skills rather than the underlying knowledge itself, making forgetting difficult unless the memorized knowledge is removed from earlier training stages; 3) compositional reasoning can recover supposedly forgotten information from retained representations, revealing strong entanglement between memorization and reasoning; 4), attribute-level experiments show that knowledge is often organized by attribute type rather than entity identity. Our findings reveal that knowledge memorization, reasoning generalization, and attribute-level representations are deeply entangled in LLMs, making selective forgetting substantially harder than previously assumed.

Our contributions are summarized as follows:

- We introduce **UnlearnBench**, a clean, controllable, and scalable benchmark for fine-grained LLM unlearning with explicit knowledge provenance, synthetic pretraining pipelines, and large-scale data for systematic evaluation under realistic memorization settings.
- We support diverse unlearning settings beyond standard entity-level unlearning, including attribute-level, relational, multi-stage, and multi-hop unlearning, enabling the study of selective forgetting and knowledge entanglement in LLMs.
- We provide pretrained and exact retraining checkpoints, along with reliable evaluation protocols.
- Using UnlearnBench, we uncover several important failure modes of existing unlearning methods under entity-level unlearning, compositional reasoning, and multi-stage memorization settings.

## 2 Related Work

**LLM Unlearning.** LLM unlearning aims to remove specific knowledge from trained models while preserving general utility. Existing methods mainly fall into three categories: optimization-based unlearning (Jia et al., 2024; Maini et al., 2024; Wang et al., 2024c; Yao et al., 2025; Wang et al., 2025b; Wuerkaixi et al., 2025; Tan et al., 2025; Entesari et al., 2025; Zhang et al., 2025), representation-level intervention and localization (Yamashita et al., 2025; Guo et al., 2024; Cha et al., 2024; Ding et al., 2024; Kim et al., 2025; Dong et al., 2024; Cai et al., 2026), and training-free or generation-time steering methods (Pawelczyk et al., 2023; Muresanu et al., 2024; Thaker et al., 2024; Wang et al., 2024b; 2025a).

**LLM Unlearning Benchmarks.** Several benchmarks have been proposed for evaluating unlearning in LLMs (Ramakrishna et al., 2025). TOFU (Maini et al., 2024) introduces fictitious-author unlearning in a controlled synthetic setting. MUSE (Shi et al., 2024) studies copyright-related

Table 1: Dataset statistics for fine-grained unlearning settings. #ForgetPerson denotes the selected forget entities number. #RetainBio denotes the retained biographies number, including both retain and test-split entities. "245k*" indicates biographies with private attribute-related sentences removed.

| Settings | #ForgetPerson | #ForgetQA | #RetainQA | #TestQA | #ForgetBio | #RetainBio |
|---|---|---|---|---|---|---|
| Entity-level (1%) | 1000 | 6000 | 294,000 | 300,000 | 5,000 | 245k + 250k |
| Entity-level (10%) | 10,000 | 60,000 | 240,000 | 300,000 | 50,000 | 200k + 250k |
| Attribute-level (partial) | 1000 | 2000 | 298,000 | 300,000 | 5000* | 250k* + 250k |
| Attribute-level (all) | 50,000 | 100,000 | 200,000 | 300,000 | 50,000* | 250k* + 250k |
| Two-hop Entity-level (1%) | 1000 | 24,000 | 1,176,000 | 1,200,000 | 5000 | 245k + 250k |
| Two-hop Neighbor (partial) | 1000 | 6000 | 1,194,000 | 1,200,000 | 5000* | 245k* + 250k |
| Two-hop Private (partial) | 1000 | 8000 | 1,192,000 | 1,200,000 | 5000* | 245k* + 250k |

unlearning and memorization, while WMDP (Li et al., 2024) focuses on hazardous knowledge removal. RWKU (Cao et al., 2024) evaluates real-world entity unlearning with adversarial probes, and OpenUnlearning (Dorna et al., 2025) standardizes multi-benchmark evaluation. Recent work further argues that existing benchmarks are weak measures of unlearning progress (Thaker et al., 2025). Besides the benchmark itself, some literature aims to propose comprehensive and effective evaluation metrics or extended evaluation tasks (Wang et al., 2024a; Scholten et al., 2024; Hu et al., 2025a; Zhang et al., 2024b; Fan et al., 2025; Xu et al., 2025; Hu et al., 2025b).

**LLM Memorization and Reasoning during Pretraining.** Prior work shows that LLMs can memorize and verbatim reproduce training data (Carlini et al., 2020; 2022), while also storing substantial factual knowledge retrievable through prompting (Petroni et al., 2019). Recent controlled studies investigate how memorization emerges during pretraining (Prashanth et al., 2024; Wei et al., 2025a). In particular, the Physics of Language Models series (Allen-Zhu & Li, 2023b;a; 2024) studies storage, retrieval, and manipulation as separable phenomena under synthetic-person pretraining settings. Other work explores the relationship between memorization and reasoning, showing that factual QA relies more heavily on memorization while reasoning tasks exhibit stronger generalization behavior (Antoniades et al., 2024). These findings motivate the need for controllable benchmarks that jointly study memorization, reasoning, and unlearning dynamics.

# 3 UnlearnBench

## 3.1 Problem Setup

We study fine-grained LLM unlearning for personalized information and relational knowledge under two settings: (i) entity-level unlearning, where all private facts of a selected set of identities $N_f$ must be forgotten; and (ii) attribute-level unlearning, where only selected private attributes (or relation types) of a set of entities should be forgotten while other knowledge is retained, including private attributes for remaining entities. Given a trained target model $\mathcal{M}_t$, unlearning aims to produce $\mathcal{M}_u$ such that performance on forget targets drops to retrain-level behavior, while preserving remaining set knowledge and generalization ability. Let $D_f$ denote the forget set that contains the unlearning targets, and $D_r$ denote the retain set that contains the remaining data. We refer to $D_t$ as the held-out test set that is used to measure the generalization ability. The objective of unlearning is to minimize the model utility degradation on $D_r$ and $D_t$, while reducing memorization and extraction performance on $D_f$ to the level of a retrained model that has never observed the forget data.

## 3.2 Benchmark Construction

Our benchmark follows a controlled synthetic person setting inspired by prior data design allen-zhu2023storage, and extends it for unlearning evaluation. We also create our own multi-hop relational knowledge dataset.

**Single-hop biographies and QA pairs** We construct a person-centric universe with up to 100,000 synthetic individuals. Each identity contains six core attributes: `birth_date`, `birth_city`, `university`, `major`, `employer`, and `work_city`. We use structured biographies constructed from predefined templates (`BioT`). To make the dataset close to the real world, we use Qwen3-30B-A3B-Instruct qwen3technicalreport model to generate biographies with lexical and syntactic variation

(BioG_m5p++). For each entity, we have five different versions of the biography to make the training data more diverse. For each individual, we derive six QA pairs, one per attribute, producing large-scale single-hop QA corpora (qaG_1hop). We mainly focus on using BioG and qaG to conduct our experiments as it is more close to the real world and contain diverse text.

**Two-hop extension: six attributes + four relationships** We further augment each entity with four relation types (mentor, colleague, advisor, neighbor). Each biography is extended from six to ten sentences by adding four directed relationship facts. The relationship sentences identify another fictional person as the target of the relation, e.g., "⟨name⟩ was mentored by ⟨target⟩" or "⟨name⟩ grew up next door to ⟨target⟩". A two-hop query first resolves the related target person, and then retrieves one of the six attributes of that related person, producing 24 QA types per source entity. We provide both direct-answer (qaG_2hop) and chain-of-thought style variants (qaG_2hop_cot). Example question forms include: *What is the birthdate of ⟨name⟩'s mentor?*, *Which university did ⟨name⟩'s colleague study?*, and *In which city did ⟨name⟩'s neighbor work?* The answer is always the corresponding attribute value of the related target person.

**Data Scale and Statistics** Both single-hop and two-hop QA datasets contain 50,000 training entities and 50,000 testing entities. For the single-hop QA dataset, it has 300,000 training and 300,000 testing QA pairs. The two-hop QA dataset contains 1,200,000 training and 1,200,000 testing QA pairs. We maintain disjoint entity splits for training, and held-out (test) evaluation. In the two-hop data, relationship edges are restricted to stay within split, so a training identity never points to a test identity. This prevents leakage from training questions into the test set and ensures that evaluation measures genuine generalization rather than accidental exposure. Table 1 summarizes the dataset size for each unlearning setting. Note that under attribute-level settings, the forget set may contain only a subset of attributes or relation types associated with the forget entities.

**Unlearning splits** Given a trained target model, we sample a forget set of identities from the training dataset with a fixed random seed and build based on QA pairs or Biography:

- **Entity-level unlearning**: remove all knowledge of selected identities (e.g., $N_f \in \{1000, 10000\}$).
- **Attribute-level unlearning**: remove only selected private attributes (or relation subtypes), retaining the rest. For single-hop setting, we define birth_date and birth_city as the private attributes. We have two settings with partial/all scopes: 1) remove only partial private attributes from the selected 1000 entities (1% of the dataset); 2) remove all private attributes from the 50,000 entities in the training dataset. For the two-hop relational knowledge unlearning setting, we add the neighbor relationship as a private attribute. We have 1) remove neighbor relationship-related knowledge, including all six direct attributes of the neighbor, from the selected 1000 entities; 2) remove all birth_day and birth_city related private knowledge from the selected 1000 entities.
- **Joint QA+Bio unlearning splits** for simultaneous forgetting in both channels.

**Comparison with Existing Benchmarks** Existing benchmarks mainly evaluate unlearning at the entity or document level, without explicitly modeling the internal structure of knowledge. As a result, they cannot systematically study whether models can selectively forget specific attributes or relations while preserving other associated knowledge. Moreover, most prior evaluations focus on single-hop factual recall or paraphrase robustness, overlooking whether seemingly forgotten knowledge can still be recovered through compositional reasoning. Table 2 summarizes the differences between UnlearnBench and representative LLM unlearning benchmark. Compared with them, UnlearnBench provides substantially larger scale, controllable synthetic generation, fine-grained attribute- and relation-level unlearning, compositional multi-hop reasoning evaluation, and multi-stage training settings. In particular, UnlearnBench is, to the best of our knowledge, the first benchmark that systematically studies both attribute-level and relational unlearning under controllable compositional knowledge settings.

### 3.3 Target Model Training Protocols

UnlearnBench supports multiple target model training settings to better reflect real-world deployment scenarios. A reliable unlearning benchmark requires that the target knowledge is both memorized and consistently extractable from the model. Since memorization alone does not guarantee extractability (Allen-Zhu & Li, 2023b), we require that the model 1) memorizes the target knowledge, 2) can reliably extract it through QA, and 3) generalizes this extraction ability to held-out entities. To

enable clean and controlled analysis of unlearning dynamics, we investigate multiple pretraining and continual pretraining regimes and study how different training stages influence memorization and selective forgetting behaviors.

**Single-hop Mixed Training** We pretrain the model from scratch and continual train it from the Huggingface initialized weights on all Bio data and half of the QA data (training dataset). Bio and QA entries are randomly sampled without requiring them to be from the same individual. We control the QA data amount with a mix ratio to be 0.8 (2:8 Bio to QA).

**Single-hop Two Stage Training (Pretrain + Finetuning)** We pretrain the LLM from scratch and continual pre-train LLM from the HuggingFace initialized model weights on the Bio data, randomly sampling and concatenating them into 512-token sentences. The pretrained Bio model is then fine-tuned using training dataset and evaluate the generalization of knowledge extraction ability on the test dataset. Here we use both full finetuning and Lora finetuning (Hu et al., 2021). This design cleanly separates knowledge acquisition from task-specific adaptation, allowing the model to first establish a stable knowledge base and then improve its ability to use that knowledge for the downstream task.

**Two-hop Knowledge Learning** We design the pipeline through various experiments as done in Appendix B. We train the two-hop CoT model with three staged CPT pipeline: we first perform pretraining on synthetic two-hop Bio data, starting from the base model to ensure that entity attributes and relational facts are absorbed into the model parameters as the knowledge backbone. On top of the Bio model, we adapt on mixed one-hop relation (also called two-hop direct) and direct attribute QA data while replaying biography data (80% QA, 20% Bio) to strengthen hop-level retrieval for multi-hop reasoning and reduce forgetting. The final stage uses two-hop CoT training QA data to encourage compositional generalization. During the QA-oriented stages, biography replay is mixed into training to reduce catastrophic forgetting and preserve the factual knowledge acquired during the initial stage. The QA stream is balanced across one-hop relation, one-hop direct attribute, and two-hop reasoning supervision, so that the overall pipeline jointly supports knowledge acquisition, hop-level skill learning, and multi-hop reasoning and generalization.

### 3.4 Evaluation Metrics

Ideally, an unlearned model should behave as if it had never seen the forget set, while showing similar behavior to a retrained model on the retain set and test set. **QA generation accuracy (QA)** measures whether the model can correctly answer a question under greedy decoding using exact-match evaluation. **Biography first-token accuracy (Bio_ft)** is a probing metric that measures whether factual knowledge remains encoded in the model by evaluating first-token prediction accuracy on biography prompts, including both attributes and relations. Compared with QA accuracy, Bio_ft more directly reflects memorized knowledge stored in model parameters, which is beyond the extractability-focused metric. We further evaluate attribute-level memorization using **BIO Attribute-Level NLL**. Given a biography context, we compute the length-normalized token-level negative log-likelihood of target attribute values. Lower NLL indicates stronger memorization, while a larger gap between private and public attributes indicates stronger forgetting of private information. To distinguish memorization from generalization, we report metrics on both $\text{QA}_{\text{train}}$ and $\text{QA}_{\text{test}}$, where $\text{QA}_{\text{test}}$ contains held-out entities. In unlearning settings, we additionally report $\text{QA}_{\text{forget}}$ and $\text{QA}_{\text{retain}}$. Effective unlearning should substantially reduce performance on $\text{QA}_{\text{forget}}$ while preserving $\text{QA}_{\text{retain}}$ and $\text{QA}_{\text{test}}$.

## 4 Experiments

**Unlearning Baselines & Evaluation Metrics** We benchmark six representative optimization-based methods: GA, GD (Maini et al., 2024), NPO (Zhang et al., 2024a), SimNPO (Fan et al., 2024), DPO (Rafailov et al., 2023), RMU (Li et al., 2024), UnDIAL (Dong et al., 2024) and one improved localization method perTA (Cai et al., 2026) using the unlearning setting described in §3.2. For the evaluation metric, we also report the Gap of $\text{QA}_{\text{retain}}$ and $\text{QA}_{\text{forget}}$ as the overall measure of the unlearning method, $\texttt{Gap=QA}_{\texttt{retain}}\texttt{-QA}_{\texttt{forget}}$. A higher gap indicates better unlearning performance.

**Two-hop Unlearning Setup** All unlearning results are obtained using the two-hop CoT QA dataset. Bio performance is also evaluated on the two-hop biography. Unless otherwise specified, evaluation is conducted on a subset of 1000 individuals for efficiency.

Table 2: Comparison of existing LLM unlearning benchmarks and **UnlearnBench**. "SC" denotes sentence continuation, while "Instructed" refers to benchmarks that directly use externally released instruction-tuned LLMs. Dataset scale is reported as a coarse estimate.

| Benchmark | TOFU (Maini et al., 2024) | MUSE (Shi et al., 2024) | WMDP (Li et al., 2024) | RWKU (Cao et al., 2024) | BLUR (?) | LUME (Ramakrishna et al., 2025) | UnlearnBench (Ours) |
|---|---|---|---|---|---|---|---|
| **Tasks** | QA | QA + SC | QA | QA | QA | QA + SC | QA + SC |
| **Training Regime** | SFT | CPT | Instructed | Instructed | - | SFT | Pretrain / CPT / SFT |
| **Synthetic Control** | ✓ | ✓ | ✗ | ✗ | ✓ | ✓ | ✓ |
| **Fine-grained** | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ |
| **Relational Structure** | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ |
| **Multi-hop** | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ |
| **Dataset Scale** | 4,000 | 28,339 | 3,668 | 128,800 | 10,000 | 4,394 | 1,500,000 |

Table 3: The memorization and QA generalization performance of target models trained under different regimes when using Qwen3-0.6B-Base model.

| Models | Bio_ft | $QA_{train}$ | $QA_{test}$ |
|---|---|---|---|
| Qwen3-0.6B-Base | 0.0 | 0.0 | 0.0 |
| Mixed-Training | 100.00 | 99.99 | 94.73 |
| Two-Stage (Bio Pretrain) | 96.67 | 0.00 | 0.00 |
| Two-Stage (Lora Finetuning) | 87.92 | 86.95 | 86.66 |
| Two-Stage (Full Finetuning) | 97.10 | 99.85 | 93.64 |
| Three-Stage (Bio Pretrain) | 98.94 | 0.00 | 0.00 |
| Three-Stage (Direct & Bio Reply) | 93.18 | 99.61 | 8.90 |
| Three-Stage (CoT & Bio Reply) | 89.06 | 99.92 | 80.86 |

### 4.1 Target Model

We train multiple target models under five training paradigms in §3.3: mixed training, two-stage LoRA finetuning, two-stage full finetuning, two-hop three-stage CoT training, and two-hop three-stage direct replay. We use Qwen3-0.6B-Base (Team, 2025) and Llama3.2-1B (Grattafiori et al., 2024) as base models, as this model scale is commonly adopted in pretraining studies. We additionally include a nanochat (Karpathy, 2025) GPT2 model (Radford et al., 2019) to validate our synthetic pretraining pipeline. The effectiveness of these training regimes is further verified in Appendix B.

**LLM Memorization Depends on the Training Paradigm.** Table 3 shows that under single-hop mixed training, the model jointly learns factual memorization and QA extraction, achieving both strong memorization and held-out QA generalization ($QA_{\text{test}} = 94.73\%$). In two-stage training, factual knowledge is first learned during Bio pretraining and later accessed through QA fine-tuning. After Stage 1, the model memorizes entity knowledge well but cannot answer QA queries directly. After Stage 2, both LoRA and full fine-tuning successfully learn QA extraction capabilities, while full finetuning achieves performance closer to mixed training. This suggests that two-stage training separates memorized knowledge from downstream QA extraction skills. This separation significantly changes the behavior of unlearning methods. For two-hop reasoning, direct replay achieves nearly perfect training accuracy but fails to generalize to held-out QA ($QA_{\text{test}} = 8.90\%$), indicating memorization of seen reasoning relations only. In contrast, CoT replay substantially improves held-out generalization ($QA_{\text{test}} = 80.86\%$), showing that explicit CoT supervision is critical for transferable compositional reasoning. Whether knowledge is stored as reusable extraction skill or localized memorized facts is mainly controlled by the training pipeline, **not model size alone**. This further affects the difficulty of selective unlearning.

