Anatomy of Unlearning: The Dual Impact of Fact Salience and Model Fine-Tuning

Anna Borisiuk[1,4], **Andrey Savchenko**[2,3,5], **Alexander Panchenko**[1,4], **Elena Tutubalina**[1,5]
[1]AIRI, [2]Sber AI Lab, [3]HSE University, [4]Skoltech,
[5]ISP RAS Research Center for Trusted Artificial Intelligence

Abstract

Machine Unlearning (MU) enables Large Language Models (LLMs) to remove unsafe or outdated information. However, existing work assumes that all facts are equally forgettable and largely ignores whether the forgotten knowledge originates from pretraining or supervised fine-tuning (SFT). In this paper, we introduce the benchmark DUET (Dual Unlearning Evaluation across Training Stages) composed of Wikidata-derived triplets annotated with fact popularity scores derived from Wikipedia link counts and LLM-based salience scores. Our experiments show that pretrained and SFT models respond differently to unlearning. An SFT step on the forget data yields smoother forgetting, more stable tuning, and 10-50% higher retention, while direct unlearning for pretrained models remains unstable and prone to relearning or catastrophic forgetting.

1 Introduction

Large Language Models have become central to modern NLP applications, yet their strong memorization of training data raises pressing questions about how to remove unsafe, outdated, or private information after deployment. Machine Unlearning aims to erase specific knowledge while preserving the model's overall competence, enabling safer and more adaptable systems (Cao and Yang, 2015; Sekhari et al., 2021; Kurmanji et al., 2023; Yuan et al., 2025). Despite rapid progress, two fundamental aspects remain underexplored. First, prior work often assumes that all facts are equally forgettable, overlooking how knowledge frequency and real-world prominence affect persistence in model parameters. Popular facts, frequent and widely distributed, may be more deeply embedded than rare ones, making them harder to erase (Wang et al., 2025; Yuan et al., 2025). Second, the role of the training paradigm is rarely systematically examined: most studies focus on supervisely fine-tuned (SFT) models built on synthetic data (Maini et al., 2024a; Shi et al., 2024), while a few recent efforts (Xu et al., 2025; Li et al., 2024) explore pretrained checkpoints without a controlled comparison to SFT. This gap motivates our research question:

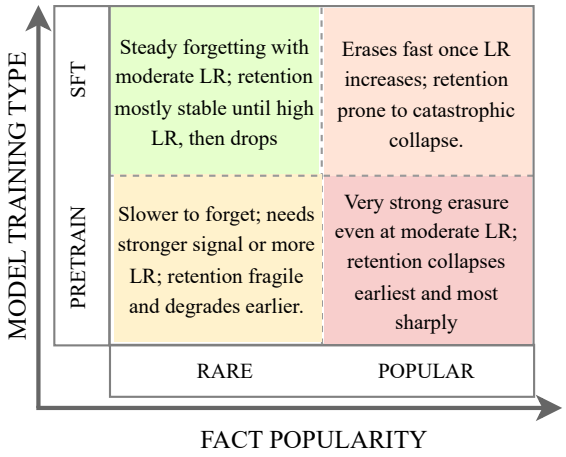
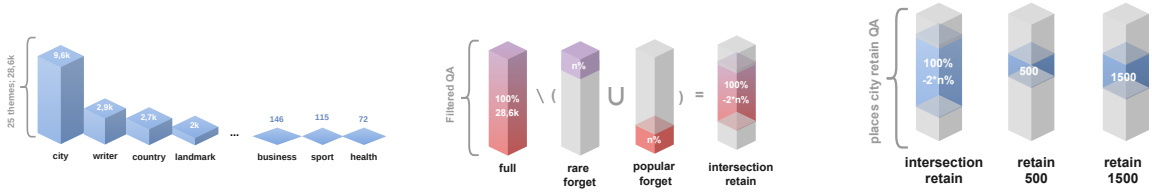


Figure 1: Unlearning landscape across fact popularity and model training type. Existing unlearning work does not account for popularity, implicitly assuming all facts are equal. Most studies evaluate forgetting on pretrained or SFT models without contrasting the two. The underexplored quadrant concerns analyzing how unlearning differs between pretrained and SFT models when fact popularity is taken into account.