### 4.2 Main Results 1: Entity-level Information and Relational Knowledge Unlearning

**Single-hop: most tested unlearning methods fail to selectively forget target entities**, while improved PerTA and UnDIAL achieve relatively better unlearning performance. However, retraining from scratch on retain-only data can successfully remove the target knowledge, suggesting that a desirable unlearning solution exists in principle. For the 1% setting, on mixed-trained Qwen3-0.6B targets, retraining achieves near-zero forget accuracy while maintaining high retain and test performance, as shown in Table 4. In contrast, GA, GD, and SimNPO can reduce forget accuracy to near zero but often collapse retain performance simultaneously. NPO, DPO, and RMU are typically conservative and under-forget, with forget accuracy remaining close to the target model. IPerTA

Table 4: Unlearning performance under entity-level unlearning with 1% and 10% forget ratios using the Qwen3-0.6B-Base Mixed and Two Stage training target models.

| Methods | Entity-10% (Mixed) | | | | Entity-1% (Mixed) | | | | Entity-1%(Two Stage-Lora) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $QA_{forget}$ | $QA_{Retain}$ | $QA_{test}$ | Gap | $QA_{forget}$ | $QA_{Retain}$ | $QA_{test}$ | Gap | $QA_{forget}$ | $QA_{Retain}$ | $QA_{test}$ | Gap |
| Target Model | 100.00 | 99.99 | 94.73 | -0.01 | 100.00 | 99.99 | 94.73 | -0.01 | 87.30 | 86.94 | 88.67 | -0.36 |
| Retrain Model | 1.01 | 99.99 | 93.18 | 98.98 | 1.00 | 98.67 | 89.00 | 97.67 | 0.80 | 91.53 | 91.15 | 90.73 |
| GA | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 87.20 | 86.60 | 86.39 | -0.60 |
| GD | 0.00 | 2.68 | 1.71 | 2.68 | 0.00 | 0.02 | 0.01 | 0.02 | 87.32 | 86.90 | 86.64 | -0.42 |
| NPO | 99.99 | 99.99 | 94.69 | 0.00 | 99.97 | 99.99 | 94.70 | 0.02 | 77.12 | 76.26 | 75.99 | -0.86 |
| SimNPO | 0.02 | 0.01 | 0.00 | -0.01 | 53.68 | 53.06 | 47.72 | -0.62 | 48.42 | 48.24 | 47.98 | -0.18 |
| DPO | 99.99 | 99.99 | 94.76 | 0.00 | 100.00 | 99.99 | 94.74 | -0.01 | 87.50 | 86.98 | 86.72 | -0.62 |
| RMU | 99.99 | 99.99 | 94.74 | 0.00 | 99.98 | 99.99 | 94.72 | 0.01 | 87.32 | 86.90 | 86.64 | -0.42 |
| UnDIAL | - | - | - | - | 40.90 | 64.77 | 58.18 | 23.87 | - | - | - | - |
| IPerTA | 79.01 | 87.41 | 62.27 | 8.40 | 82.32 | 87.76 | 65.78 | 5.44 | 87.80 | 88.00 | 87.69 | 0.20 |

achieves intermediate forgetting with better utility preservation compared to destructive GA runs. The same pattern holds at a larger scale (10%): retraining remains the strongest reference; GA and GD incur substantial collateral damage; NPO, DPO, and RMU remain conservative; and IPerTA and UnDIAL improve over conservative baselines but still falls short of retraining. These results suggest that, under entity-level unlearning, optimization-based methods fail when forget and retain data share correlated features, consistent with Mavrothalassitis et al. (2025). This indicates that current unlearning methods fail not because selective forgetting is impossible, but because they cannot accurately identify and remove the underlying memorized knowledge—especially in challenging settings where forget and retain samples exhibit highly similar patterns and the target knowledge is strongly memorized. To bridge the gap between controlled synthetic settings and real-world data, we further validate our benchmark pipeline on real-world datasets (Appendix D.8), showing that these failure modes persist under more realistic data distributions.

**Two-hop relational knowledge unlearning is substantially harder due to compositional reasoning over retained knowledge.** Table 5 presents the unlearning results for the Stage 3 CoT reasoning model under both direct two-hop QA and CoT two-hop QA settings. Unlike single-hop unlearning, Stage 3 training mainly learns transferable compositional reasoning and generalization abilities rather than isolated factual mappings. As a result, forgetting becomes significantly more difficult. When only QA data is removed, even retraining fails to achieve effective forgetting, with the retrained model still obtaining 88.42% on $QA_{forget}$. This indicates that compositional reasoning can recover forgotten relations from the retained biography knowledge. Only when both retain-only QA data and retain-only biography data are used during retraining can successful forgetting be achieved, reducing $QA_{forget}$ to 8.60% (Appendix B). Among existing unlearning methods, RMU and Iterative PerTA show relatively better forgetting behavior, producing a noticeable gap between $QA_{forget}$ and $QA_{retain}$. This effect becomes more visible under CoT supervision, suggesting partial suppression of reasoning paths. However, most methods either retain the target knowledge (e.g., DPO) or globally degrade both forget and retain knowledge (e.g., GA), rather than achieving reliable selective forgetting. We also conduct experiments on the Stage 2 model using both CoT and direct two-hop data (Table 19). However, the results remain unsatisfactory, further suggesting that two-hop reasoning introduces substantially stronger knowledge entanglement, where forgotten information can still be extracted through retained intermediate relations. Overall, two-hop reasoning introduces substantially stronger knowledge entanglement, where forgotten information can still be reconstructed through retained intermediate relations. This makes entity-level selective forgetting fundamentally more challenging than in single-hop settings.

### 4.3 Main Results 2: Multi-stage Memorization and Unlearning

**When the model learns question answering as a transferable skill through fine-tuning on a Bio pretrained base model, selective unlearning may become difficult. In contrast, when factual knowledge is learned directly through mixed training, selective forgetting becomes more feasible.** We study unlearning behavior under a two-stage training setting for both LoRA fine-tuning and full fine-tuning models. In the two-stage LoRA setting, factual knowledge is accessed through shared QA extraction pathways rather than separable person-specific parameter subsets. The learned QA capability generalizes across both seen and unseen entities, making selective person-level forgetting

Table 5: Entity-level unlearning performance under two-hop relational QA settings. We evaluate unlearning using either direct two-hop QA or CoT two-hop QA during Stage 3. "QA only (Stage 3)" denotes models unlearned with QA only, while "Bio+QA (Stage 3)" additionally includes biography. $QA_{forget}$, $QA_{retain}$, and $QA_{test}$ denote generation accuracy on the CoT forget, retain, and test sets, respectively. $Bio_ft_F$ and $Bio_ft_R$ denote biography first-token accuracy on the forget and retain sets.

| Method | Two-hop Direct QA | | | | | Two-hop CoT QA | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $QA_{forget}$ | $QA_{retain}$ | $QA_{test}$ | $Bio_ft_F$ | $Bio_ft_R$ | $QA_{forget}$ | $QA_{retain}$ | $QA_{test}$ | $Bio_ft_F$ | $Bio_ft_R$ |
| **QA only (Stage 3)** | | | | | | | | | | |
| Target | 99.94 | 99.92 | 80.86 | 88.53 | 89.20 | 99.94 | 99.92 | 80.86 | 88.53 | 89.20 |
| Retrain (S2+QA) | 88.42 | 99.92 | 79.83 | 95.00 | 96.33 | 88.42 | 99.92 | 79.83 | 95.00 | 96.33 |
| GA | 0.0 | 0.0 | 0.0 | 64.25 | 63.47 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| GD | 0.0 | 0.0 | 0.0 | 17.28 | 12.98 | 0.20 | 2.07 | 0.60 | 39.57 | 40.85 |
| NPO | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 92.75 | 92.18 | 44.52 | 87.75 | 89.20 |
| SimNPO | 38.87 | 38.90 | 18.90 | 93.25 | 94.40 | 0.0 | 0.0 | 0.0 | 5.37 | 5.25 |
| DPO | 99.95 | 99.92 | 80.99 | 88.43 | 89.12 | 99.94 | 99.92 | 80.94 | 88.20 | 88.97 |
| RMU | 97.94 | 99.89 | 79.55 | 87.65 | 89.85 | 89.70 | 99.90 | 80.75 | 88.97 | 99.94 |
| IPerTA | 3.07 | 5.06 | 2.62 | 1.70 | 1.22 | 12.90 | 21.76 | 5.51 | 33.98 | 34.10 |
| **Bio+QA (Stage 3)** | | | | | | | | | | |
| Retrain (S2+Bio+QA) | 8.60 | 99.92 | 81.48 | 92.03 | 96.62 | 8.60 | 99.92 | 81.48 | 92.03 | 96.62 |
| GA (ep1) | 0.0 | 0.0 | 0.0 | 15.00 | 14.33 | 0.0 | 0.0 | 0.0 | 84.67 | 87.67 |
| GD | 0.0 | 0.0 | 0.0 | 0.67 | 1.00 | 0.0 | 0.0 | 0.0 | 40.67 | 40.33 |
| NPO | 86.92 | 87.08 | 41.25 | 82.00 | 80.00 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| SimNPO | 0.0 | 0.0 | 0.0 | 17.33 | 15.67 | 45.42 | 44.17 | 19.67 | 92.00 | 95.00 |
| DPO | 100.00 | 99.83 | 81.17 | 88.67 | 89.33 | 100.00 | 99.83 | 81.33 | 88.67 | 90.00 |
| RMU (ep3) | 99.08 | 99.83 | 79.08 | 88.00 | 88.67 | 100.00 | 99.83 | 80.67 | 89.00 | 90.00 |
| IPerTA | 0.92 | 1.00 | 1.00 | 2.67 | 2.67 | 0.47 | 2.22 | 0.00 | 1.00 | 1.00 |

structurally difficult. Only QA data can not remove the memorized knowledge. As shown in Table 4, most all optimization-based methods exhibit conservative behavior and suffer from under-forgetting. The same trend is also shown under partial attribute-level unlearning in Table 15.

For the full finetuning setting, we observe that entity-level unlearning consistently fails when only QA data is used as the forget set. Even retraining cannot reliably remove the target knowledge in this setting. The reason is that Stage 2 mainly learns a generalizable QA extraction skill, whereas the underlying factual knowledge has already been memorized during Stage 1 biography pretraining. As a result, as long as the Stage 1 model still retains knowledge about the forget entities, retraining Stage 2 on retain-only QA data cannot fully eliminate such knowledge. In contrast, successful forgetting can be achieved when both Bio and QA data are removed during retraining, and the Stage 1 model itself is retrained using retain-only Bio data. This result suggests that forgetting fundamentally depends on whether the memorized entity knowledge remains in the pretrained model. We further investigate whether existing unlearning methods can directly remove such knowledge. Specifically, we apply unlearning methods on both Bio and QA data after Stage 2 training. However, most existing methods still fail to achieve reliable selective forgetting (Table 18).

### 4.4 Main Results 3: Attribute-level Unlearning

**Single-hop: RMU performs best overall under the attribute-level all setting, while other methods partially separate public and private attributes but struggle with fine-grained selective unlearning (Partial).** Figure 1 summarizes the trade-off between forgetting selected private attributes and preserving public knowledge. Under the attribute-level all setting, RMU achieves the strongest forgetting performance while largely preserving public attributes in both mixed and two-stage training settings. GD and IPerTA partially separate private and public attributes, but with substantially lower utility preservation. In contrast, under the partial forget-people setting, where only the private attributes of 1,000 individuals are removed, most methods fail to selectively forget target people while preserving the same attributes for retain people. Although GA, GD, SimNPO, and IPerTA introduce a gap between private and public QA performance under mixed training, the degradation also propagates to retain and test private attributes, indicating that these methods remove attribute knowledge globally rather than selectively by person identity (Also see Table 14). Retraining achieves near-optimal forgetting under the mixed setting, but this advantage largely disappears in

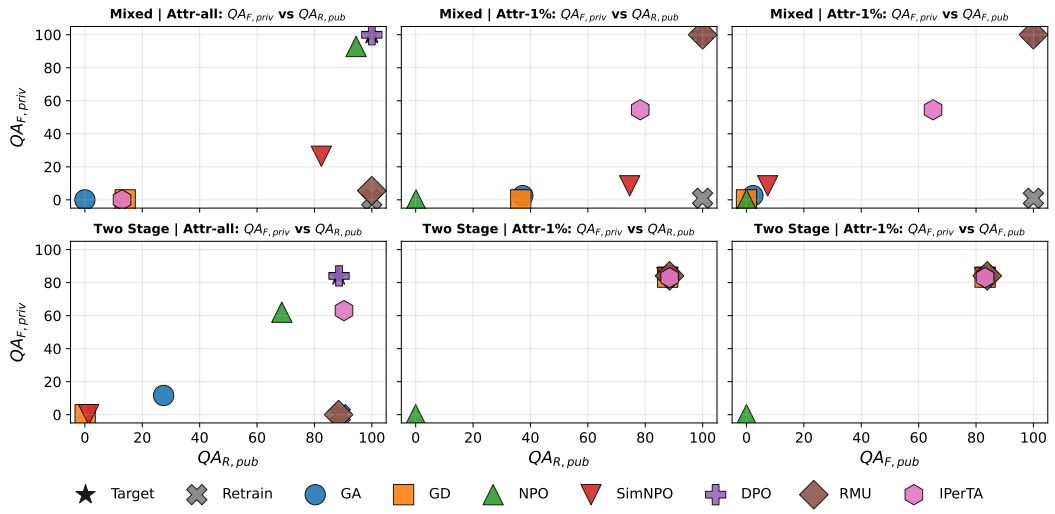


Figure 1: Attribute-level Unlearning under Mixed and Two Stage Training Qwen Models. A lower y-axis means stronger forgetting; higher x-axis means better preservation utility.

the two-stage setting, where both retraining and existing unlearning methods maintain high forget set performance. This suggests that, under two-stage memorization, the model primarily learns a transferable QA extraction skill, making selective attribute-level forgetting difficult.

**Two-hop: Neighbor-related relationships are harder to selectively remove, while private attributes (e.g., birth information) across relations show relatively stronger unlearning performance.** In the two-hop setting, we evaluate unlearning using both direct QA and CoT QA forget sets. For relation-level unlearning (Appendix Table 20), most methods fail to selectively forget neighbor-related knowledge. Performance drops are largely uniform across $\text{QA}_{\text{F, neighbor}}$, $\text{QA}_{\text{F, other}}$, and $\text{QA}_{\text{retain}}$, indicating entangled representations across relation types. In particular, even the best-performing method (GD) only reduces neighbor-related forget accuracy to 42.52%, while the accuracy on other relations remains 78.32% and retain performance is 70.17%. This suggests that GD can only partially suppress target relations while preserving non-target knowledge. For private attribute unlearning (Appendix Table 21), methods such as GA, NPO, and SimNPO achieve partial forgetting under direct evaluation while largely maintaining retain performance. However, under CoT unlearning, performance further deteriorates or becomes inconsistent, suggesting that reasoning amplifies instability. Although a gap between forget and retain sets still exists, the overall performance drops significantly, indicating substantial degradation of general model capability. These results highlight that effective unlearning in two-hop settings requires structure-aware methods that can balance three competing objectives: removing target knowledge, preserving non-target knowledge, and generalizing to unseen queries. Relation-specific and private-attribute-level two-hop forgetting remains an open challenge.

## 5 Conclusions

We introduce **UnlearnBench**, a benchmark supporting entity-level, attribute-level, relational, and multi-hop unlearning under explicit memorization pipelines, enabling systematic evaluation of unlearning behavior. Through extensive experiments, we uncover consistent failure modes in existing unlearning methods under entity-level forgetting, multi-stage memorization, compositional reasoning, and knowledge entanglement settings, even when exact retraining succeeds. Our results further show that the memorization process fundamentally affects unlearning behavior: selective forgetting becomes substantially harder when question answering is learned as a transferable retrieval skill, while mixed training makes selective unlearning more feasible. Overall, our findings suggest that selective unlearning in modern LLMs is more challenging than existing benchmarks indicate. We hope UnlearnBench and the released checkpoints can support future research on reliable and fine-grained LLM unlearning.

## AI USAGE STATEMENT

We used LLMs to generate synthetic biographies and questions from predefined factual schemas. AI tools also assisted with manuscript drafting and language editing. The authors take responsibility for the final text, claims, and research artifacts.

## ETHICS STATEMENT

The core benchmark uses fictional entities and synthetic biographies. Supplementary experiments use the existing MUSE benchmark (Shi et al., 2024).

## REPRODUCIBILITY STATEMENT

Dataset construction and splits are described in Appendix A.1, training protocols in Appendix B, unlearning objectives in Appendix C, and evaluation metrics in Appendix A.3. Detailed results and hardware information appear in Appendix D.

## REFERENCES

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.2, knowledge manipulation. *arXiv preprint arXiv:2309.14402*, 2023a.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.1, knowledge storage and extraction. *arXiv preprint arXiv:2309.14316*, 2023b.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.3, knowledge capacity scaling laws. *arXiv preprint arXiv:2404.05405*, 2024.

Antonis Antoniades, Xinyi Wang, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, and William Yang Wang. Generalization v.s. memorization: Tracing language models' capabilities back to pretraining data. *ArXiv*, abs/2407.14985, 2024. URL https://api.semanticscholar.org/CorpusID:271328219.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: Perfect linear concept erasure in closed form. *arXiv preprint arXiv:2306.03819*, 2023.

Karuna Bhaila, Minh-Hao Van, and Xintao Wu. Soft prompting for unlearning in large language models. *arXiv preprint arXiv:2406.12038*, 2024.

Chengyi Cai, Zesheng Ye, Jiangchao Yao, Jianzhong Qi, Bo Han, Xiaolu Zhang, Feng Liu, and Jun Zhou. Per-parameter task arithmetic for unlearning in large language models. *arXiv preprint arXiv:2601.22030*, 2026.

Pengfei Cao, Chenhao Wang, Zhitao He, Hongbang Yuan, Jiachun Li, Yubo Chen, Kang Liu, Jun Zhao, et al. Rwku: Benchmarking real-world knowledge unlearning for large language models. *Advances in Neural Information Processing Systems*, 37:98213–98263, 2024.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Xiaodong Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *USENIX Security Symposium*, 2020. URL https://api.semanticscholar.org/CorpusID:229156229.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. Quantifying memorization across neural language models. *ArXiv*, abs/2202.07646, 2022. URL https://api.semanticscholar.org/CorpusID:246863735.

Sungmin Cha, Sungjun Cho, Dasol Hwang, and Moontae Lee. Towards robust and parameter-efficient knowledge unlearning for llms. *arXiv preprint arXiv:2408.06621*, 2024.

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

Yuheng Chen, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. Knowledge localization: Mission not accomplished? enter query localization! *arXiv preprint arXiv:2405.14117*, 2024.

Minseok Choi, Daniel Rim, Dohyun Lee, and Jaegul Choo. Snap: Unlearning selective knowledge in large language models with negative instructions. *arXiv preprint arXiv:2406.12329*, 2024.

Aghyad Deeb and Fabien Roger. Do unlearning methods remove information from language model weights? *arXiv preprint arXiv:2410.08827*, 2024.

Zhijie Deng, Chris Yuhao Liu, Zirui Pang, Xinlei He, Lei Feng, Qi Xuan, Zhaowei Zhu, and Jiaheng Wei. Guard: Generation-time llm unlearning via adaptive restriction and detection. *arXiv preprint arXiv:2505.13312*, 2025.

Chenlu Ding, Jiancan Wu, Yancheng Yuan, Jinda Lu, Kai Zhang, Alex Su, Xiang Wang, and Xiangnan He. Unified parameter-efficient unlearning for llms. *arXiv preprint arXiv:2412.00383*, 2024.

Yijiang River Dong, Hongzhou Lin, Mikhail Belkin, Ramon Huerta, and Ivan Vulić. Undial: Self-distillation with adjusted logits for robust unlearning in large language models. *arXiv preprint arXiv:2402.10052*, 2024.

Vineeth Dorna, Anmol Mekala, Wenlong Zhao, Andrew McCallum, Zachary C Lipton, J Zico Kolter, and Pratyush Maini. Openunlearning: Accelerating llm unlearning via unified benchmarking of methods and metrics. *arXiv preprint arXiv:2506.12618*, 2025.

Ronen Eldan and Mark Russinovich. Who's harry potter? approximate unlearning for llms. 2023.

Taha Entesari, Arman Hatami, Rinat Khaziev, Anil Ramakrishna, and Mahyar Fazlyab. Constrained entropic unlearning: A primal-dual framework for large language models. *arXiv preprint arXiv:2506.05314*, 2025.

Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Simplicity prevails: Rethinking negative preference optimization for llm unlearning. *arXiv preprint arXiv:2410.07163*, 2024.

Chongyu Fan, Jinghan Jia, Yihua Zhang, Anil Ramakrishna, Mingyi Hong, and Sijia Liu. Towards llm unlearning resilient to relearning attacks: A sharpness-aware minimization perspective and beyond. *arXiv preprint arXiv:2502.05374*, 2025.

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

Tianle Gu, Kexin Huang, Ruilin Luo, Yuanqi Yao, Yujiu Yang, Yan Teng, and Yingchun Wang. Meow: Memory supervised llm unlearning via inverted facts. *arXiv preprint arXiv:2409.11844*, 2024.

Phillip Guo, Aaquib Syed, Abhay Sheshadri, Aidan Ewart, and Gintare Karolina Dziugaite. Mechanistic unlearning: Robust knowledge unlearning and editing via mechanistic localization. *arXiv preprint arXiv:2410.12949*, 2024.

Bahareh Harandizadeh, Abel Salinas, and Fred Morstatter. Risk and response in large language models: Evaluating key threat categories. *arXiv preprint arXiv:2403.14988*, 2024.

Yihuai Hong, Lei Yu, Haiqin Yang, Shauli Ravfogel, and Mor Geva. Intrinsic test of unlearning using parametric knowledge traces. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 19524–19546, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Shengyuan Hu, Yiwei Fu, Steven Wu, and Virginia Smith. Unlearning or obfuscating? jogging the memory of unlearned llms via benign relearning. In *International Conference on Learning Representations*, volume 2025, pp. 8857–8888, 2025a.

Shengyuan Hu, Neil Kale, Pratiksha Thaker, Yiwei Fu, Steven Wu, and Virginia Smith. Blur: A benchmark for llm unlearning robust to forget-retain overlap. *arXiv preprint arXiv:2506.15699*, 2025b.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.

Jiabao Ji, Yujian Liu, Yang Zhang, Gaowen Liu, Ramana Kompella, Sijia Liu, and Shiyu Chang. Reversing the forget-retain objectives: An efficient llm unlearning framework from logit difference. *Advances in Neural Information Processing Systems*, 37:12581–12611, 2024.

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

Andrej Karpathy. nanochat: The best chatgpt that $100 can buy, 2025. URL https://github.com/karpathy/nanochat.

Yejin Kim, Eunwon Kim, Buru Chang, and Junsuk Choe. Improving fisher information estimation and efficiency for lora-based llm unlearning. *arXiv preprint arXiv:2508.21300*, 2025.

Hwiyeong Lee, Uiji Hwang, Hyelim Lim, and Taeuk Kim. Does localization inform unlearning? a rigorous examination of local parameter attribution for knowledge unlearning in language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21868–21880, 2025.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, et al. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*, 2024.

Chris Liu, Yaxuan Wang, Jeffrey Flanigan, and Yang Liu. Large language model unlearning via embedding-corrupted prompts. *Advances in Neural Information Processing Systems*, 37:118198–118266, 2025.

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

Ioannis Mavrothalassitis, Pol Puigdemont, Noam Levi, and Volkan Cevher. Ascent fails to forget. *ArXiv*, abs/2509.26427, 2025. URL https://api.semanticscholar.org/CorpusID:281682873.

Anmol Mekala, Vineeth Dorna, Shreya Dubey, Abhishek Lalwani, David Koleczek, Mukund Rungta, Sadid Hasan, and Elita Lobo. Alternate preference optimization for unlearning factual knowledge in large language models. *arXiv preprint arXiv:2409.13474*, 2024.

Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. Can llms keep a secret? testing privacy implications of language models via contextual integrity theory. *arXiv preprint arXiv:2310.17884*, 2023.

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

Seth Neel and Peter Chang. Privacy issues in large language models: A survey. *arXiv preprint arXiv:2312.06717*, 2023.

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. In-context unlearning: Language models as few shot unlearners. *arXiv preprint arXiv:2310.07579*, 2023.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. Language models as knowledge bases? *ArXiv*, abs/1909.01066, 2019. URL https://api.semanticscholar.org/CorpusID:202539551.

USVSN Sai Prashanth, Alvin Deng, Kyle O'Brien, V JyothirS, Mohammad Aflah Khan, Jaydeep Borkar, Christopher A. Choquette-Choo, Jacob Ray Fuehne, Stella Biderman, Tracy Ke, Katherine Lee, and Naomi Saphra. Recite, reconstruct, recollect: Memorization in lms as a multifaceted phenomenon. *ArXiv*, abs/2406.17746, 2024. URL https://api.semanticscholar.org/CorpusID:270710931.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. URL https://api.semanticscholar.org/CorpusID:160025533.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

Anil Ramakrishna, Yixin Wan, Xiaomeng Jin, Kai-Wei Chang, Zhiqi Bu, Bhanukiran Vinzamuri, Volkan Cevher, Mingyi Hong, and Rahul Gupta. Lume: Llm unlearning with multitask evaluations. *arXiv preprint arXiv:2502.15097*, pp. 10, 2025.

Jonas B Sandbrink. Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools. *arXiv preprint arXiv:2306.13952*, 2023.

Yan Scholten, Stephan Günnemann, and Leo Schwinn. A probabilistic perspective on unlearning and alignment for large language models. *arXiv preprint arXiv:2410.03523*, 2024.

Nazanin Mohammadi Sepahvand, Eleni Triantafillou, Hugo Larochelle, Doina Precup, James J Clark, Daniel M Roy, and Gintare Karolina Dziugaite. Selective unlearning via representation erasure using domain adversarial training. In *The Thirteenth International Conference on Learning Representations*, 2025.

William F Shen, Xinchi Qiu, Meghdad Kurmanji, Alex Iacob, Lorenzo Sani, Yihong Chen, Nicola Cancedda, and Nicholas D Lane. Lunar: Llm unlearning via neural activation redirection. *arXiv preprint arXiv:2502.07218*, 2025.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A Smith, and Chiyuan Zhang. Muse: Machine unlearning six-way evaluation for language models. *arXiv preprint arXiv:2407.06460*, 2024.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Beyond memorization: Violating privacy via inference with large language models. *arXiv preprint arXiv:2310.07298*, 2023.

Yize Sui, Jing Ren, Wenjing Yang, Ruochun Jin, Liyang Xu, Xiyao Liu, and Ji Wang. Elastic robust unlearning of specific knowledge in large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Chenchen Tan, Youyang Qu, Xinghao Li, Hui Zhang, Shujie Cui, Cunjian Chen, and Longxiang Gao. Wisdom is knowing what not to say: Hallucination-free llms unlearning via attention shifting. *arXiv preprint arXiv:2510.17210*, 2025.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Pratiksha Thaker, Yash Maurya, Shengyuan Hu, Zhiwei Steven Wu, and Virginia Smith. Guardrail baselines for unlearning in llms. *arXiv preprint arXiv:2403.03329*, 2024.

Pratiksha Thaker, Shengyuan Hu, Neil Kale, Yash Maurya, Zhiwei Steven Wu, and Virginia Smith. Position: Llm unlearning benchmarks are weak measures of progress. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 520–533. IEEE, 2025.

Qizhou Wang, Bo Han, Puning Yang, Jianing Zhu, Tongliang Liu, and Masashi Sugiyama. Towards effective evaluations and comparisons for llm unlearning methods. *arXiv preprint arXiv:2406.09179*, 2024a.

Shang Wang, Tianqing Zhu, Dayong Ye, and Wanlei Zhou. When machine unlearning meets retrieval-augmented generation (rag): Keep secret or forget knowledge? *arXiv preprint arXiv:2410.15267*, 2024b.

Yaxuan Wang, Jiaheng Wei, Chris Yuhao Liu, Jinlong Pang, Quan Liu, Ankit Parag Shah, Yujia Bao, Yang Liu, and Wei Wei. Llm unlearning via loss adjustment with only forget data. *arXiv preprint arXiv:2410.11143*, 2024c.

Yaxuan Wang, Chris Yuhao Liu, Quan Liu, Jinglong Pang, Wei Wei, Yujia Bao, and Yang Liu. Dragon: Guard llm unlearning in context via negative detection and reasoning. *arXiv preprint arXiv:2511.05784*, 2025a.

Yaxuan Wang, Yuhao Liu, Quan Liu, Jinlong Pang, Wei Wei, Yujia Bao, and Yang Liu. Dragon: Guard llm unlearning in context via negative detection and reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 145078–145126, 2026a.

Yue Wang, Qizhou Wang, Feng Liu, Wei Huang, Yali Du, Xiaojiang Du, and Bo Han. Gru: Mitigating the trade-off between unlearning and retention for llms. *arXiv preprint arXiv:2503.09117*, 2025b.

Yuze Wang, Yujia Tong, Ke Xu, Jingling Yuan, Jiawei Jiang, and Chuang Hu. Sparsity-aware unlearning for large language models. *arXiv preprint arXiv:2602.00577*, 2026b.

Johnny Tian-Zheng Wei, Ameya Godbole, Mohammad Aflah Khan, Ryan Wang, Xiaoyuan Zhu, James Flemings, Nitya Kashyap, Krishna P. Gummadi, Willie Neiswanger, and Robin Jia. Hubble: a model suite to advance the study of llm memorization. *ArXiv*, abs/2510.19811, 2025a. URL https://api.semanticscholar.org/CorpusID:282272225.

Rongzhe Wei, Peizhi Niu, Hans Hao-Hsun Hsu, Ruihan Wu, Haoteng Yin, Mohsen Ghassemi, Yifan Li, Vamsi K Potluru, Eli Chien, Kamalika Chaudhuri, et al. Do llms really forget? evaluating unlearning with knowledge correlation and confidence awareness. *arXiv preprint arXiv:2506.05735*, 2025b.

Xinwei Wu, Junzhuo Li, Minghui Xu, Weilong Dong, Shuangzhi Wu, Chao Bian, and Deyi Xiong. Depn: Detecting and editing privacy neurons in pretrained language models. *arXiv preprint arXiv:2310.20138*, 2023.

Abudukelimu Wuerkaixi, Qizhou Wang, Sen Cui, Wutong Xu, Bo Han, Gang Niu, Masashi Sugiyama, and Changshui Zhang. Adaptive localization of knowledge negation for continual llm unlearning. In *Forty-second International Conference on Machine Learning*, 2025.

Haoming Xu, Ningyuan Zhao, Liming Yang, Sendong Zhao, Shumin Deng, Mengru Wang, Bryan Hooi, Nay Oo, Huajun Chen, and Ningyu Zhang. Relearn: Unlearning via learning for large language models. *arXiv preprint arXiv:2502.11190*, 2025.

Tomoya Yamashita, Akira Ito, Yuuki Yamanaka, Masanori Yamada, Takayuki Miura, and Toshiki Shibahara. Sparse-autoencoder-guided internal representation unlearning for large language models. *arXiv preprint arXiv:2509.15631*, 2025.

Zhengbang Yang, Yisheng Zhong, Junyuan Hong, and Zhuangdi Zhu. Catnip: Llm unlearning via calibrated and tokenized negative preference alignment. *arXiv preprint arXiv:2602.02824*, 2026.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. *Advances in Neural Information Processing Systems*, 37:105425–105475, 2025.

Chenlong Zhang, Zhuoran Jin, Hongbang Yuan, Jiaheng Wei, Tong Zhou, Kang Liu, Jun Zhao, and Yubo Chen. Rule: Reinforcement unlearning achieves forget-retain pareto optimality. *arXiv preprint arXiv:2506.07171*, 2025.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. *arXiv preprint arXiv:2404.05868*, 2024a.

Zhiwei Zhang, Fali Wang, Xiaomin Li, Zongyu Wu, Xianfeng Tang, Hui Liu, Qi He, Wenpeng Yin, and Suhang Wang. Catastrophic failure of llm unlearning via quantization. *arXiv preprint arXiv:2410.16454*, 2024b.

Alan Zhu, Parth Asawa, Jared Quincy Davis, Lingjiao Chen, Boris Hanin, Ion Stoica, Joseph E Gonzalez, and Matei Zaharia. Bare: Leveraging base language models for few-shot synthetic data generation. *arXiv preprint arXiv:2502.01697*, 2025.

## APPENDIX CONTENTS

The Appendix is organized as follows.


- **Appendix A**: UnlearnBench Details
  - **Section § A.1** Benchmark Construction
  - **Section § A.2** Benchmark Comparison
  - **Section § A.3** Evaluation Metrics
- **Appendix B**: Target Model Training Protocols
  - **Section § B.1** Single-hop Memorization Regimes
  - **Section § B.2** Two-hop Relational Memorization
  - **Section § B.3** Retrained Oracle Models
- **Appendix C**: Benchmarking Unlearning Method
  - Optimization-based Baselines
  - Localization and Representation-level Unlearning
- **Appendix D**: More Experimental Results
  - **Section § D.1** Single-hop Entity-level Unlearning
  - **Section § D.2** Single-hop Attribute-level Unlearning
  - **Section § D.3** Single-hop Two-Stage Unlearning (Detailed)
  - **Section § D.4** Two-hop Entity-Level Unlearning
  - **Section § D.5** Two-hop Attribute-level Unlearning
  - **Section § D.6** Unlearn Both Bio and QA Dataset
  - **Section § D.7** Unlearning Process Analysis
  - **Section § D.8** Extend to Real-World Data
  - **Section § D.9** Additional Experiments
    - Sensitivity Study of the Dataset Generator
    - Randomized Attribute-level Unlearning Setting
    - Robustness Analysis under Different Attackers and Probes
    - Inference Time Unlearning Results
- **Appendix E**: Related Work
- **Appendix F**: Discussion


## LIMITATIONS

Although **UnlearnBench** provides a clean and controllable environment for studying fine-grained LLM unlearning, several limitations remain. First, our benchmark primarily relies on synthetic pretraining corpora and controlled QA generation pipelines. While this design enables precise control over knowledge provenance and memorization dynamics, synthetic settings may not fully capture the complexity of real-world web-scale corpora. To partially address this, we **additionally extend our training and evaluation pipeline to real-world datasets** in Appendix D.8. Second, our experiments mainly focus on small and medium-scale language models due to computational constraints. Evaluating larger frontier-scale models under controlled pretraining settings remains a difficult but important direction for future work. Finally, our work focuses on benchmark construction and empirical analysis rather than proposing new unlearning algorithms. We hope the released pretrained models, retraining checkpoints, challenging unlearning settings, and evaluation protocols can support future research on more reliable and interpretable LLM unlearning.

## BROADER IMPACT

This work aims to advance reliable and fine-grained LLM unlearning, which has important positive societal implications for privacy protection, copyrighted content removal, and regulatory compliance.

By providing a clean and controllable benchmark, UnlearnBench enables more rigorous evaluation of whether sensitive information can be effectively removed from LLMs. At the same time, our findings suggest that selective unlearning in LLMs is substantially more difficult than existing evaluations indicate. Imperfect unlearning methods may create a false sense of data removal while sensitive knowledge remains recoverable through reasoning or related representations. We hope this work encourages the development of more reliable and transparent unlearning methods and evaluation protocols.

# A UnlearnBench Details

## A.1 Benchmark Construction

UnlearnBench is a controllable synthetic benchmark for studying fine-grained and compositional LLM unlearning. The benchmark contains biography corpora, single-hop QA data, and two-hop relational QA data. All data are generated from fictional entities with deterministic construction rules and fixed random seeds. We mainly generate two types of data: 1) template-generated data and 2) LLM-paraphrased data generated by Qwen models. This allows us to study unlearning under both clean templated distributions and more natural rewritten text distributions (our focus).

**Templated Biography and QA Data** Each fictional person is associated with six attributes. For every entity, we generate a six-sentence biography, where each sentence describes one attribute. The template-based biography dataset is denoted as `BioT`. To improve memorization and extraction robustness, we apply multiple sentence templates and sentence-order permutation In the standard `m5p` setting, each person has five rewritten and permuted biography variants. From each biography, we construct six single-hop QA pairs, one for each attribute. The resulting QA dataset is denoted as `qaT_1hop`. Example QA templates are shown in Listing 2.

```
Marco Makenzie Kaye pursued professional opportunities in <
Cambridge, MA>. He was part of a dynamic work environment at <
Moderna>. He arrived in this world on <July 2, 1966>, a day to
be remembered. He conducted research on cutting-edge <Statistics>
 topics. He studied at <Santa Clara University>. He has a strong
 attachment to <Mobile, AL>.
```

Listing 1: One example biography from BioT.

```
birth_date: What is the birth date/date of birth of {name}?
birth_city: What city was {name} born in?
university: Which university did {name} study?
major: What major did {name} study?
employer: Which company did {name} work for?
work_city: In which city did {name} work?
```

Listing 2: QA example for six attributes.

```
Marco Makenzie Kaye's commitment to excellence, integrity, and
innovation drives his success in the field of <Statistics>. He
dedicated his career to scientific advancement and innovation,
eventually becoming a valued professional at <Moderna>. He
continued his academic and professional journey at the <Santa
Clara University>, where he contributed to impactful research.
He currently resides in <Cambridge, MA>, where he continue to
contribute his expertise and dedication to his work,
exemplifying excellence in science and innovation. His life and
work reflect enduring commitment and excellence, beginning on <
July 2, 1966>. He is a dedicated botanist and innovator whose
passion for plant science began in his hometown of <Mobile, AL>.
```

Listing 3: One example biography from BioG.