How do fact popularity and model training type jointly influence machine unlearning performance?

To answer this, we introduce **DUET**, a benchmark of 28.6k Wikidata-derived question-answer pairs annotated for fact popularity using Wikipedia statistics and model-perceived salience. DUET enables controlled comparison of unlearning performance across popularity levels and between *Pretrain* and *SFT* versions of the same architecture. We systematically analyze how these two factors interact and find that popular facts are substantially harder to erase, while pretrained and SFT models respond qualitatively differently to the same unlearning signals (cf. Figure 1). These findings

Findings of the Association for Computational Linguistics: ACL 2026, pages 25846–25859



(a) distribution of topics (b) distribution of entity frequency (c) retain sets

Figure 2: **Overview of the DUET benchmark.** (a) Topic distribution across 25 semantic themes (28.6k validated QA pairs), dominated by the *places city* domain. (b) The filtering and stratification to derive rare and popular forget sets and the retain intersection. (c) Compact retain subsets (*retain 500* and *retain 1500*) drawn from *places city*, providing resource-efficient yet structurally consistent evaluation settings.

highlight that unlearning is shaped not only by the chosen algorithm but also by where and how knowledge resides in the model.

The code[1] and the dataset[2] are available online.

2 Related Work

Machine unlearning aims to selectively remove the impact of specific data points from a trained model without requiring complete retraining (Mantelero, 2013; Cao and Yang, 2015; Sekhari et al., 2021). The objective is to obtain a model that behaves as if the forgotten data were never part of its training set. Early work explored MU in classical machine learning, while more recent research targets textual unlearning in LLMs (Sekhari et al., 2021; Kurmanji et al., 2023). A large body of studies focus on the LLaMA family-model (Yuan et al., 2025; Maini et al., 2024b; Wang et al., 2025; Si et al., 2023), which have become a standard testbed due to their effectiveness in controlled unlearning experiments (Maini et al., 2024b; Yuan et al., 2025). For this reason, our experiments also center on LLaMA-3.1-8B variants.

Unlearning benchmarks were proposed to evaluate MU methods. Many of them rely on synthetic data, creating controlled but artificial scenarios. For example, TOFU (Maini et al., 2024b) introduces 200 fictitious author profiles and constructs forget sets at 1%, 5%, and 10% scales, with the remainder used to measure retention. While valuable for privacy-like setups, its synthetic nature does not capture the complexity of real-world knowledge distributions. Other synthetic benchmarks include (Geng et al., 2025; Hu et al., 2024; Si et al., 2023).

Benchmarks based on real-world data address different objectives but largely ignore knowledge prevalence. For example, WMDP (Li et al., 2024) provides 3,668 multiple-choice questions to assess forgetting of hazardous knowledge, while MUSE (Shi et al., 2024) emphasizes a multifaceted evaluation that covers dimensions such as efficacy and efficiency. These resources are essential, yet their focus is orthogonal to the role of fact popularity.

Limitations of existing datasets stems from the fact that, they typically assumes all facts equally forgettable, overlooking how strongly frequent, widely distributed knowledge may be embedded compared to rare facts. Second, little attention has been given to the role of the training paradigm: many benchmarks rely on SFT models built on synthetic data, e.g., (Maini et al., 2024a), as pretrained models alone are often not directly evaluable, while other works, e.g., (Xu et al., 2025), attempt unlearning directly on pretrained models. Consequently, the joint effect of fact popularity and model training type has not been systematically studied.

3 DUET Benchmark

Data Source We construct **DUET** from 57k Wikidata-derived factual triplets spanning 25 semantic topics, including places, cities, human writers, countries, landmarks, industries, and health symptoms (Figure 2a). For each fact, we compute a *popularity score* as the sum of Wikipedia sitelinks of its subject and object entities, capturing external prominence.