**LLM Generated Biography and Setup** For the generated biography dataset (`BioG`), we use Qwen3-30B-A3B-Instruct-2507 to paraphrase the biographies into more natural language. The model receives structured attribute records and generates short biography passages while preserving all factual values exactly. We apply rejection sampling to filter generations that fail factual consistency checks. Unless otherwise specified, we generate five rewritten biographies per entity. The generated biographies and QA pairs keep the same underlying facts as the template-generated versions while introducing more diverse linguistic patterns. For the standard single-hop `bioG` setup, we generate $P = 5$ rewritten biographies per person. Unless otherwise specified, we use the following decoding and runtime parameters: temperature $= 1.0$, top-$p = 1.0$, top-$k = -1$, min-$p = 0.0$, presence penalty $= 0.0$, repetition penalty $= 1.0$, maximum generation length $= 200$ tokens, tensor parallelism $= 2$, data parallelism $= 4$, and SGLang memory fraction $= 0.80$. The default random seed is $12345$. The resulting output is saved as `bioG_m5p.jsonl`, and the same model is also used to paraphrase the corresponding single-hop QA questions, with the answer strings kept unchanged.

**Two-hop Relational Data** To study compositional reasoning and relational unlearning, we extend each biography with four relationship facts: mentor, colleague, academic advisor, and neighbor. Each relationship points to another fictional entity. As a result, each two-hop biography contains six attribute sentences and four relationship sentences. The relationship assignments satisfy four constraints. First, self-links are not allowed. Second, source and target entities must belong to the same split. Third, the four relationship targets for one entity must be distinct. Fourth, for each relationship type separately, every entity appears exactly once as a target. These constraints ensure balanced relational coverage and prevent information leakage across splits.

**Two-hop QA Construction** Two-hop questions ask about an attribute of a related person. For every combination of four relationship types and six target attributes, we generate one QA type, resulting in 24 two-hop QA types per entity. The answer is always the corresponding attribute value of the related target person. We additionally construct a chain-of-thought (CoT) variant for two-hop QA. Instead of directly predicting the final attribute value, the model first resolves the relationship target and then retrieves the target attribute. This decomposition separates relational reasoning from attribute retrieval and improves compositional generalization.

**Data Splits and Scale** We maintain disjoint entity splits for forgetting, retaining, and held-out evaluation. In the two-hop setting, all relationship edges remain within the same split, ensuring that training entities never connect to held-out test entities. This prevents information leakage from training data into evaluation data. In the standard setup, biography pretraining uses 100K fictional entities with five biography variants per entity, resulting in 500K biography documents. The single-hop QA dataset contains 50K training entities with six QA pairs per entity, resulting in 300K training QA pairs. For two-hop QA, each entity contributes 24 QA types and five question variants per type, resulting in approximately 6M training QA examples.

**Joint QA and Biography Unlearning** To evaluate cross-format forgetting, we additionally construct mixed unlearning datasets that combine QA forget samples and biography forget samples. This setting tests whether unlearning methods can consistently suppress target knowledge across both QA generation and biography memorization.

**IDK Refusal Construction** For each scenario, we construct a forget set and a retain set from the training data. For preference-based unlearning methods, we additionally construct a refusal-style preferred set containing "I don't know" (IDK) responses. We first extract the questions corresponding to the forget set, and then replace their answers with refusal templates (e.g., I don't know" or Unknown") sampled from a pre-collected file.

**Reproducibility** All datasets and splits are generated using deterministic scripts with fixed random seeds. The generated data are stored as parquet and jsonl files. We will release the dataset construction pipeline, split configurations, and evaluation scripts for reproducibility.

### A.2 Benchmark Comparison

Table 2 summarizes the differences between UnlearnBench and representative LLM unlearning benchmarks, including TOFU (Maini et al., 2024), MUSE (Shi et al., 2024), WMDP (Li et al., 2024), RWKU (Cao et al., 2024), BLUR (Hu et al., 2025b), and LUME (Ramakrishna et al., 2025). Existing benchmarks have made important progress in evaluating factual forgetting, verbatim memorization, privacy leakage, hazardous knowledge removal, and robustness under forget–retain overlap. However, most of them evaluate unlearning through direct single-hop queries or sentence completion, and therefore do not explicitly test whether forgotten knowledge can be reconstructed through compositional reasoning.

UnlearnBench differs from prior benchmarks in five key aspects. First, it provides substantially larger scale while preserving full synthetic control over the data generation process. Second, it introduces structured knowledge with explicit attribute and relation decomposition, allowing us to define forget and retain sets at the entity, attribute, and relation levels. Third, it evaluates both direct memorization and query-based extraction through sentence completion and QA tasks. Fourth, it supports multiple training regimes, including pretraining, continual pretraining, and supervised fine-tuning, which enables us to study how unlearning behavior changes across different stages of model development. Finally, UnlearnBench explicitly evaluates multi-hop and compositional recovery, where a model may fail direct recall but still recover forgotten information through relational reasoning chains.

This compositional setting is the main distinction from prior work. TOFU studies controlled fictitious author unlearning, but focuses on direct QA over synthetic profiles. MUSE evaluates six practical evaluations for unlearning, including verbatim memorization, knowledge memorization, privacy leakage, utility preservation, scalability, and sequential unlearning, but its removal targets are long-form books and news articles rather than structured relational knowledge. WMDP focuses on hazardous knowledge removal using multiple-choice questions in biosecurity, cybersecurity, and chemical security. RWKU provides a realistic benchmark over real-world famous people and includes adversarial probes, but its targets are not generated under a controllable compositional structure. BLUR emphasizes forget-retain overlap and relearning robustness, which is complementary to our focus, but it does not study attribute-level or relational unlearning. LUME expands unlearning evaluation to multiple tasks including creative documents, synthetic biographies, and real biographies, but still does not explicitly evaluate whether forgotten knowledge can be recovered through multi-hop reasoning. Therefore, UnlearnBench is, to our knowledge, the first benchmark that systematically studies fine-grained attribute-level and relation-level unlearning under controllable compositional knowledge settings.

**Dataset Scale** In line with prior unlearning benchmarks, we report dataset scale primarily by the number of entities and QA-style evaluation instances. **Note the reported number is of coarse estimation.** Specifically, TOFU (Maini et al., 2024) is defined as QA pairs built from autobiographies of 200 fictitious authors with around 4,000 rows; MUSE (Shi et al., 2024) focuses on news articles and novels and provides multiple QA-oriented subsets (e.g., KnowMem forget/retain QA splits) with 28,339 in public releases; WMDP (Li et al., 2024) is a multiple-choice QA benchmark (bio/cyber/chem) with 3,668 total questions; RWKU (Cao et al., 2024) is a real-world knowledge benchmark with 200 unlearning targets and 128,800 subsets in public releases; BLUR (Hu et al., 2025b) extends existing benchmarks with harder combined forget, retain and relearning queries (10,000); and LUME (Ramakrishna et al., 2025) explicitly covers three tasks including synthetic biographies with sensitive PII and real biographies, resulting 4,394 rows. Following this benchmark style, our UnlearnBench scale is computed as: 100,000 persons, 5 biographies per person, giving 500,000 biography documents, and 10 attribute-related QA items per person, giving 1,000,000 QA pairs in total. Therefore, the coarse scale can be reported as 1,500,000 (million-scale), plus the CoT version for the two-hop setting.

### A.3 Evaluation Metrics

Ideally, an unlearned model should behave as if it had never seen the forget set, while showing similar behavior to a retrained model on the retain set and test set. We evaluate target model training and unlearning with a set of complementary metrics that distinguish knowledge memorization from knowledge extraction. **QA generation accuracy (QA)** is a generation-based metric that measures whether the model can answer a question correctly under greedy decoding; accuracy is computed by

exact match between the generated answer and the gold answer. This serves as our main downstream metric for testing whether the model can successfully use learned knowledge in a task-oriented setting. **Biography first token accuracy (Bio_ft)** is a retrieval-oriented probing metric that evaluates whether target facts remain encoded in the model by measuring first-token accuracy on biography prompts, including both attribute values and relational targets. Compared with QA generation accuracy, Bio_ft more directly reflects the presence or absence of factual knowledge in the model parameters, while P reflects the model's ability to deploy that knowledge during generation.

To separate memorization from generalization, we report these metrics on both $\text{QA}_{\text{train}}$ and $\text{QA}_{\text{test}}$, where $\text{QA}_{\text{train}}$ contains entities seen during task-level supervision and $\text{QA}_{\text{test}}$ contains held-out entities. High $\text{QA}_{\text{train}}$ but low $\text{QA}_{\text{test}}$ indicates overfitting to supervised examples, whereas strong performance on both suggests successful knowledge generalization. In the unlearning setting, we further evaluate performance on forget and retain subsets. Effective unlearning should substantially reduce $\text{QA}_{\text{forget}}$, indicating removal or suppression of the targeted knowledge, while preserving high performance on the retain set, indicating minimal collateral damage to unrelated knowledge. Together, these metrics provide a unified view of whether CPT or PPT improves knowledge acquisition and whether unlearning selectively removes the intended information without broadly degrading the model.

**Bio_ft Motivation.** Bio_ft is included to measure verbatim memorization of biography facts, rather than QA-style retrievability. Following Allen-Zhu & Li (2023b), it probes whether the model assigns the correct first token to a biography continuation, which directly tests whether the corresponding fact is stored in the model. This is necessary because QA accuracy alone cannot distinguish absence of knowledge from failure of retrieval. For example, Table 3 shows that the Two-Stage (Bio Pretrain) model has Bio_ft = 96.67 but $QA_{\text{train}}$ = $QA_{\text{test}}$ = 0.00, meaning the biography facts are memorized but not elicitable through QA prompts. We use first-token accuracy instead of full-sequence exact match because attribute values are often multi-token and variable-length, so full-generation accuracy would confound memorization with decoding and tokenization artifacts. Bio_ft therefore serves as a targeted probe of memorized biography content, while QA evaluates whether that content can be retrieved in downstream question answering.

**Attribute-Level Memory NLL (AL-MNLL).** To evaluate attribute-level forgetting and retention, we use a length-normalized conditional negative log-likelihood metric on attribute values. For a biography context $x$ and target attribute string $a = (a_1, \ldots, a_m)$, with start token position $p$ (located by token-subsequence matching), we define

$$\text{nNLL}(a \mid x) = -\frac{1}{m} \sum_{t=1}^{m} \log P(a_t \mid x_{\leq p-1}, a_{<t}) \,. \tag{1}$$

Lower nNLL indicates stronger memory of the attribute value.

## B Target Model Training Protocols

A meaningful unlearning benchmark requires the target model to genuinely acquire the target knowledge while maintaining reliable extraction and generalization ability. Following the storage extraction distinction of Allen-Zhu & Li (2023b), we require target models to satisfy two properties before unlearning: 1) Knowledge storage: the model memorizes the target factual knowledge, measured by Bio_ft; 2) Knowledge extraction and generalization: the stored knowledge can be consistently extracted through question answering on both training and held-out entities. To study how different memorization regimes affect selective forgetting, we construct several target model training protocols spanning both single-hop factual knowledge and compositional two-hop relational reasoning.

**LLM Memorization Depends on the Training Paradigm** We study memorization and unlearning under mixed training, two-stage training, and multi-hop reasoning settings. Our results show that the training paradigm fundamentally changes how knowledge is stored and how difficult it is to unlearn. Under mixed training, models can jointly learn knowledge memorization and QA extraction skills (skills includes generalize to unseen QA.) in the single-hop setting. However, such capabilities fail to generalize to two-hop reasoning. In particular, two-hop reasoning requires explicit replay of compositional knowledge together with both direct QA supervision and chain-of-thought (CoT) QA

Table 6: Single-hop target model training regimes. Bio_ft denotes BIO first-token accuracy. $QA_{train}$ and $QA_{test}$ evaluate extraction on seen and held-out entities respectively. "T" refers to the template QA dataset.

| Models | Bio_ft | $QA_{train}$ | $QA_{test}$ | $QA_{train}$ (T) | $QA_{test}$ (T) |
|---|---|---|---|---|---|
| Qwen3-0.6B-Base | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Mixed Training | 100.00 | 99.99 | 94.73 | 86.64 | 80.02 |
| Two-Stage (BIO pretraining) | 96.67 | 0.00 | 0.00 | 0.00 | 0.00 |
| Two-Stage (LoRA finetuning) | 87.92 | 86.95 | 86.66 | 55.54 | 55.01 |
| Two-Stage (Full finetuning) | 97.10 | 99.85 | 93.64 | 74.20 | 69.21 |
| Llama3.2-1B Base | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Mixed Training | 99.81 | 99.99 | 92.70 | 66.15 | 60.90 |

supervision. In contrast, two-stage training separates knowledge memorization from downstream QA extraction. This separation significantly changes the behavior of unlearning methods.

**Pretraining Loss.** For pretraining and continual pretraining, given a token sequence $x = [x_1, ..., x_n]$, the autoregressive language modeling objective is $\mathcal{L}_{pre}(x; \theta) = -\sum_{i=1}^{n} \log P_\theta(x_i|x_{<i})$, where $x_{<i} := [x_1, \ldots, x_{i-1}]$ and $P_\theta$ denotes the next-token distribution parameterized by $\theta$.

**Fine-tuning Loss** For a prompt-response or question-answer pair $(x, y)$, the loss function on $y$ for fine-tuning is $\mathcal{L}(x, y; \theta) = \sum_{i=1}^{|y|} \ell(h_\theta(x, y_{<i}), y_i)$, where $\ell(\cdot)$ is the cross-entropy loss, and $h_\theta(x, y_{<i}) := \mathbb{P}(y_i|(x, y_{<i}); \theta)$ is the predicted probability of the token $y_i$ given by an LLM $M_\theta$ parameterized by $\theta$, with the input prompt $x$ and the already generated tokens $y_{<i} := [y_1, ..., y_{i-1}]$.

### B.1 Single-hop Memorization Regimes

Table 6 summarizes the single-hop target model training settings.

**Mixed training: knowledge memorization regime.** The first setting jointly trains BIO and QA data from scratch in a single continual pretraining stage. Under this setting, the model directly memorizes question-answer mappings together with the underlying biographies. As shown in Table 6, mixed training achieves strong memorization and generalization performance, reaching over 94% QA accuracy on held-out entities. The model behaves similarly to a domain-specialized memorization system, where factual associations are directly encoded into model parameters.

**Two-stage training: extraction-skill regime.** We next separate knowledge storage and extraction into two stages. The model is first pretrained only on biographies, then finetuned using LoRA QA supervision and full finetuning. Interestingly, BIO pretraining alone produces nearly perfect factual storage while yielding almost zero QA extraction accuracy. This demonstrates that **memorization alone does not guarantee extractability**. After LoRA finetuning, QA performance on unseen entities becomes nearly identical to training performance (86.95% vs. 86.66%). This suggests that LoRA finetuning primarily learns a transferable extraction skill rather than storing entity-specific knowledge. The model learns how to retrieve structured attributes from memorized representations instead of memorizing QA pairs independently.

### B.2 Two-hop Relational Memorization

Compared with single-hop factual extraction, compositional relational reasoning is substantially harder to acquire. We therefore conduct a systematic training study using GPT2-NC depth = 12 to understand which training conditions enable reliable two-hop reasoning generalization. Table 7 summarizes the major experimental findings.

**Naive mixed training fails to generalize.** Directly mixing biographies and two-hop CoT data during pretraining leads to both poor memorization and compositional generalization.

Table 7: GPT2 target-model training methods and held-out performance ($QA_{\text{test}}$).

| Method | Pipeline family | Task | $QA_{\text{test}}$ |
|---|---|---|---|
| GPT2 mixed training | 1-hop mixed (joint BIO + QA) | 1-hop QA | 96.06% |
| GPT2 mixed training | 2-hop mixed (joint BIO + CoT QA) | 2-hop CoT QA | 8.30% |
| GPT2 direct 2-hop LoRA | 2-hop two-stage (BIO pretrain + direct QA FT) | 2-hop CoT QA | 2.73% |
| GPT2 staged 2-hop CoT replay | 2-hop staged (strong BIO pretrain + direct replay + CoT replay) | 2-hop CoT QA | 97.92% |

Table 8: Two-hop target model training regimes.

| **Datasets** | **Bio** | **1-hop Direct** | | **1-hop Relation** | | **2-hop Relation (CoT)** | |
|---|---|---|---|---|---|---|---|
| **Models** | Bio_ft | $\text{QA}_{\text{train}}$ | $\text{QA}_{\text{test}}$ | $\text{QA}_{\text{train}}$ | $\text{QA}_{\text{test}}$ | $\text{QA}_{\text{train}}$ | $\text{QA}_{\text{test}}$ |
| Qwen3-0.6B-Base | 0.00 | 0.00 | 0.0 | 0.00 | 0.00 | 0.00 | 0.00 |
| Three-Stage (Bio pretrain) | 98.94 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Three-Stage (Direct replay) | 93.18 | 99.98 | 96.52 | 99.88 | 8.89 | 99.61 | 8.90 |
| Three-Stage (CoT) | 89.06 | 0.0 | 0.0 | 0.0 | 0.0 | 99.92 | 80.86 |

**Format consistency is critical for compositional transfer.** We observe that simply mixing direct relation QA with CoT reasoning does not improve transfer. The model fails to connect the intermediate relation resolution process with the final compositional reasoning objective.

**Two-hop reasoning requires both strong memory and structured extraction.** The final successful protocol combines three stages used in Section 3.3. Under this setting, the model reaches 80.86% two-hop QA accuracy on held-out entities as indicated in Table 8. Importantly, neither memorization nor reasoning supervision alone is sufficient; successful compositional extraction emerges only when both are jointly aligned.

### B.3 Retrained Oracle Models

For every unlearning setting, we additionally train a retrained oracle model using the same initialization, optimization schedule, and hyperparameters as the corresponding target model, while removing all forget examples from every training stage. For mixed-training protocols, retraining uses only retain BIO and retain QA data. For staged pipelines, both pretraining and finetuning stages are reconstructed using retain-only splits, or alternatively reuse the original stage-1 checkpoint while restricting later stages to retain-only supervision. These retrained models serve as the gold-standard lower bound for ideal selective forgetting.

## C Benchmarking Unlearning Methods

In this section, we formulate all the unlearning methods benchmarked in this paper.

### C.1 Optimization-based Baselines

We evaluate a representative set of finetuning-based unlearning objectives under a unified training/evaluation pipeline. Let $D_f$ and $D_r$ denote forget and retain sets, and let $\theta_t$ be the target-model parameters before unlearning. All methods start from the same target checkpoint and are tuned with matched optimization settings where applicable.

**Gradient Ascent(GA) (Maini et al., 2024)** GA is the direct baseline: maximize loss on forget samples only, so model likelihood on $D_f$ decreases.

$$L_{\text{GA}} = -\frac{1}{|D_f|} \sum_{(x_f, y_f) \in D_f} \mathcal{L}(x_f, y_f; \theta)$$

**Gradient Difference(GD) (Maini et al., 2024)** GD adds an explicit retain term to GA, balancing forget suppression and utility preservation in one objective.

$$L_{\text{GD}} = -\frac{1}{|D_f|}\sum_{(x_f,y_f)\in D_f}\mathcal{L}(x_f,y_f;\theta) + \frac{1}{|D_r|}\sum_{(x_r,y_r)\in D_r}\mathcal{L}(x_r,y_r;\theta)$$

**Negative Preference Optimization(NPO) (Zhang et al., 2024a)** NPO reformulates unlearning as preference-style suppression of the forget response $y_f$ relative to a frozen reference model $\pi_{\text{ref}}$. We report both NPO and the retain-regularized variant NPO-RT.

$$L_{\text{NPO}} = -\frac{2}{\beta}E_{D_f}\Big[\log\sigma\Big(-\beta log\frac{\pi_\theta(y_f \mid x_f)}{\pi_{ref}(y_f \mid x_f)}\Big)\Big]$$

$$L_{\text{NPO-RT}} = \frac{1}{|D_r|}\sum_{(x_r,y_r)\in D_r}\mathcal{L}(x_r,y_r;\theta) - \frac{2}{\beta}E_{D_f}\Big[\log\sigma\Big(-\beta log\frac{\pi_\theta(y_f \mid x_f)}{\pi_{ref}(y_f \mid x_f)}\Big)\Big]$$

**Direct preference optimization (DPO) (Rafailov et al., 2023)** DPO uses pairwise preferences on forget prompts. For each $x_f$, we construct a preferred refusal response $y_p$ and a dispreferred original response $y_f$, then optimize relative preference against $\pi_{\text{ref}}$.

$$L_{\text{DPO},\beta}(\theta) = -\frac{2}{\beta}E_{D_{pair}}\Big[\log\sigma\Big(\beta\log\frac{\pi_\theta(y_p \mid x_f)}{\pi_{ref}(y_p \mid x_f)} - \beta\log\frac{\pi_\theta(y_f \mid x_f)}{\pi_{ref}(y_f \mid x_f)}\Big)\Big]$$

where $\sigma(t) = \frac{1}{1+e^{-t}}$ is the sigmoid function, $\beta > 0$ is the inverse temperature, $\pi_\theta := \prod_{i=1}^{|y|} h_\theta(x, y_{<i})$ is the predicted probability of the response $y$ to prompt $x$ given by LLM $M_\theta$, $\pi_{ref}$ is the predicted probability given by reference model $M_{\theta_t}$.

**Representation Misdirection for Unlearning(RMU) (Li et al., 2024)** RMU pushes the hidden representation at layer $\ell$ towards a random vector and minimizes the squared difference between the unlearned model and the original model.

$$L_{\text{RMU}} = \frac{1}{|D_f|}\sum_{x\in D_f}\|M_\ell(x;\theta) - c\cdot u\|_2^2 + \frac{\alpha}{|D_r|}\sum_{x\in D_r}\|M_\ell(x;\theta) - M_\ell(x;\theta_t)\|_2^2$$

where $M_\ell(x;\theta)$ is the function that returns the hidden representation of LLM $M_\theta$ at a layer $\ell$, and $u$ is a fixed random unit vector sampled uniformly from $[0, 1)$.

### C.2 Localization & Representation-level Unlearning

**Improved PerTA (IPerTA).** We introduce an improved iterative PerTA method based on Cai et al. (2026). In each iteration, we first compute per-parameter Fisher importance scores and identify parameters whose importance exceeds a predefined threshold. These parameters are then reset to their pretrained values to suppress forget-related knowledge while preserving retain-relevant representations. After the reset step, we perform a repair stage consisting of 200 to 300 finetuning steps on retained BIO and QA data to recover collateral utility degradation. This iterative process is repeated with adaptive thresholds to progressively balance forgetting and retention.

## D More Experimental Results

### D.1 Single-hop Entity-level Unlearning

**Computational cost.** All experiments ran on a shared cluster of 8 A5000 GPUs (24 GB each).

Table 9: Unlearning performance under the **entity-level 1%** forget setting using the Qwen3-0.6B-Base **mixed training** model. $\text{Bio_ft}_{\text{F}}$, $\text{Bio_ft}_{\text{R}}$, and $\text{Bio_ft}_{\text{test}}$ denote first-token biography accuracy on the forget, retain, and test sets, respectively. This table extends the results presented in Table 4.

| **Methods** | $\text{QA}_{\text{forget}}$ | $\text{QA}_{\text{Retain}}$ | $\text{QA}_{\text{test}}$ | Gap | $\text{Bio_ft}_{\text{F}}$ | $\text{Bio_ft}_{\text{R}}$ | $\text{Bio_ft}_{\text{test}}$ |
|---|---|---|---|---|---|---|---|
| Target Model | 100.00 | 99.99 | 94.73 | -0.01 | 99.73 | 99.74 | 99.65 |
| Retrained Model | 1.00 | 98.67 | 89.00 | 97.67 | 15.81 | 95.20 | 92.52 |
| GA | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| GD | 0.00 | 0.02 | 0.01 | 0.02 | 7.82 | 20.24 | 18.72 |
| NPO | 99.97 | 99.99 | 94.70 | 0.02 | 99.77 | 99.75 | 99.65 |
| SimNPO | 53.68 | 53.06 | 47.72 | -0.62 | 98.36 | 98.14 | 97.53 |
| DPO | 100.00 | 99.99 | 94.74 | -0.01 | 99.73 | 99.74 | 99.64 |
| RMU | 99.98 | 99.99 | 94.72 | 0.01 | 99.75 | 99.76 | 99.66 |
| I PerTA | 82.32 | 87.76 | 65.78 | 5.44 | 66.05 | 72.04 | 62.18 |

Table 10: Unlearning performance under the **entity-level 10%** forget setting using the Qwen3-0.6B-Base **mixed training** model. $\text{Bio_ft}_{\text{F}}$, $\text{Bio_ft}_{\text{R}}$, and $\text{Bio_ft}_{\text{test}}$ denote first-token biography accuracy on the forget, retain, and test sets, respectively. This table extends the results presented in Table 4.

| **Methods** | $\text{QA}_{\text{forget}}$ | $\text{QA}_{\text{Retain}}$ | $\text{QA}_{\text{test}}$ | Gap | $\text{Bio_ft}_{\text{F}}$ | $\text{Bio_ft}_{\text{R}}$ | $\text{Bio_ft}_{\text{test}}$ |
|---|---|---|---|---|---|---|---|
| Target Model | 100.00 | 99.99 | 94.73 | -0.01 | 99.74 | 99.74 | 99.65 |
| Retrained Model | 1.01 | 99.99 | 93.18 | 98.98 | 14.73 | 99.84 | 99.76 |
| GA | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.0 | 0.00 |
| GD | 0.00 | 2.68 | 1.71 | 2.68 | 2.03 | 14.53 | 10.74 |
| NPO | 99.99 | 99.99 | 94.69 | 0.00 | 99.75 | 99.73 | 99.63 |
| SimNPO | 0.02 | 0.01 | 0.00 | -0.01 | 1.63 | 1.58 | 1.12 |
| DPO | 99.99 | 99.99 | 94.76 | 0.00 | 99.73 | 99.74 | 99.65 |
| RMU | 99.99 | 99.99 | 94.74 | 0.00 | 99.74 | 99.75 | 99.65 |
| IPerTA | 79.01 | 87.41 | 62.27 | 8.40 | 63.88 | 73.54 | 60.16 |

**Mixed training Model** Table 9 and Table 10 show the unlearning performance under mixed training CPT model. Biography-level forgetting is substantially harder than QA forgetting. Exact retraining can effectively suppress biography memorization on the forget set while largely preserving retain and test performance, reducing $\text{Bio_ft}_{\text{F}}$ from nearly 100% to around 15% under both the 1% and 10% forget settings. In contrast, most existing unlearning methods either fail to remove biography knowledge (NPO, DPO, RMU) or collapse the overall memorization ability globally (GA, GD, SimNPO). IPerTA achieves partial biography forgetting, but retain and test biography performance also significantly degrade. These results suggest that memorized biography knowledge also remains highly entangled under mixed training, making selective biography unlearning challenging for existing methods.

**Two Stage Training Model** Table 11 show the unlearning performance under two stage training model. Biography memorization becomes even harder to selectively remove. Exact retraining still achieves strong biography forgetting, reducing $\text{Bio_ft}_{\text{F}}$ from 88.95% to 17.61%, while largely preserving retain and test performance. In contrast, nearly all existing unlearning methods fail to suppress biography memorization, with Bio_ft remaining close to the target model across forget, retain, and test splits. Although SimNPO and IPerTA slightly reduce biography accuracy globally, they do not create a meaningful gap between forget and retain performance.

### D.2 Single-hop Attribute-level Unlearning

**Single-hop: RMU performs best overall under the attribute-level all setting, while other methods partially separate public and private attributes but struggle with fine-grained selective**

Table 11: Unlearning performance under the **entity-level 1%** forget setting using the Qwen3-0.6B-Base **Two Stage** model. Bio_ft$_F$, Bio_ft$_R$, and Bio_ft$_{test}$ denote first-token biography accuracy on the forget, retain, and test sets, respectively. This table extends the results presented in Table 4.

| **Methods** | QA$_{forget}$ | QA$_{Retain}$ | QA$_{test}$ | Gap | Bio_ft$_F$ | Bio_ft$_R$ | Bio_ft$_{test}$ |
|---|---|---|---|---|---|---|---|
| Target Model | 87.30 | 86.94 | 88.67 | -0.36 | 88.95 | 87.92 | 87.88 |
| Retrained Model | 0.80 | 91.53 | 91.15 | 90.73 | 17.61 | 81.58 | 81.35 |
| GA | 87.20 | 86.60 | 86.39 | -0.60 | 87.59 | 87.74 | 87.67 |
| GD | 87.32 | 86.90 | 86.64 | -0.42 | 87.84 | 87.57 | 87.85 |
| NPO | 77.12 | 76.26 | 75.99 | -0.86 | 88.40 | 88.46 | 88.46 |
| SimNPO | 48.42 | 48.24 | 47.98 | -0.18 | 84.95 | 85.04 | 85.02 |
| DPO | 87.50 | 86.98 | 86.72 | -0.62 | 87.84 | 87.96 | 87.85 |
| RMU | 87.32 | 86.90 | 86.64 | -0.42 | 87.93 | 88.03 | 87.94 |
| IPerTA | 87.80 | 88.00 | 87.69 | 0.20 | 82.84 | 82.86 | 82.92 |

**unlearning (Partial).** For the attribute-level all unlearning setting, RMU shows the best performance under both mixed and two-stage training settings (Table 12 and Table 13). Methods like GD and iterative PerTA can distinguish private attributes from public ones to some extent, with around a 13% gap. GA shows a 15% gap under the two-stage setting. For the mixed training, test-time private attribute accuracy of the retrained model also degrades because private attributes are never observed during training (Table 12). Consequently, the model fails to learn how to extract private attribute related knowledge. For the two-stage training, the unlearning methods can achieve unlearning even on private attributes, which may indicate that knowledge is more organized by attribute type, not by person.

We further evaluate attribute-level unlearning under the partial forget-people setting, where only the private attributes of 1,000 forget people are removed. Table 14 shows that GA, GD, SimNPO, and Iterative PerTA achieve partial forgetting under the mixed training model, as reflected by the noticeable gap between private QA and public QA performance. Retraining further achieves near-optimal unlearning performance. However, private QA performance also degrades across the retain and test sets, indicating that these unlearning methods tend to remove the same attribute globally rather than selectively forgetting it for specific individuals. A similar phenomenon is observed in the two-stage training setting, including for the retrained model (Table 15). Nevertheless, unlearning methods still maintain relatively high performance on the forget set. This may be because only a small subset of private attributes is removed, while the model mainly learns a generalizable QA extraction skill during Stage 2, allowing the overall QA capability to remain largely preserved.

**Beyond Qwen Model** Table 16 and Table 17 show that the overall unlearning behavior on Llama3.2-1B is highly consistent with the Qwen-based models. Under both entity-level and attribute-level settings, existing methods exhibit similar failure modes: methods such as DPO and RMU largely preserve the original memorization ability and fail to achieve effective forgetting, while GA, GD, and NPO tend to globally damage model utility instead of selectively removing target knowledge. SimNPO again shows unstable behavior, partially reducing memorization while also harming generalization performance. The attribute-level results further show that selective forgetting of private attributes remains challenging. Overall, these results suggest that the main observations of UnlearnBench are not specific to Qwen models and can generalize across different model families and architectures.

### D.3 Single-hop Two-Stage Unlearning (Detailed)

Table 18 further analyzes unlearning under the two-stage training setting. The results show that forgetting only the second-stage QA fine-tuning is insufficient for effective unlearning. When retraining only the QA stage while reusing the original stage 1 biography model (Retrain Original s1), the model still preserves strong QA and biography memorization ability on the forget set. In contrast, retraining both stages jointly using mixed Bio+QA forget data (Retrain s1) successfully removes both QA and biography memorization, reducing QA$_{forget}$ to 0.33 and Bio_ft$_F$ to 13.98 while preserving

Table 12: Unlearning performance on UnlearnBench under the **attribute-level all** unlearning setting using the Qwen3-0.6B-base **mixed training** model. $QA_{F,priv}$ denotes performance on all private attributes (birth date and birth city) in the forget set (50,000 entities). In this setting, $QA_{R,pub}$ corresponds to the retain public attributes. The Gap measures the difference between $QA_{R,pub}$ and $QA_{F,priv}$. $Bio_nll_F$ denotes the NLL on forget biographies, where higher values indicate better forgetting, while lower $Bio_nll_R$ indicates stronger memorization of the retain biographies. This table shows the detailed number results presented in Figure 1.

| **Methods** | $QA_{F,priv}$ | $QA_{R,pub}$ | Gap | $QA_{T,priv}$ | $QA_{T,pub}$ | $Bio_nll_F$ | $Bio_nll_R$ |
|---|---|---|---|---|---|---|---|
| Target Model | 99.99 | 99.99 | 0.00 | 94.54 | 94.82 | 4.69 | 3.89 |
| Retrained Model | 0.06 | 99.99 | 99.93 | 0.04 | 90.17 | 17.20 | 0.55 |
| GA | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1434.20 | 1296.62 |
| GD | 0.00 | 14.05 | 14.05 | 0.00 | 8.09 | 1134.70 | 368.62 |
| NPO | 92.89 | 94.52 | 1.63 | 70.04 | 67.00 | 5.10 | 3.94 |
| SimNPO | 26.40 | 82.40 | 56.00 | 20.80 | 75.04 | 5.83 | 4.83 |
| DPO | 99.99 | 99.99 | 0.00 | 94.56 | 94.84 | 4.66 | 3.83 |
| RMU | 5.52 | 99.99 | **94.47** | 4.58 | 93.92 | 5.04 | 3.94 |
| IPerTA | 0.00 | 12.95 | 12.95 | 0.00 | 4.76 | 2515.76 | 2247.65 |

Table 13: Unlearning performance on UnlearnBench under the **attribute-level all** unlearning setting using the Qwen3-0.6B-base **Two Stage training** model. $QA_{F,priv}$ denotes performance on all private attributes (birth date and birth city) in the forget set (50,000 entities). In this setting, $QA_{R,pub}$ corresponds to the retain public attributes. The Gap measures the difference between $QA_{R,pub}$ and $QA_{F,priv}$. $Bio_nll_F$ denotes the NLL on forget biographies, where higher values indicate better forgetting, while lower $Bio_nll_R$ indicates stronger memorization of the retain biographies. "Original S1" denotes that the retrained model is initialized from the original Stage 1 Bio pretrained model, which still contains all memorized biography knowledge. This table shows the detailed number results presented in Figure 1.

| **Methods** | $QA_{F,priv}$ | $QA_{R,pub}$ | Gap | $QA_{T,priv}$ | $QA_{T,pub}$ | $Bio_nll_F$ | $Bio_nll_R$ |
|---|---|---|---|---|---|---|---|
| Target Model | 83.98 | 88.43 | 4.45 | 83.67 | 88.17 | 4.35 | 3.64 |
| Retrained Model (Original S1) | 0.0 | 89.00 | 89.00 | 0.0 | 89.00 | 6.75 | 3.93 |
| GA | 11.63 | 27.44 | 15.81 | 11.67 | 27.38 | 6.81 | 5.34 |
| GD | 0.0 | 0.10 | 0.10 | 0.0 | 0.11 | 243.61 | 101.07 |
| NPO | 62.07 | 68.63 | 6.56 | 61.37 | 68.41 | 4.54 | 3.45 |
| SimNPO | 0.0 | 1.33 | 1.33 | 0.0 | 1.28 | 17.69 | 9.61 |
| DPO | 84.06 | 88.57 | 4.51 | 83.79 | 88.33 | 4.34 | 3.63 |
| RMU | 0.03 | 88.42 | 88.39 | 0.03 | 88.16 | 49.88 | 4.94 |
| Iterative PerTA | 63.01 | 90.30 | 27.29 | 62.64 | 89.83 | 3.98 | 3.44 |

Table 14: Unlearning performance on UnlearnBench under the **attribute-level 1%** unlearning setting using the Qwen3-0.6B-base **Mixed training** model. $QA_{F,priv}$ denotes performance on private attributes (birth date and birth city) in the selected forget set (1000 entities). In this setting, $QA_{F,pub}$, $QA_{R,priv}$, $QA_{R,pub}$ corresponds to the retain attributes. $Bio_nll_F$ denotes the NLL on forget biographies, where higher values indicate better forgetting, while lower $Bio_nll_R$ indicates stronger memorization of the retain biographies. This table shows the detailed number results presented in Figure 1.

| **Methods** | $QA_{F,priv}$ | $QA_{F,pub}$ | $QA_{R,priv}$ | $QA_{R,pub}$ | $QA_{T,priv}$ | $QA_{T,pub}$ | $Bio_nll_F$ | $Bio_nll_R$ (↓) |
|---|---|---|---|---|---|---|---|---|
| Target Model | 100.00 | 100.00 | 99.99 | 99.99 | 94.54 | 94.82 | 4.37 | 4.11 |
| Retrained Model | 0.95 | 100.00 | 99.99 | 99.99 | 91.75 | 93.51 | 16.84 | 0.00 |
| GA | 2.55 | 37.25 | 2.25 | 37.34 | 0.87 | 25.63 | 16.47 | 31.43 |
| GD | 0.00 | 36.20 | 0.01 | 36.62 | 0.00 | 20.61 | 635.43 | 223.07 |
| NPO | 0.00 | 0.07 | 0.00 | 0.12 | 0.00 | 0.60 | 118.94 | 110.65 |
| SimNPO | 8.50 | 74.75 | 7.39 | 74.58 | 5.31 | 66.57 | 8.10 | 9.48 |
| DPO | 100.00 | 100.00 | 99.99 | 99.99 | 94.54 | 94.83 | 4.56 | 4.11 |
| RMU | 100.00 | 100.00 | 99.99 | 99.99 | 94.54 | 94.81 | 4.64 | 4.32 |
| IPerTA | 54.55 | 78.03 | 65.02 | 78.25 | 34.90 | 49.91 | 642.42 | 476.24 |