Filtering and QA Generation We remove instances where the same subject–relation pair yields multiple candidate answers to guarantee a unique answer per question, keeping only the answer associated with the most popular object. Each remaining triplet is converted into a natural-language question–answer pair (*subject–relation* $\rightarrow$ *object*), following a similar disambiguation procedure as

[1] https://github.com/Anya-wUw/DUET

[2] https://huggingface.co/datasets/SwetieePawsss/DUET

in (Huang et al., 2025). To verify that the knowledge is already accessible to pretrained models, we prompt *LLaMA-3.1-8B* and retain only 28.6k triplets where the model's generated answer reaches BERT cosine similarity above 0.6 with the gold answer, ensuring that unlearning targets facts the model has actually internalized rather than knowledge it never encoded in the first place.

Popularity Validation To confirm that Wikipedia-based popularity correlates with model-internal salience, we compare it with factual judgments from *LLaMA-3.3-70B*. The model rates each fact's prominence on a three-point scale: -1 (disagree), 0 (unknown), and 1 (agree). The two signals are strongly correlated (80.65% on 5% forget splits and 81.95% on the *places city* subset; see Figure 4), validating the reliability of our external metric. Among existing QA datasets, the closest in spirit is **PopQA** (Mallen et al., 2023), which also quantifies factual popularity; however, DUET is roughly twice as large and specifically designed for controlled unlearning experiments across training stages.

Popularity-Based Splits Following methodology from (Maini et al., 2024a; Dontsov et al., 2025), we stratify DUET into unlearning tasks at 1%, 5%, and 10% scales. For a given proportion N, we define three complementary subsets (Figure 2b):

1. **Rare forget set** is composed of bottom $N\%$ of least popular facts.
2. **Popular forget set** is composed of top $N\%$ of most popular facts.
3. **Retain intersection set** is composed of the remaining $(100 - 2N)\%$ of data, used to evaluate the preservation of unaffected knowledge.

City Sets To support efficient experimentation, we replicate the same popularity-based partitioning (rare vs. popular for forget and retain) within the most prominent domain, *places city* (9.6k samples). We construct smaller sets for rapid validation from this category: *retain 500* and *retain 1500*. These subsets are randomly sampled and filtered with a stricter BERT similarity threshold (> 0.7), selected empirically to balance coverage and factual confidence: lower thresholds admit noisy paraphrases, while higher thresholds excessively reduce set size. This ensures full alignment with the structure of the complete benchmark (Figure 2c).

4 Experiments

4.1 Experimental Setup

We conduct our experiments using the LLaMA-3.1-8B model as the base architecture. To obtain comparable Pretrained and SFT variants under identical conditions, we use the released checkpoint as the Pretrained model and train an SFT variant on the full DUET dataset (28.6k samples) using LoRA. Parameter-efficient fine-tuning achieves performance close to full-model training while requiring substantially fewer resources (Hu et al.; Sun et al., 2023; Ding et al., 2023), and LoRA-based unlearning has emerged as the standard approach for large models (Cha et al., 2025; Liu et al., 2025).

We apply LoRA to the unlearning step as well. Full-parameter SFT produces substantially less stable unlearning: on the city-forget split with NPO at learning rate of $2 \cdot 10^{-5}$, the full-SFT model retains a ROUGE-L of 0.997 on the forget set, while LoRA reduces it to 0.364. On multi-domain data, full-SFT models collapse catastrophically across all algorithms, dropping ROUGE-L to 0-0.14 on the forget set, whereas LoRA models maintain stable forgetting and near-unchanged retention. Full results are in Appendix B.1.

For unlearning experiments, we focus on the largest topical category, places city (9.6k samples), and apply three widely used algorithms: Gradient Ascent (GA) (Jang et al., 2022), Gradient Difference (GD) (Liu et al., 2022), and Negative Preference Optimization (NPO) (Zhang et al., 2024). We perform a grid search over learning rates from 10^{-6} to $5 \cdot 10^{-5}$ and training epochs from 1 to 3. Learning rates above $5 \cdot 10^{-5}$ consistently led to catastrophic forgetting, while rates below 10^{-6} resulted in negligible unlearning. We identify two epochs as a practical compromise (cf. Figure 3): 1 epoch yields insufficient forgetting (cf. Figure 5), while three epochs cause excessive knowledge degradation. We report primary results for N of 5%; experiments for N of 1% and 10% are provided in Appendix B.2 and show consistent qualitative trends.