Table 15: Unlearning performance on UnlearnBench under the **attribute-level 1%** unlearning setting using the Qwen3-0.6B-base **Two Stage training** model. $QA_{F,priv}$ denotes performance on private attributes (birth date and birth city) in the selected forget set (1000 entities). In this setting, $QA_{F,pub}$, $QA_{R,priv}$, $QA_{R,pub}$ corresponds to the retain attributes. $Bio_nll_F$ denotes the NLL on forget biographies, where higher values indicate better forgetting, while lower $Bio_nll_R$ indicates stronger memorization of the retain biographies. This table shows the detailed number results presented in Figure 1.

| **Methods** | $QA_{F,priv}$ | $QA_{F,pub}$ | $QA_{R,priv}$ | $QA_{R,pub}$ | $QA_{T,priv}$ | $QA_{T,pub}$ | $Bio_nll_F$ | $Bio_nll_R$ (↓) |
|---|---|---|---|---|---|---|---|---|
| Target Model | 83.90 | 89.05 | 83.96 | 88.41 | 83.67 | 88.17 | 4.12 | 3.32 |
| GA | 84.00 | 89.10 | 83.95 | 88.43 | 83.65 | 88.16 | 4.54 | 3.22 |
| GD | 83.30 | 88.60 | 83.17 | 87.81 | 82.85 | 87.48 | 4.53 | 3.30 |
| NPO | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 48.21 | 44.45 |
| SimNPO | 83.00 | 88.05 | 83.08 | 87.54 | 82.72 | 87.35 | 4.94 | 3.37 |
| DPO | 84.25 | 89.18 | 84.03 | 88.50 | 83.73 | 88.27 | 4.46 | 3.20 |
| RMU | 84.10 | 89.08 | 83.91 | 88.46 | 83.64 | 88.19 | 4.51 | 3.24 |
| IPerTA | 83.00 | 89.92 | 83.17 | 88.49 | 82.43 | 89.02 | 4.53 | 3.36 |

Table 16: Unlearning performance under the **entity-level 1%** forget setting using the Llama3.2-1B **Mixed** model. $Bio_ft_F$, $Bio_ft_R$, and $Bio_ft_{test}$ denote first-token biography accuracy on the forget, retain, and test sets, respectively.

| **Methods** | $QA_{forget}$ | $QA_{Retain}$ | $QA_{test}$ | Gap |
|---|---|---|---|---|
| Target Model | 99.99 | 99.98 | 92.70 | -0.01 |
| GA | 3.75 | 4.33 | 1.40 | 0.58 |
| GD | 0.00 | 0.00 | 0.00 | 0.00 |
| NPO | 0.13 | 0.03 | 0.02 | -0.10 |
| SimNPO | 38.75 | 39.43 | 32.45 | 0.68 |
| DPO | 99.97 | 99.97 | 92.93 | 0.00 |
| RMU | 99.92 | 99.98 | 93.03 | 0.06 |

Table 17: Unlearning performance on UnlearnBench under the **attribute-level 1%** unlearning setting using the Llama3.2-1B **Mixed training** model. $QA_{F,priv}$ denotes performance on private attributes (birth date and birth city) in the selected forget set (1000 entities). In this setting, $QA_{F,pub}$, $QA_{R,priv}$, $QA_{R,pub}$ corresponds to the retain attributes.

| **Methods** | $QA_{F,priv}$ | $QA_{F,pub}$ | $QA_{R,priv}$ | $QA_{R,pub}$ | $QA_{T,priv}$ | $QA_{T,pub}$ |
|---|---|---|---|---|---|---|
| Target Model | 100.00 | 100.00 | 99.99 | 100.00 | 95.31 | 91.27 |
| GA | 72.45 | 99.93 | 72.65 | 99.91 | 60.20 | 90.26 |
| GD | 0.00 | 10.12 | 0.00 | 10.09 | 0.00 | 0.07 |
| NPO | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| SimNPO | 89.90 | 99.90 | 89.90 | 99.91 | 0.00 | 0.00 |
| DPO | 99.95 | 99.97 | 99.97 | 99.97 | 93.93 | 92.07 |
| RMU | 99.95 | 99.93 | 99.97 | 100.00 | 93.85 | 92.26 |

Table 18: Analysis of unlearning under the two-stage training setting, where the target model is first pretrained on biography data and then fully fine-tuned on QA data. We study different forget sources, including Bio-only, QA-only, and mixed Bio+QA forgetting. "Retrain (Original s1)" denotes retraining only the second-stage QA fine-tuning model while reusing the original stage 1 biography model, whereas "Retrain (Retrain s1)" retrains both stages jointly.

| Method | $QA_{forget}$ | $QA_{Retain}$ | $QA_{Test}$ | Gap | $Bio_ft_F$ | $Bio_ft_R$ |
|---|---|---|---|---|---|---|
| Target | 99.88 | 99.88 | 93.67 | 0.00 | 97.49 | 97.03 |
| **Bio-Only** | | | | | | |
| GD (780) | 89.72 | 89.98 | 83.18 | 0.26 | 93.82 | 93.62 |
| SimNPO (780) | 99.50 | 99.38 | 92.42 | 0.12 | 96.13 | 95.21 |
| RMU (780) | 99.88 | 99.88 | 93.33 | 0.00 | 97.31 | 97.67 |
| **QA-Only** | | | | | | |
| Retrain (Original s1) | 95.33 | 99.33 | 94.67 | 4.00 | 94.03 | 96.41 |
| GD (3750) | 68.67 | 67.67 | 67.00 | -1.00 | 96.29 | 94.33 |
| SimNPO (935) | 99.88 | 99.88 | 93.00 | 0.00 | 96.97 | 96.33 |
| RMU (935) | 99.88 | 99.88 | 93.33 | 0.00 | 96.96 | 97.00 |
| **Bio+QA Mix** | | | | | | |
| Retrain (Retrain s1) | 0.33 | 99.33 | 98.33 | 99.00 | 13.98 | 96.89 |
| GD (1000) | 44.03 | 45.85 | 41.30 | 1.82 | 88.09 | 88.01 |
| GD (1715) | 34.20 | 35.15 | 32.50 | 0.95 | 81.48 | 82.21 |
| SimNPO (1000) | 99.68 | 99.63 | 92.80 | -0.05 | 96.54 | 95.71 |
| SimNPO (1715) | 99.72 | 99.62 | 92.82 | -0.10 | 96.44 | 95.71 |
| RMU(1000) | 99.87 | 99.87 | 93.65 | 0.00 | 97.56 | 97.10 |
| RMU(1715) | 99.85 | 99.87 | 93.67 | 0.02 | 97.56 | 97.08 |

retain performance. Similar fail trends are observed for existing unlearning methods: using only Bio or QA forget data cannot effectively remove the memorized knowledge, suggesting that the knowledge learned during biography pretraining and QA fine-tuning remains highly connected. These results indicate that under two-stage training, effective unlearning requires removing knowledge from both memorization stages simultaneously.

Alternatively, an interesting future direction is to design unlearning methods with dynamic thresholds or adaptive objectives that jointly monitor biography and QA losses during training. Since memorized knowledge can be distributed across different training stages and representations, simply applying a fixed forgetting objective may be insufficient to fully reverse the memorization process. A dynamic mechanism that progressively balances biography forgetting and QA forgetting could potentially remove the target knowledge more effectively while preserving general utility. This may provide a more principled way to reverse multi-stage memorization in LLMs.

### D.4 Two-hop Entity-Level Unlearning

Table 19 shows the entity-level unlearning results on the two-hop setting using the stage 2 direct-reply model. Although the target model achieves poor generalization on the held-out test set, it strongly memorizes the training distribution, allowing us to focus on the forgetting and retain behavior directly. Similar to the single-hop setting, methods such as DPO and RMU almost completely preserve the original memorization ability under both direct and CoT evaluations, indicating little effective unlearning. GA and GD can partially suppress memorization under the direct two-hop setting, but they also significantly damage retain performance, suggesting global degradation instead of selective forgetting. Interestingly, the CoT evaluation appears substantially harder to unlearn: several methods that reduce direct QA memorization still retain strong performance under CoT prompting. This suggests that CoT decomposition may provide an alternative retrieval pathway for memorized relational knowledge, making two-hop unlearning more difficult under reasoning-based evaluation.

Table 19: Entity-level unlearning results on two-hop Direct QA and CoT QA data using the stage-2 direct-reply fine-tuned model as the target model. The model achieves strong memorization on the training distribution but weak generalization on held-out test data. Therefore, we mainly focus on the forgetting and retain behavior rather than test set generalization. The evaluation is conducted on a randomly sampled subset of 50 identities for efficiency.

| | **Two-hop Direct QA** | | | | | **Two-hop CoT QA** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Method | $QA_{forget}$ | $QA_{retain}$ | $QA_{test}$ | $Bio_ft_F$ | $Bio_ft_R$ | $QA_{forget}$ | $QA_{retain}$ | $QA_{test}$ | $Bio_ft_F$ | $Bio_ft_R$ |
| Target | 99.67 | 99.58 | 8.90 | 91.67 | 91.67 | 99.67 | 99.58 | 8.90 | 91.67 | 91.67 |
| GA | 4.33 | 5.67 | 1.17 | 93.00 | 94.33 | 89.08 | 89.33 | 7.83 | 93.33 | 95.67 |
| GD | 0.42 | 0.83 | 0.42 | 76.67 | 77.67 | 35.75 | 38.75 | 3.75 | 84.00 | 83.33 |
| NPO | 0.33 | 0.58 | 0.08 | 2.67 | 3.33 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| SimNPO | 66.92 | 68.17 | 5.83 | 94.33 | 94.67 | 86.58 | 87.50 | 7.92 | 91.33 | 94.00 |
| DPO | 99.67 | 99.58 | 8.83 | 91.67 | 91.67 | 99.67 | 99.58 | 8.83 | 92.00 | 92.00 |
| RMU | 99.67 | 99.58 | 8.83 | 92.00 | 91.67 | 99.67 | 99.58 | 8.75 | 91.67 | 91.67 |

Table 20: Two-hop relation-level unlearning results targeting neighbor-related knowledge under both direct QA and CoT QA evaluation settings. $QA_{F,neighbor}$ denotes questions involving the target neighbor relationship selected for forgetting, while $QA_{F,other}$ refers to other non-target relationships associated with the same forget entities.

| | **Two-hop Direct QA** | | | | **Two-hop CoT QA** | | | |
|---|---|---|---|---|---|---|---|---|
| Method | $QA_{F,\ neighbor}$ | $QA_{F,\ other}$ | $QA_{retain}$ | $QA_{test}$ | $QA_{F,\ neighbor}$ | $QA_{F,\ other}$ | $QA_{retain}$ | $QA_{test}$ |
| Target | 100.00 | 99.92 | 99.92 | 80.86 | 100.00 | 99.92 | 99.92 | 80.86 |
| GA | 81.52 | 81.98 | 81.72 | 56.54 | 98.98 | 99.04 | 98.92 | 71.06 |
| GD | 42.52 | 78.32 | 70.17 | 36.31 | 0.00 | 0.00 | 0.00 | 0.00 |
| NPO | 71.82 | 73.29 | 73.28 | 44.29 | 0.52 | 2.10 | 1.62 | 0.26 |
| SimNPO | 36.58 | 36.61 | 36.52 | 16.51 | 0.00 | 0.00 | 0.00 | 0.00 |
| DPO | 100 | 99.93 | 99.92 | 81.01 | 100 | 99.92 | 99.92 | 80.97 |
| RMU | 100 | 99.93 | 99.92 | 80.75 | 100 | 99.93 | 99.92 | 80.79 |
| PerTA* | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| IPerTA | 0.58 | 0.66 | 0.66 | 0.67 | 1.12 | 2.63 | 2.42 | 1.09 |

### D.5 Two-hop Attribute-level Unlearning

Table 20 and Table 21 present fine-grained two-hop unlearning results under relation-level and private attribute level settings, respectively. We evaluate both direct QA and CoT QA formats to study whether existing unlearning methods can selectively remove relational or private knowledge while preserving other associated information.

These four tables extend Table 22, Table 23, Table 24 and Table 25 by additionally reporting biography-level memorization performance. Overall, the biography results show similar trends to the QA results. Methods such as DPO and RMU largely preserve the original biography memorization ability, indicating little effective forgetting. In contrast, aggressive forgetting methods such as GD, PerTA, and Improved PerTA substantially reduce biography memorization, but often collapse both forget and retain performance simultaneously, suggesting global memorization destruction rather than selective unlearning. Compared with direct QA evaluation, the CoT setting generally leads to stronger degradation of biography memorization, further indicating that reasoning-based supervision and relational retrieval are highly entangled with the underlying biography representations.

### D.6 Unlearn Both Bio and QA dataset

Table 26 and Table 27 present the unlearning results when jointly forgetting both QA and biography data. For BIO first-token evaluation, we randomly sample 1000 identities for efficiency. Under the entity-level setting, exact retraining remains the only method that effectively removes both QA and biography memorization while preserving retain performance. Most existing methods either fail to forget the target knowledge (e.g., DPO and RMU) or collapse the entire model utility globally (e.g.,

Table 21: Two-hop attribute-level unlearning results targeting private attributes (birth date and birth city) under both direct QA and CoT QA evaluation settings. $QA_{F,priv}$ denotes questions involving the target private attributes selected for forgetting, while $QA_{F,other}$ refers to the remaining non-target attributes associated with the same forget entities.

| | **Direct 2-hop** | | | | **CoT 2-hop** | | | |
|---|---|---|---|---|---|---|---|---|
| Method | $QA_{F,priv}$ | $QA_{F,other}$ | $QA_{retain}$ | $QA_{test}$ | $QA_{F,priv}$ | $QA_{F,other}$ | $QA_{retain}$ | $QA_{test}$ |
| Target | 99.97 | 99.92 | 99.92 | 80.86 | 99.97 | 99.92 | 99.92 | 80.86 |
| GA | 19.41 | 72.38 | 54.34 | 33.28 | 14.28 | 58.84 | 44.20 | 19.45 |
| GD | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| NPO | 61.49 | 96.46 | 84.58 | 53.78 | 22.26 | 48.87 | 40.40 | 10.61 |
| SimNPO | 36.72 | 91.51 | 73.19 | 52.41 | 0.00 | 0.00 | 0.00 | 0.00 |
| DPO | 99.97 | 99.92 | 99.92 | 81.00 | 99.97 | 99.92 | 99.92 | 80.91 |
| RMU | 99.99 | 99.92 | 99.92 | 80.80 | 99.97 | 99.92 | 99.92 | 80.88 |
| PerTA* | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Iterative PerTA | 0.20 | 1.41 | 0.96 | 0.98 | 15.50 | 14.32 | 16.04 | 3.65 |

Table 22: Extension of Table 20 with additional biography-level memorization results under the direct two-hop QA setting. $Bio_ft_F$ and $Bio_ft_R$ denote biography first-token accuracy on the forget and retain sets, respectively.

| Method | $QA_{F,\ neighbor}$ | $QA_{F,\ other}$ | $QA_{retain}$ | $QA_{test}$ | $Bio_ft_F$ | $Bio_ft_R$ |
|---|---|---|---|---|---|---|
| Target | 100.00 | 99.92 | 99.92 | 80.86 | 89.47 | 90.15 |
| GA | 81.52 | 81.98 | 81.72 | 56.54 | 90.00 | 91.36 |
| GD | 42.52 | 78.32 | 70.17 | 36.31 | 66.84 | 65.22 |
| NPO | 71.82 | 73.29 | 73.28 | 44.29 | 90.00 | 90.46 |
| SimNPO | 36.58 | 36.61 | 36.52 | 16.51 | 90.00 | 89.73 |
| DPO | 100 | 99.93 | 99.92 | 81.01 | 89.47 | 90.19 |
| RMU | 100 | 99.93 | 99.92 | 80.75 | 88.95 | 89.98 |
| PerTA | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.68 |
| IPerTA | 0.58 | 0.66 | 0.66 | 0.67 | 0.53 | 0.54 |

GA, GD, SimNPO, and PerTA variants). Under the attribute-level setting, RMU achieves partial separation between private and public attributes, while NPO and SimNPO show unstable behavior across different attributes and BIO metrics. The PerTA variants strongly suppress memorization but also destroy retain performance.

### D.7 Unlearning Process Analysis

We analyze the training dynamics of different unlearning methods under the 1-hop mixed-training setting. Across most methods, forget and retain performance decrease together rather than separating cleanly, indicating that the target and non-target knowledge are highly entangled. We further observe a clear **cliff phenomenon**: instead of gradually forgetting the target knowledge, many methods maintain high performance for a long period and then suddenly collapse to near zero on both forget and retain sets. This behavior is particularly evident for GA and NPO. In contrast, GD shows smoother degradation, but forget and retain performance still decrease at similar rates, indicating global knowledge destruction rather than selective forgetting. Overall, we do not observe a stable intermediate stage where the model forgets the target knowledge while preserving retain knowledge. These results suggest that existing unlearning methods struggle to achieve a balanced trade-off between forgetting and retention under our challenging benchmarks.