We use the *fast retain* subset of 500 samples to evaluate retention quality. ROUGE-L is the primary metric for forgetting and retention; we additionally validate it via an LLM-as-a-Judge protocol (DeepSeek v1), scoring *Accuracy* (factual alignment with the reference) and *Fluency* (linguistic naturalness). Both metrics are reported for LLaMA in Table 1; Gemma results are in Appendix B.3.

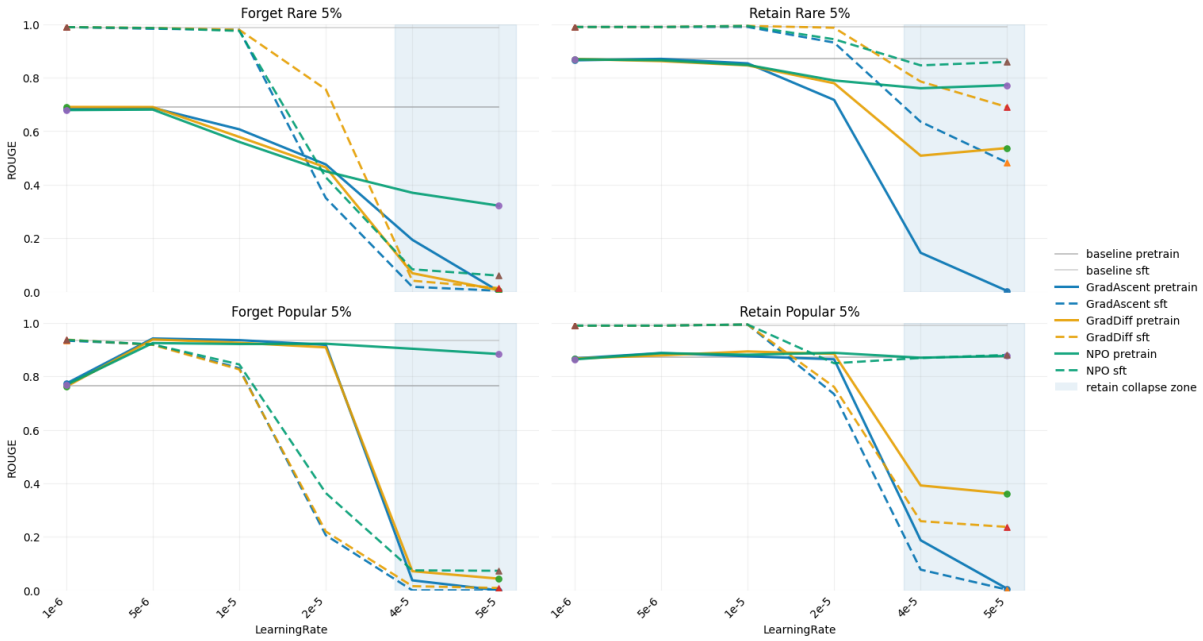


Figure 3: **Unlearning results for the city domain** and, top $N = 5\%$ rare/popular entities. Top: rare; bottom: popular. Left: forget (`city_forget_{rare,popular}_5`); right: retain (`city_fast_retain_500`). ROUGE is reported across learning rates. Pretrain is shown as a solid line with circles; SFT as a dashed line with stars.

4.2 Key Findings

Finding 1: Pretrained and SFT models exhibit opposite behavior on popular facts When unlearning popular entities, the Pretrained model displays counter-intuitive behavior: instead of forgetting, it improves performance on the forget set, with ROUGE-L scores increasing across all tested epochs (1-3). This suggests the model treats unlearning signals as additional fine-tuning on familiar knowledge. In stark contrast, the SFT model behaves as expected: unlearning consistently decreases ROUGE-L on the forget set. This divergence persists regardless of epoch count, indicating a fundamental difference in how the two model training types process unlearning interventions on well-known facts.

For rare entities, both models behave conventionally: forgetting metrics decline exponentially with training. As learning rates increase, both architectures eventually reach catastrophic forgetting thresholds where performance collapses uniformly.

Takeaway 1: For *popular entities*, a preliminary SFT step enables more stable and reliable unlearning.

Finding 2: retention degradation differs dramatically by fact popularity On the retain set, Pretrained models show relatively stable ROUGE-L scores regardless of whether rare or popular facts are removed: popularity has little effect on retained knowledge quality, though forgetting is accompanied by a sharp drop once catastrophic thresholds are reached.

In contrast, SFT models display the opposite trend: unlearning popular facts leads to catastrophic forgetting of the retain set, while unlearning rare facts does not cause such drastic quality loss. Thus, SFT models are overall more robust, reducing the risk of catastrophic retention collapse by roughly a factor of two compared to rare-fact forgetting. Pretrained models, by comparison, behave more uniformly across fact types but are prone to sudden quality drops when forgetting escalates.

Takeaway 2: *SFT models are more robust* on the retain set, showing roughly half the risk of catastrophic forgetting compared to rare-fact removal, while pretrained models degrade more abruptly regardless of fact type.

Unlearning preserves overall capability: the worst-case deviation on MMLU and HellaSwag does not exceed 3% across all tested configurations (cf. Appendix B.4 and Figure 7). Representation-level analyses in Appendix B.5 reveal that rare and popular facts respond asymmetrically to unlearning at the token and hidden-state levels: popular facts

Train type	Algo	Pop.	ROUGE-L		Judge Acc.		Judge Flu.	
			Forget ↓	Retain ↑	Forget ↓	Retain ↑	Forget ↑	Retain ↑
Pretrain	w/o	rare	0.691	0.871	0.584	0.839	0.701	0.840
SFT	w/o	rare	0.987	0.990	0.773	0.862	0.741	0.649
Pretrain	GA	rare	0.476 (−0.22)	0.717 (−0.15)	0.368 (−0.22)	0.670 (−0.17)	0.646 (−0.06)	0.748 (−0.09)
SFT	GA	rare	0.350 (−0.64)	0.931 (−0.06)	0.507 (−0.27)	0.842 (−0.02)	0.642 (−0.10)	0.615 (−0.03)
Pretrain	GD	rare	0.465 (−0.23)	0.779 (−0.09)	0.358 (−0.23)	0.747 (−0.09)	0.577 (−0.12)	0.627 (−0.21)
SFT	GD	rare	0.756 (−0.23)	0.987 (−0.00)	0.572 (−0.20)	0.853 (−0.01)	0.694 (−0.05)	0.612 (−0.04)
Pretrain	NPO	rare	0.450 (−0.24)	0.790 (−0.08)	0.349 (−0.23)	0.745 (−0.09)	0.608 (−0.09)	0.618 (−0.22)
SFT	NPO	rare	0.428 (−0.56)	0.944 (−0.05)	0.521 (−0.25)	0.828 (−0.03)	0.637 (−0.10)	0.589 (−0.06)
Pretrain	w/o	pop	0.766	0.871	0.915	0.839	0.899	0.840
SFT	w/o	pop	0.935	0.990	0.862	0.862	0.606	0.649
Pretrain	GA	pop	0.918 (+0.15)	0.864 (−0.01)	0.804 (−0.11)	0.752 (−0.09)	0.742 (−0.16)	0.720 (−0.12)
SFT	GA	pop	0.206 (−0.73)	0.733 (−0.26)	0.261 (−0.60)	0.555 (−0.31)	0.604 (−0.00)	0.673 (+0.02)
Pretrain	GD	pop	0.908 (+0.14)	0.884 (+0.01)	0.817 (−0.10)	0.760 (−0.08)	0.777 (−0.12)	0.753 (−0.09)
SFT	GD	pop	0.220 (−0.72)	0.760 (−0.23)	0.254 (−0.61)	0.588 (−0.27)	0.644 (+0.04)	0.669 (+0.02)
Pretrain	NPO	pop	0.922 (+0.16)	0.888 (+0.02)	0.793 (−0.12)	0.823 (−0.02)	0.763 (−0.14)	0.758 (−0.08)
SFT	NPO	pop	0.364 (−0.57)	0.849 (−0.14)	0.276 (−0.59)	0.683 (−0.18)	0.703 (+0.10)	0.690 (+0.04)

Table 1: LLaMA-3.1-8B at $N = 5\%$, $lr = 2 \times 10^{-5}$. ROUGE-L measures lexical forgetting and retention. Judge Accuracy measures factual alignment with the reference answer; Judge Fluency measures linguistic naturalness (DeepSeek v1). Parentheses show change from the matching *w/o* baseline.

sustain higher token probability under gradient-based pressure, while rare facts show larger shifts in hidden-state similarity after unlearning.