### D.8 Extend to Real-World Data

To evaluate whether the observations from our controlled synthetic benchmark generalize to real-world data, we further apply our training and evaluation pipeline to the MUSE dataset (Shi et al.,

Table 23: Extension of Table 20 with additional biography-level memorization results under the CoT two-hop QA setting. $\text{Bio_ft}_\text{F}$ and $\text{Bio_ft}_\text{R}$ denote biography first-token accuracy on the forget and retain sets, respectively.

| Method | $\text{QA}_{\text{F, neighbor}}$ | $\text{QA}_{\text{F, other}}$ | $\text{QA}_{\text{retain}}$ | $\text{QA}_{\text{test}}$ | $\text{Bio_ft}_\text{F}$ | $\text{Bio_ft}_\text{R}$ |
|---|---|---|---|---|---|---|
| Target | 100.00 | 99.92 | 99.92 | 80.86 | 89.47 | 90.15 |
| GA | 98.98 | 99.04 | 98.92 | 71.06 | 92.63 | 92.68 |
| GD | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| NPO | 0.52 | 2.10 | 1.62 | 0.26 | 62.63 | 56.50 |
| SimNPO | 0.00 | 0.00 | 0.00 | 0.00 | 4.74 | 4.39 |
| DPO | 100 | 99.92 | 99.92 | 80.97 | 89.47 | 89.81 |
| RMU | 100 | 99.93 | 99.92 | 80.79 | 88.42 | 90.40 |
| PerTA | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.54 |
| IPerTA | 1.12 | 2.63 | 2.42 | 1.09 | 1.05 | 0.69 |

Table 24: Extension of Table 21 with additional biography-level memorization results under the direct two-hop QA setting. $\text{Bio_ft}_\text{F}$ and $\text{Bio_ft}_\text{R}$ denote biography first-token accuracy on the forget and retain sets, respectively.

| Method | $\text{QA}_{\text{F,priv}}$ | $\text{QA}_{\text{F,other}}$ | $\text{QA}_{\text{retain}}$ | $\text{QA}_{\text{test}}$ | $\text{Bio_ft}_\text{R}$ | $\text{Bio_ft}_\text{F}$ |
|---|---|---|---|---|---|---|
| Target | 99.97 | 99.92 | 99.92 | 80.86 | 89.47 | 90.15 |
| GA | 19.41 | 72.38 | 54.34 | 33.28 | 91.58 | 91.43 |
| GD | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 |
| NPO | 61.49 | 96.46 | 84.58 | 53.78 | 90.53 | 91.01 |
| SimNPO | 36.72 | 91.51 | 73.19 | 52.41 | 91.58 | 92.18 |
| DPO | 99.97 | 99.92 | 99.92 | 81.00 | 89.47 | 90.23 |
| RMU | 99.99 | 99.92 | 99.92 | 80.80 | 88.42 | 90.04 |
| PerTA | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| IPerTA | 0.20 | 1.41 | 0.96 | 0.98 | 0.00 | 0.06 |

2024). We use the original unlearning setting in the original muse paper with 100 forget QA samples and 100 retain QA samples, and train target models for 5 epochs under different memorization strengths. Table 32 presents the unlearning results. We observe several trends that are consistent with our findings in the synthetic setting.

**Unlearning performance strongly depends on memorization strength.** Under high memorization, retraining substantially reduces $\text{QA}_{\text{forget}}$ (from 43% to 1%) while preserving utility, whereas most approximate methods (e.g., NPO, SimNPO) fail to reduce forget accuracy. Under low memorization, optimization-based methods (e.g., GD, SimNPO) can achieve near-zero forgetting, but at the cost of severe utility collapse ($\text{QA}_{\text{retain}}$ drops to 0%), indicating destructive unlearning behavior.

**The trade-off between forgetting and utility persists.** Across both memorization regimes, methods that aggressively reduce $\text{QA}_{\text{forget}}$ tend to significantly degrade $\text{QA}_{\text{retain}}$ and Bio_ft, while conservative methods preserve utility but fail to forget. This mirrors the failure modes observed in our main experiments.

**Retraining remains the most reliable oracle.** Consistent with the synthetic benchmark, retraining achieves the best balance between forgetting and retention, further validating that the desired unlearning solution exists but is not captured by current approximate methods.

Overall, these results demonstrate that the key phenomena identified in UnlearnBench, namely the dependence on memorization strength, the forget retain trade-off, and the gap between retraining and approximate unlearning, also emerge in real-world datasets. This suggests that our benchmark captures fundamental challenges of LLM unlearning beyond controlled synthetic settings.

Table 25: Extension of Table 21 with additional biography-level memorization results under the CoT two-hop QA setting. Bio_ft$_F$ and Bio_ft$_R$ denote biography first-token accuracy on the forget and retain sets, respectively.

| Method | $QA_{F,priv}$ | $QA_{F,other}$ | $QA_{retain}$ | $QA_{test}$ | Bio_ft$_R$ | Bio_ft$_F$ |
|---|---|---|---|---|---|---|
| Target | 99.97 | 99.92 | 99.92 | 80.86 | 89.47 | 90.15 |
| GA | 14.28 | 58.84 | 44.20 | 19.45 | 86.84 | 90.22 |
| GD | 0.00 | 0.00 | 0.00 | 0.00 | 8.42 | 10.66 |
| NPO | 22.26 | 48.87 | 40.40 | 10.61 | 67.37 | 63.22 |
| SimNPO | 0.00 | 0.00 | 0.00 | 0.00 | 26.32 | 24.63 |
| DPO | 99.97 | 99.92 | 99.92 | 80.91 | 90.00 | 90.04 |
| RMU | 99.97 | 99.92 | 99.92 | 80.88 | 89.47 | 90.07 |
| PerTA | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 |
| IPerTA | 15.50 | 14.32 | 16.04 | 3.65 | 18.42 | 17.25 |

Table 26: Joint QA+BIO unlearning results under the **entity-level 1%** forget setting using the Qwen3-0.6B **Mixed** model. Bio_ft$_F$ and Bio_ft$_R$ denote biography first-token accuracy on the forget and retain sets, respectively.

| Method | $QA_{Forget}$ | $QA_{Retain}$ | $QA_{Test}$ | Bio_ft$_F$ | Bio_ft$_R$ |
|---|---|---|---|---|---|
| Target | 100.00 | 99.99 | 94.73 | 99.67 | 100.00 |
| Retrain | 1.00 | 98.67 | 89.00 | 15.81 | 92.67 |
| GA | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| GD | 0.00 | 0.00 | 0.00 | 0.33 | 39.72 |
| NPO | 99.00 | 98.67 | 81.83 | 97.66 | 98.83 |
| DPO | 100.00 | 99.50 | 93.17 | 99.83 | 100.00 |
| SimNPO | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| RMU | 99.83 | 99.50 | 93.50 | 99.83 | 100.00 |
| IPerTA-r1 | 0.00 | 0.83 | 1.50 | 5.68 | 5.00 |
| IPerTA-r2 | 0.67 | 0.67 | 0.33 | 4.68 | 3.84 |
| IPerTA-r3 | 1.17 | 1.17 | 0.50 | 4.51 | 2.83 |
| IPerTA-r4 | 0.67 | 2.00 | 0.50 | 5.84 | 4.17 |

## D.9 Additional Experiment

### D.9.1 Generated Dataset Quality Sensitivity

To further examine generator sensitivity, we conduct an additional subset study using a stronger generator GPT5.6-Luna, while keeping the entities and attributes fixed. Specifically, we regenerate biographies for 1,000 entities, resulting in 5,000 biographies, and single-hop QA pairs for 5,000 entities, resulting in 30,000 QA examples. We evaluate the generated data using four quality metrics. Bio Diversity is measured by self-BLEU-4: for each entity, each of its five biographies is treated as the candidate and the other four as references, then averaged across entities. Lower self-BLEU indicates higher surface diversity. Validity rate measures whether the generated text faithfully realizes the controlled schema: a biography is valid if the full name and all eight schema attributes can be detected under normalized matching rules, and a QA example is valid if it is a single non-empty line containing the full name verbatim. Duplicate rate measures exact duplicates. Answer consistency (AC) measures whether the generated question asks for the intended attribute specified by its qa_type; since gold answers are copied from the entity schema, AC captures question drift rather than answer-generation errors.

Table 33 shows that the original Qwen generator already produces high-quality data: both generators achieve perfect validity and zero duplication, with comparable biography diversity and answer consistency. This suggests that our benchmark is not highly sensitive to the choice of generator. Moreover, since the target model can successfully memorize and learn from these data, the generated dataset is sufficient for evaluating unlearning behavior under our controlled benchmark setting. Importantly, the role of the generator in our benchmark is to produce natural-language realizations

Table 27: Joint QA+BIO attribute-level partial unlearning results using the Qwen3-0.6B **Mixed** model. $QA_{F,priv}$ denotes the target private attributes (birth date and birth city) for the forget entities, while $QA_{F,pub}$ denotes the remaining public attributes associated with the same entities. $Bio_nll_F$ and $Bio_nll_R$ denote the biography negative log-likelihood on the forget and retain sets, respectively.

| Method | $QA_{F,priv}(\downarrow)$ | $QA_{F,pub}$ | $QA_{R,priv}$ | $QA_{R,pub}$ | $QA_{T,priv}$ | $QA_{T,pub}$ | $QA_{test}$ | $Bio_nll_F$ | $Bio_nll_R(\downarrow)$ |
|---|---|---|---|---|---|---|---|---|---|
| Target Model | 100.00 | 100.00 | 99.99 | 99.99 | 94.54 | 94.82 | 94.68 | 4.37 | 4.11 |
| Retrain Model | 0.95 | 100.00 | 99.99 | 99.99 | 91.75 | 93.51 | 92.63 | 16.84 | 0.00 |
| GA | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 2239.16 | 2195.25 |
| GD | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 2064.16 | 60.71 |
| NPO | 44.00 | 64.25 | 42.50 | 61.25 | 25.50 | 39.75 | 35.00 | 41.89 | 20.17 |
| DPO | 100.00 | 100.00 | 99.50 | 99.50 | 93.00 | 93.25 | 93.17 | 4.26 | 4.04 |
| SimNPO | 0.00 | 25.00 | 0.00 | 22.25 | 0.00 | 13.25 | 8.83 | 378.17 | 176.71 |
| RMU | 12.00 | 78.00 | 8.00 | 78.50 | 9.50 | 70.00 | 51.95 | 49.15 | 4.09 |
| PerTA-r1 | 0.00 | 1.00 | 0.00 | 1.25 | 0.00 | 1.00 | 0.67 | 537.00 | 518.40 |
| PerTA-r2 | 0.00 | 1.50 | 0.00 | 0.50 | 0.00 | 0.50 | 0.33 | 255.50 | 258.68 |
| PerTA-r3 | 0.00 | 1.75 | 0.00 | 1.75 | 1.00 | 0.25 | 0.50 | 179.47 | 179.24 |
| PerTA-r4 | 0.00 | 1.25 | 0.00 | 0.75 | 1.00 | 0.50 | 0.67 | 146.33 | 145.22 |

Table 28: GA exhibits a sharp phase transition from full retention to complete collapse.

| Step | $QA_{forget}$ | $QA_{retain}$ | $QA_{test}$ |
|---|---|---|---|
| | GA (1e-5) | | |
| 500 | 0.00 | 0.00 | 0.00 |
| | GA (1e-6) | | |
| 500 | 100.00 | 99.33 | 94.00 |
| 1k | 100.00 | 99.33 | 93.00 |
| 1.5k | 100.00 | 99.33 | 89.33 |
| 2k | 13.33 | 12.67 | 5.33 |
| 2.5k | 0.00 | 0.00 | 0.00 |
| 3750 | 0.00 | 0.00 | 0.00 |

of a controlled underlying schema, while the entities, attributes, relations, and gold answers are fixed by construction. So the unlearning evaluation is primarily determined by whether models memorize, retain, and forget the controlled entity-attribute knowledge, rather than by uncontrolled generator-specific facts. The fact that target models can successfully learn from these generated data further confirms that the dataset is suitable for evaluating fine-grained unlearning behavior.

### D.9.2 Randomized Attribute-level Unlearning Setting

As indicated in our paper, these optimization-based methods struggle under attribute-level unlearning. This is exactly the kind of failure mode our fine-grained attribute-level evaluation is designed to expose. We added a randomized attribute-level setting: for each of the 1,000 target entities, we randomly sample two private attributes from the full 1-hop attribute pool birth date, birth city, university, major, employer, work city. Thus, each attribute type is private for roughly one-third of the entities and public for the rest, which decorrelates private or public status from attribute type.

Lower Forget is better, while higher Retain and Test are better. The results (Table 34) show that no optimization-based method achieves both fine-grained forgetting and preservation. GA and ECO reduce Forget, but they also strongly reduce Retain, indicating over-forgetting of target-public attributes. SimNPO preserves Retain and Test, but barely forgets the private attributes. IPerTA shows partial selectivity, but still leaves substantial private-attribute accuracy and hurts non-target performance. This randomized setting is more challenging than the fixed attribute-level setting, as the method must distinguish private from public attributes while also preserving the entity-attribute association. The fixed setting already reveals this difficulty, where existing methods fail to match the retrained model's retention behavior in Figure 1. The randomized setting further shows that this failure is not simply due to the fixed choice of birth date and birth city.

Selective attribute-level forgetting is possible in principle: the retrained model in Figure 1 forgets the designated private entity-attribute pairs while preserving private/public and non-target knowledge.

Table 29: GD gradually suppresses both forget and retain performance, leading to full collapse.

| Step | $QA_{forget}$ | $QA_{retain}$ | $QA_{test}$ |
|---|---|---|---|
| GD (1e-5) | | | |
| 500 | 57.67 | 64.33 | 42.33 |
| 1k | 0.00 | 2.67 | 1.00 |
| 1.5k | 0.00 | 0.33 | 0.33 |
| 2k | 0.00 | 0.00 | 0.00 |
| 3750 | 0.00 | 0.00 | 0.00 |

Table 30: NPO fails to forget and preserves near-target performance across all splits.

| Step | $QA_{forget}$ | $QA_{retain}$ | $QA_{test}$ |
|---|---|---|---|
| NPO (5e-3) | | | |
| 500 | 96.33 | 94.67 | 81.33 |
| 1000 | 96.67 | 95.67 | 81.33 |
| NPO (1e-5) | | | |
| 250–3500 | 100.00 | 99.67 | 94.33 |
| 3750 | 100.00 | 99.67 | 94.00 |

Thus, the failure comes from current approximate unlearning methods, which often suppress related attribute knowledge too broadly. This reinforces the need for our fine-grained benchmark, which evaluates whether a method forgets the intended entity-attribute pairs rather than merely reducing aggregate accuracy.

In summary, this result strengthens our benchmark contribution: it decorrelates private status from attribute type, exposes global or over-broad attribute suppression as a concrete failure mode, and shows why fine-grained entity-attribute-level evaluation is necessary.

### D.9.3 Additional Experiments using different probes

The two-hop relation-level setting is itself a stronger recovery probe. It tests whether supposedly forgotten information can be reconstructed through retained facts and a reasoning chain. This directly targets a failure mode that pure output-level suppression cannot solve: even if direct QA access is blocked, the model may still infer the forgotten answer from retained representations and learned reasoning ability.

Besides the Bio_ft and two-hop relation-level probe, we add three stronger removal-oriented probes on entity-level checkpoints: membership-inference-style loss/AUC, benign relearning, and a lightweight linear representation probe (Allen-Zhu & Li, 2023b). These probes test whether evaluations beyond QA extractability change the conclusions drawn from QA and Bio_ft (Table 35).

For MIA, we compute likelihood-based scores and report AUC for distinguishing forget biographies from retain biographies. Retraining gives the desired selective signature 0.0, while SimNPO remains at the target-model level (0.502), indicating residual member-like memorization. GD obtains a low AUC, but with severe retain-side degradation, suggesting broad damage rather than selective removal. For relearning, following benign relearning evaluation (Hu et al., 2025a), we fine-tune unlearned checkpoints on retain QA only and measure whether $QA_{forget}$ recovers. Retraining remains stable, while SimNPO recovers to 99.9% and GD to 56.1%, showing that low QA accuracy can reflect fragile suppression or damage rather than durable removal. For the linear probe, inspired by Allen-Zhu & Li (2023b), we train a lightweight logistic probe on retain entities to predict attributes from hidden states and evaluate it on forget entities. Retraining reduces forget-probe accuracy to 2.3%, while SimNPO remains close to the target model (85.6%), meaning the forgotten attributes are still linearly recoverable.

Overall, these stronger probes do not change the conclusions from QA and Bio_ft; they make the failure modes clearer. SimNPO reduces QA extractability, but its MIA score remains at the target-model level, its forget attributes remain highly linearly decodable, and $QA_{forget}$ quickly recovers under retain-only relearning. This indicates suppression rather than durable removal. GD and IPerTA reduce

Table 31: NPO exhibits delayed collapse in two-hop: strong memorization followed by sudden failure.

| Step | $QA_{forget}$ | $QA_{retain}$ | $QA_{test}$ | $Bio_ft_r$ |
|---|---|---|---|---|
| 5000 | 99.92 | 99.50 | 73.42 | 91.33 |
| 7500 | 97.25 | 97.33 | 55.42 | 90.00 |
| 8000 | 96.50 | 96.58 | 51.67 | 89.67 |
| 8.5k–15k | 0.00 | 0.00 | 0.00 | 0.00 |

Table 32: Unlearning results on the real-world MUSE dataset using the Qwen3-0.6B-Based mixed training regime under different memorization strengths.

| Method | C.1 Verbatim Mem. | C2. Knowledge Mem. | C4. Utility Preserv. | $QA_{forget}$ | $QA_{Retain}$ | $Bio_ft_F$ | $Bio_ft_R$ |
|---|---|---|---|---|---|---|---|
| **High memorization** | | | | | | | |
| Target (10000) | 35.19 | 44.30 | 40.33 | 43.00 | 39.00 | 100.00 | 100.00 |
| Retrain (10000) | 13.58 | 3.75 | 63.20 | 1.00 | 52.00 | 31.31 | 100.0 |
| GD | 16.82 | 32.69 | 30.73 | 29.00 | 32.00 | 93.94 | 97.37 |
| NPO | 34.82 | 40.97 | 39.81 | 43.00 | 39.00 | 100.00 | 100.00 |
| SimNPO | 27.20 | 35.78 | 35.72 | 40.00 | 39.00 | 100.00 | 100.00 |
| **Low memorization** | | | | | | | |
| Target (1000) | 45.38 | 22.96 | 24.50 | 39.00 | 36.00 | 100.00 | 99.47 |
| Retrain (1000) | 16.07 | 6.19 | 43.28 | 1.00 | 38.00 | 26.26 | 97.89 |
| GD | 1.13 | 0.00 | 0.00 | 0.00 | 0.00 | 6.06 | 10.53 |
| NPO | 30.11 | 19.66 | 21.55 | 29.00 | 23.00 | 96.97 | 95.26 |
| SimNPO | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.01 | 0.53 |

some forget-side signals, but not selectively: GD mainly reflects broad model damage, while IPerTA remains far from the retrained oracle. ECO matches oracle-like MIA and linear-probe scores despite changing no parameters because it is inference-time gated, so the probes observe the intervened forward pass. Thus, these probes should be interpreted together with the method mechanism and retrained oracle. Taken together, they strengthen UnlearnBench's diagnosis of suppression, damage, and selective removal.

We also avoid overstating representation probes. Prior work notes that linear probes are useful diagnostics but not definitive certificates: success shows accessibility, while failure does not prove absence of knowledge (Belinkov, 2022). This is consistent with Allen-Zhu & Li (2023b), which shows that biography facts may be memorized but not linearly encoded at the entity-name representation without sufficient augmentation. Thus, representation probing is complementary to not a replacement for the controlled benchmark with retraining references, memorization probes, QA extraction, and reasoning-based recovery.

#### D.9.4 Inference-Time Interventions: ECO and Filter Prompting

We evaluate ECO (Liu et al., 2025) and Filter Prompting on the single-hop mixed-trained Qwen3-0.6B model. Both methods use the same target classifier, with identical flag rates in each setting. Filter Prompting instructs the model to answer incorrectly when a query is flagged, whereas ECO applies an embedding-level intervention to the flagged prompt. This comparison holds target detection fixed while varying the intervention applied to detected queries.