Thus, the popularity of the fact and the type of model training jointly determine unlearning behavior. Pretrained models respond uniformly but abruptly to unlearning signals, with narrow and unstable hyperparameter windows that make reliable forgetting difficult to achieve. SFT models are strongly shaped by fact popularity: popular facts resist forgetting at low learning rates and cause larger retention drops at high ones, while rare facts are removed smoothly across the full range. Despite this sensitivity, SFT models exhibit substantially better forgetting dynamics overall, with retention quality 10 to 50% higher than pretrained counterparts at the same learning rate (see Figure 3).

Both effects persist under distillation-based unlearning: UNDIAL experiments (cf. Appendix B.6) show that SFT retains near-baseline retention while forgetting progresses smoothly, and the rare/popular asymmetry holds under a self-distillation mechanism, confirming that these patterns are not tied to gradient-based objectives alone. These findings underscore the need to jointly evaluate data composition and model training type and highlight the practical benefit of a preliminary SFT step for achieving more stable, controllable unlearning.

We further evaluate the same setup on Gemma-7B and Qwen-2.5-7B to test model-specific effects. As detailed in Appendix B.7, Gemma shows the same qualitative patterns as LLaMA: rare facts are easier to forget, while SFT variants yield smoother but more sensitive forgetting dynamics. Multi-domain experiments in Appendix B.8 further confirm that the rare/popular asymmetry is not an artifact of the city-heavy distribution, but holds across balanced domain subsets.

5 Conclusion

We introduced a Wikidata-based benchmark annotated with fact popularity. It enabled us to conduct experiments, which uncovered that unlearning dynamics are jointly depend on fact popularity and model training type. Besides, we found that pretrained models are unstable, prone to abrupt degradation, and even unintended relearning of popular facts. By contrast, SFT models provide a smoother forgetting, more reliable hyperparameter tuning, and up to 10–50% higher retention quality.

These results challenge the common assumption that all facts are equally forgettable and that model training type is irrelevant. We argue that future MU methods must consider both the composition of the forget set and the origin of knowledge in the model. The proposed dataset offers enables a more reliable evaluation and a deeper understanding of how popularity and training paradigm interact in unlearning.

Limitations

Our claims hold under the following boundary conditions. (i) Scope: experiments target LLaMA-3.1-8B and the *Places City* forget set with a compact retain set; broader domains and larger models are future work. (ii) Popularity labels: we rely on Wikipedia signals and model salience; these proxies may drift over time and across languages. (iii) Metrics: we report ROUGE-L for free-form answers; complementary factuality and safety judgments, including human evaluation, are planned follow-ups. These limits do not alter the central result that popularity and training regime jointly shape unlearning outcomes.

Ethics Statement

Our data come from public sources (Wikidata, Wikipedia). We do not collect sensitive attributes, and no human subjects were involved. We view machine unlearning as a contribution to AI safety and data governance because it enables the removal of unsafe, outdated, or private content from deployed models. We also used ChatGPT 5 for minor language and grammatical edits; all research design, analysis, and interpretation were conducted by the authors.

Acknowledgments

The work was supported by a grant, provided by the Ministry of Economic Development of the Russian Federation in accordance with the subsidy agreement (agreement identifier 000000C313925P4G0002) and the agreement with the Ivannikov Institute for System Programming of the Russian Academy of Sciences dated June 20, 2025 No. 139-15-2025-011.

References

Yinzhi Cao and Junfeng Yang. 2015. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pages 463–480. IEEE.