**ECO yields stronger selective QA suppression than Filter Prompting.** Table 36 shows that sharing the same target detector does not imply comparable suppression of target answers. At the entity level, ECO reduces forget accuracy to 0.48%, compared with 93.53% for Filter Prompting, while both retain 99.99% retain accuracy and 94.72% test accuracy. The distinction persists under partial attribute deletion: ECO achieves 7.95% target-private accuracy while preserving 94.97% target-public accuracy and 99.99% non-target retain accuracy; Filter Prompting leaves target-private accuracy at 99.10%. Thus, ECO distinguishes designated and retained attributes of the same entities, although target-public accuracy is 5.03 percentage points lower than with Filter Prompting. Under all-entity attribute deletion, ECO leaves more residual private-attribute accuracy (19.32%) while preserving 99.72% public-attribute accuracy, whereas Filter Prompting again leaves private-attribute

Table 33: Ablation on the data generator.

| Generator | Bio Diversity | Bio+QA Validity | Bio+QA Duplicate | AC |
|---|---|---|---|---|
| Qwen3-30B-A3B-Instruct | 0.356 | 1.000 | 0.000 | 0.9876 |
| GPT5.6-Luna | 0.367 | 1.000 | 0.000 | 0.9704 |

Table 34: Randomized attribute-level unlearning 1%.

| Method | $QA_{forget}$ | $QA_{Retain}$ | $QA_{test}$ |
|---|---|---|---|
| GA | 10.8 | 11.1 | 7.9 |
| SimNPO | 97.8 | 97.6 | 89.7 |
| IPerTA | 63.5 | 74.9 | 44.5 |
| ECO | 4.5 | 6.1 | 94.7 |

accuracy high (99.22%). These results demonstrate that the benchmark captures substantial differences between inference-time interventions under matched target detection. They characterize selective suppression through the evaluated QA interface, rather than removal of knowledge from model weights, which these interventions leave unchanged.

## E Related Work

**LLM Unlearning.** LLM unlearning aims to remove specific knowledge from trained models while preserving overall utility. Existing unlearning approaches can be broadly categorized into three lines. First line is optimization-based unlearning methods, including finetuning-based unlearning with specialized objectives (Jia et al., 2024; Maini et al., 2024; Wang et al., 2024c; Yao et al., 2025; Wang et al., 2025b; Wuerkaixi et al., 2025; Tan et al., 2025; Entesari et al., 2025; Zhang et al., 2025) or a set of modified responses (Choi et al., 2024; Gu et al., 2024; Mekala et al., 2024), representation redirection approaches (Eldan & Russinovich, 2023; Chen & Yang, 2023; Li et al., 2024; Ji et al., 2024; Shen et al., 2025; Sepahvand et al., 2025; Wang et al., 2026b), preference-style methods and its variants (Rafailov et al., 2023; Zhang et al., 2024a; Fan et al., 2024; Sui et al., 2025; Yang et al., 2026). Another line is about localization and representation-level control. They attempt parameter or circuit localization for selective editing or unlearning, including internal representation interventions (Yamashita et al., 2025), mechanistic localization (Guo et al., 2024), parameter-efficient unlearning (Cha et al., 2024; Ding et al., 2024; Kim et al., 2025), and model editing-related methods (Wu et al., 2023; Belrose et al., 2023; Ilharco et al., 2022; Dong et al., 2024). PerTA (Cai et al., 2026) weights the task vector subtraction per parameter based on how important each parameter is for forget vs retain via diagonal Fisher information approximation. At the same time, several studies question whether localized edits imply true deletion in weights or robust behavioral forgetting (Deeb & Roger, 2024; Chen et al., 2024; Hong et al., 2025; Lee et al., 2025). Finally, there's also a training-free or generation time unlearning line, which avoid altering model weights by steering model behavior through prompt intervention (Pawelczyk et al., 2023; Muresanu et al., 2024; Thaker et al., 2024; Wang et al., 2024b; 2026a), embedding manipulation (Bhaila et al., 2024; Liu et al., 2025), or generation time redirection (Deng et al., 2025).

Despite these advances, several core challenges remain. Current unlearning methods are often evaluated on uncontrolled corpora and inconsistent training settings, making comparisons difficult and obscuring the underlying failure modes. Moreover, most existing evaluations focus on suppressing single-hop QA recall, rather than testing whether forgotten knowledge can still be recovered through reasoning. As a result, current methods struggle to disentangle memorized knowledge from generalizable reasoning abilities. Our empirical study evaluates representative unlearning methods under a unified, controlled, and fully synthetic protocol. The results show that unlearning performance is highly regime-dependent, and that existing methods fail substantially on our more challenging compositional and fine-grained benchmark settings.

Table 35: Experimental results using three removal-oriented probes on entity-level checkpoints.

| Method | $QA_{forget}$ | MIA Bio AUC (F vs R) | Retain-only Relearning $QA_{forget}$ | Linear Probe Forget Acc. (%) |
|---|---|---|---|---|
| Target | 100.00 | 0.50 | 100.00 | 87.9 |
| Retrain | 1.00 | 0.00 | 1.60 | 2.30 |
| GD | 0.0 | 0.063 | 56.1 | 7.1 |
| SimNPO | 53.68 | 0.502 | 99.9 | 85.6 |
| IPerTA | 82.32 | 0.482 | 85.32 | 11.2 |
| ECO | 0.48 | 0.0 | - | 2.7 |

Table 36: ECO and Filter Prompting under single-hop entity-level and fixed attribute-level deletion settings. All values are QA accuracy (%). Lower forget accuracy and higher retained/test accuracy are preferred. In the partial attribute setting, $QA_{F,pub}$ measures public attributes of the target entities and $QA_{retain}$ measures the non-target retain set. In the all-entity attribute setting, $QA_{R,pub}$ measures retained public attributes. The designated private attributes are birth date and birth city. Dashes denote inapplicable columns. Results evaluate the systems with their inference-time interventions enabled.

| **Entity-level (1,000 target entities)** | | | | |
|---|---|---|---|---|
| Method | $QA_{forget}\downarrow$ | — | $QA_{retain}\uparrow$ | $QA_{test}\uparrow$ |
| Filter Prompting | 93.53 | — | 99.99 | 94.72 |
| ECO | 0.48 | — | 99.99 | 94.72 |
| **Attribute-level: partial (1,000 target entities)** | | | | |
| Method | $QA_{F,priv}\downarrow$ | $QA_{F,pub}\uparrow$ | $QA_{retain}\uparrow$ | $QA_{test}\uparrow$ |
| Filter Prompting | 99.10 | 100.00 | 99.99 | 94.74 |
| ECO | 7.95 | 94.97 | 99.99 | 94.72 |
| **Attribute-level: all (50,000 target entities)** | | | | |
| Method | $QA_{F,priv}\downarrow$ | — | $QA_{R,pub}\uparrow$ | $QA_{test}\uparrow$ |
| Filter Prompting | 99.22 | — | 99.98 | 94.74 |
| ECO | 19.32 | — | 99.72 | 94.72 |

**LLM Unlearning Benchmarks.** Evaluating unlearning in LLMs remains a critical challenge. Most previous studies have focused this evaluation on specific tasks such as question answering (Maini et al., 2024; Li et al., 2024) or sentence completion (Eldan & Russinovich, 2023; Shi et al., 2024; Ramakrishna et al., 2025) and try to propose comprehensive metrics suits to measure the unlearning progress (Hu et al., 2025b; Cao et al., 2024). TOFU (Maini et al., 2024) introduced fictitious-author unlearning as a controlled benchmark. Subsequent work broadened scope and metrics, including copyright content unlearning in MUSE (Shi et al., 2024), hazardous-knowledge evaluation in WMDP (Li et al., 2024), and real-world entity-focused benchmarks such as RWKU (Cao et al., 2024). OpenUnlearning (Dorna et al., 2025) further standardized multi-benchmark comparisons. However, previous research (Thaker et al., 2025) indicates that existing benchmarks are weak measures of unlearning progress. They generally treat knowledge independently, overlooking the interdependencies between target facts and related knowledge in LLMs. (Wei et al., 2025b) proposes a knowledge unlearning evaluation framework that more accurately captures the implicit structure of real-world knowledge by representing relevant factual contexts as knowledge graphs with associated confidence scores. Most existing unlearning benchmarks focus on single-hop factual recall, where the target knowledge can be directly queried. However, real-world knowledge is often compositional, requiring multi-step reasoning. Recent work in multi-hop QA and chain-of-thought (CoT) reasoning demonstrates that models can: retrieve latent knowledge through intermediate steps,reconstruct answers even when direct recall appears suppressed. This suggests a critical gap: a model may appear to forget a fact in isolation but still recover it through reasoning chains. Current unlearning benchmarks largely ignore this phenomenon, leading to an incomplete assessment of true forgetting.

Besides the benchmark itself, some literature aim to propose comprehensive and effective evaluations (Wang et al., 2024a), such as non-deterministic evaluations (Scholten et al., 2024) and benign

relearning (Hu et al., 2025a). (Zhang et al., 2024b) reveals that applying quantization to an unlearned model can restore the "forgotten" information and propose a quantization-robust method. (Fan et al., 2025) proposes an unlearning method utilizing sharpness-aware minimization, which is robust to relearning attacks (Xu et al., 2025).

Our benchmark complements these lines by explicitly varying *training regime* (skill-like generalization vs memorization-heavy storage), *hop complexity*, and *unlearning granularity* (entity vs attribute). Unlike leaderboard-style studies that primarily aggregate method scores, our focus is to characterize *when* unlearning is feasible versus structurally difficult under controlled training-regime shifts, and to provide a benchmark axis that reveals failure modes hidden by single-number reporting.

**LLM Memorization and Reasoning during Pretraining.** Understanding how LLMs memorize and retrieve training data has become increasingly important due to privacy, copyright, and safety concerns. Early work demonstrates that LLMs can memorize and even verbatim extract training data (Carlini et al., 2020; 2022). Other studies further show that pretrained language models store substantial factual knowledge that can be retrieved through prompting (Petroni et al., 2019). Recent controlled studies investigate how memorization abilities emerge during pre-training (Prashanth et al., 2024; Wei et al., 2025a). In particular, the Physics of Language Models series (Allen-Zhu & Li, 2023b;a; 2024) introduces synthetic-person pretraining paradigms and studies storage, retrieval, and manipulation as separable phenomena. HUBBLE (Wei et al., 2025a) finds that dilute sensitive data by increasing the training corpus size and order sensitive data to appear earlier in training can help to address memorization risks. Recent work also explores the relationship between memorization and reasoning generalization, showing that factual QA tasks rely more heavily on memorization while reasoning tasks exhibit stronger generalization behavior (Antoniades et al., 2024). These findings motivate the need for clean and controllable benchmarks to systematically study memorization and unlearning dynamics.

# F Discussion

**More about the Benchmark Design Rationales.** Existing LLM unlearning benchmarks have several limitations. First, knowledge entanglement is difficult to measure in current settings. When an unlearning method performs well, it is often unclear whether the improvement comes from the method itself or from the fact that the target knowledge was easy to remove. This issue is further complicated by how the target knowledge is introduced. Some benchmarks inject knowledge through simple fine-tuning, which does not reflect the realistic training process of LLMs, where knowledge may be acquired through pretraining, mid-training, or post-training. As a result, it is difficult to trace where the target knowledge comes from and how it is organized in the model. Since the acquisition and organization of knowledge directly affect the difficulty of selective removal, a clean and controllable benchmark is necessary before making reliable progress in LLM unlearning. Second, existing benchmarks mostly focus on shallow single-hop QA suppression, while overlooking more challenging settings such as compositional reasoning and multi-stage memorization. They also lack fine-grained evaluations, such as attribute-level or relation-level forgetting, which are important for measuring knowledge entanglement.

Our benchmark is designed to address these gaps. It introduces a clean and controllable evaluation pipeline that supports fine-grained and compositional unlearning under harder memorization regimes. Using this benchmark, we obtain several important findings: the memorization stage and training recipe can significantly affect unlearning behavior; strong knowledge entanglement persists in reasoning models; and selective forgetting is harder than previously assumed.

The use of synthetic data is also intentional. Our goal is to isolate the unlearning problem itself in a controlled setting. By using continual pretraining on synthetic data, we can ensure that the target knowledge is newly introduced and not already entangled with the pretraining corpus of the base model (Qwen3-0.6B). This makes the source and structure of the target knowledge more traceable, allowing us to more cleanly evaluate whether an unlearning method can remove the intended knowledge while preserving related or retained knowledge. We view this benchmark as a controlled stress test: it is not intended to replace real-world evaluation, but to provide a clean and challenging setup for evaluating the potential and limitations of unlearning methods.

**Scope of the benchmark.** The benchmark does not cover every possible form of unlearning, such as concept-level, skill-level, style-level, or domain-level unlearning. These settings involve different definitions of the forget set, different notions of success, and different evaluation protocols. Our contribution is instead a controlled and clean benchmark for fine-grained factual unlearning. We deliberately center the benchmark on synthetic biographies because this gives us full control over what knowledge is introduced, when it is introduced, how strongly it is memorized, and what should be forgotten or retained. This controlled design allows us to isolate a fundamental question: how does unlearning depend on the way knowledge is memorized, retrieved, and reused through reasoning?

Although the entities are synthetic persons, the benchmark is not limited to removing names. It includes entity-level, attribute-level, and relation-level forgetting, as well as single-hop, two-hop, and multi-stage training settings. These settings test whether a method can selectively remove structured factual associations, whether retained knowledge can recover forgotten information, and how the training regime affects unlearning difficulty. We believe these mechanisms are fundamental to many broader unlearning scenarios, including domain or concept unlearning, where the key difficulty is often not the domain label itself but how the relevant knowledge is stored and entangled in the model.

Future extensions include the above situations. For example, concept-level unlearning could use synthetic concepts with controlled attributes and neighboring concepts; skill-level unlearning could introduce procedural abilities or reasoning patterns; style-level unlearning could vary style while preserving content; and domain-level unlearning could use synthetic domains with known entities, relations, and rules. Across these extensions, the key idea remains the same: control when knowledge is introduced, how it is memorized, what should be forgotten, and what should be retained.

**We prioritize controlled knowledge provenance over model scale so that unlearning results can be attributed to the method rather than differences in how the model acquired its knowledge.** Evaluating larger models, reasoning models, and instruction-tuned models is an important direction. However, in this paper, our primary goal is to build a clean and controllable benchmark for studying the unlearning problem itself. Therefore, we focus on models trained from scratch or continually pretrained from base models, where the target knowledge provenance, memorization stage, and forget-retain splits can be precisely controlled. **Directly using instruction-tuned or advanced reasoning models would introduce substantial ambiguity**, since their knowledge may come from pretraining, instruction tuning, alignment, or reasoning-oriented training. This makes it difficult to isolate the source of the target knowledge and attribute the results to the unlearning method itself. **A small-scale, well-controlled setting is a necessary first step:** if a method already struggles here, it is unlikely to reliably scale to larger and less controlled models. Conversely, methods that perform well under this controlled setup can be further tested on larger-scale models in future work. This design also makes the benchmark relatively inexpensive to run. We also include a controlled reasoning setting by training a 2-hop reasoning checkpoint from scratch and through mixed continual pretraining, which keeps the reasoning chain and knowledge source explicit and analyzable, compared with advanced reasoning models such as DeepSeek-R1.

**Benchmark outcomes depend primarily on the controlled facts and forget/retain structure, while the data generator mainly affects how those facts are expressed.** The quality of the generated text may affect linguistic naturalness and diversity, but the benchmark evaluation is not directly determined by the raw capability of the generator model (Zhu et al., 2025). Our dataset is generated through a structured pipeline: entities, attributes, relations, forget/retain splits, and oracle answers are produced from a controlled schema (Appendix A), while the LLM generator is mainly used to verbalize these predefined facts into biographies and QA pairs. Thus, a stronger generator may improve fluency and surface-level diversity, but should not change the underlying unlearning targets or evaluation labels.

We also explicitly encourage diversity during generation. Following Allen-Zhu & Li (2023b), we generate five biography variants for each person, reducing dependence on a single surface realization of the same fact. Therefore, we expect the main failure modes observed in our benchmark, including under-forgetting, utility collapse, and reasoning-based recovery, to remain, as they arise primarily from controlled knowledge provenance, forget/retain structure, and compositional relations rather than generator-specific wording. Moreover, our result suggests that our benchmark is not highly sensitive to the choice of generator.

**Diagnostic usefulness.** The either overly destructive or overly conservative results are the diagnostic findings. Our benchmark is designed as a controlled and clean stress test for unlearning: the target knowledge is deliberately well-memorized, isolated, and measurable, so that we can test whether a method can selectively remove it while preserving retained knowledge.

Under entity-level forgetting, many existing optimization-based methods fail this stress test, aligning with previous findings (Mavrothalassitis et al., 2025). They either do not remove the target knowledge, or they damage the model broadly, leading to the apparent conservative/destructive pattern. The retraining oracle reaches the desired behavior, showing that the target point is attainable; the issue is that current methods often cannot reach it.

The benchmark itself is not inherently binary. For example, the prompt-based method ECO achieves strong selective behavior on entity-level unlearning and a favorable forgetting-utility balance on attribute-level unlearning, indicating that when a method does achieve selective unlearning, the benchmark captures it clearly. The benchmark is also intentionally hierarchical. Single-hop entity unlearning tests whether a method can remove directly memorized facts. The two-hop setting is harder: even if direct QA access is suppressed, the model may recover forgotten information through compositional reasoning over retained facts. This exposes the entanglement between memorization and reasoning.

**Relative difficulty of Single-hop and Two-hop Settings. Although many baselines fail in both settings, the two tasks are difficult for different reasons, and our benchmark is designed to expose this distinction.**

In the single-hop setting, a valid unlearning solution exists: the retraining oracle removes the target knowledge while preserving retained knowledge. Thus, when optimization-based methods fail, this reflects a method-level failure: current methods cannot localize and remove the target entity knowledge selectively under our controlled stress test.

In the two-hop and multi-stage settings, the challenge is fundamentally stronger. Removing only the direct QA supervision is insufficient, because the model may reconstruct the supposedly forgotten relation through compositional reasoning over retained biography knowledge. In other words, the forgotten answer is not stored only in the direct QA mapping; it is entangled with retained memorized facts and reasoning paths. This makes unlearning harder than standard single-hop entity forgetting, especially for models that have acquired reasoning ability from broader data distributions.

This distinction is one of the main findings of our work: the difficulty of unlearning depends not only on the amount of forget data, but also on how the knowledge was acquired and organized inside the model. Both settings require better methods than current optimization-based baselines, but the two-hop setting additionally requires methods that can address reasoning-mediated recovery, potentially through more precise localization, suppression, or revision of the relevant internal representations.