Sungmin Cha, Sungjun Cho, Donggyu Hwang, and Moontae Lee. 2025. Towards robust and parameter-efficient knowledge unlearning for LLMs. In *The Thirteenth International Conference on Learning Representations*.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, and 1 others. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature machine intelligence*, 5(3):220–235.

Yijiang River Dong, Hongzhou Lin, Mikhail Belkin, Ramon Huerta, and Ivan Vulić. 2025. Undial: Self-distillation with adjusted logits for robust unlearning in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8827–8840.

Alexey Dontsov, Dmitrii Korzh, Alexey Zhavoronkin, Boris Mikheev, Denis Bobkov, Aibek Alanov, Oleg Rogov, Ivan Oseledets, and Elena Tutubalina. 2025. CLEAR: Character unlearning in textual and visual modalities. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20582–20603, Vienna, Austria. Association for Computational Linguistics.

Jiahui Geng, Qing Li, Herbert Woisetschlaeger, Zongxiong Chen, Fengyu Cai, Yuxia Wang, Preslav Nakov, Hans-Arno Jacobsen, and Fakhri Karray. 2025. A comprehensive survey of machine unlearning techniques for large language models. *arXiv preprint arXiv:2503.01854*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Shengyuan Hu, Yiwei Fu, Zhiwei Steven Wu, and Virginia Smith. 2024. Unlearning or obfuscating? jogging the memory of unlearned llms via benign relearning. *arXiv preprint arXiv:2406.13356*.

Baixiang Huang, Canyu Chen, Xiongxiao Xu, Ali Payani, and Kai Shu. 2025. Can knowledge editing really correct hallucinations? In *The Thirteenth International Conference on Learning Representations*.

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

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. 2023. Towards unbounded machine unlearning. *Advances in neural information processing systems*, 36:1957–1987.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, and 1 others. 2024. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*.

Bo Liu, Qiang Liu, and Peter Stone. 2022. Continual learning and private unlearning. In *Conference on Lifelong Learning Agents*, pages 243–254. PMLR.

Yezi Liu and 1 others. 2025. LUNE: Efficient LLM unlearning via LoRA fine-tuning with negative examples. In *Socially Responsible and Trustworthy Foundation Models Workshop at NeurIPS 2025*.

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

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

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822.

Alessandro Mantelero. 2013. The eu proposal for a general data protection regulation and the roots of the â right to be forgottenâ. *Computer Law & Security Review*, 29(3):229–235.

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. 2021. Remember what you want to forget: Algorithms for machine unlearning. *Advances in Neural Information Processing Systems*, 34:18075–18086.

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

Nianwen Si, Hao Zhang, Heyu Chang, Wenlin Zhang, Dan Qu, and Weiqiang Zhang. 2023. Knowledge unlearning for llms: Tasks, methods, and challenges. *arXiv preprint arXiv:2311.15766*.

Xianghui Sun, Yunjie Ji, Baochang Ma, and Xiangang Li. 2023. A comparative study between full-parameter and lora-based fine-tuning on chinese instruction data for instruction following large language model. *arXiv preprint arXiv:2304.08109*.

Lingzhi Wang, Xingshan Zeng, Jinsong Guo, Kam-Fai Wong, and Georg Gottlob. 2025. Selective forgetting: Advancing machine unlearning techniques and evaluation in language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 843–851.

Xiaoyu Xu, Xiang Yue, Yang Liu, Qingqing Ye, Haibo Hu, and Minxin Du. 2025. Unlearning isn't deletion: Investigating reversibility of machine unlearning in llms. *arXiv preprint arXiv:2505.16831*.

Hongbang Yuan, Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2025. Towards robust knowledge unlearning: An adversarial framework for assessing and improving unlearning robustness in large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25769–25777.

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

A Machine Unlearning: Definition and Algorithms

A.1 Problem Definition

MU aims to remove the influence of specific data from a trained model without retraining it from scratch. Given a language model f_θ with parameters θ and a dataset $\mathcal{D} = \{(q, a)\}$ of question–answer pairs, the goal is to make the model forget the knowledge contained in a subset of data, while keeping its general capabilities intact. Formally, we define two disjoint subsets:

- the **forget set** $\mathcal{D}_f \subset \mathcal{D}$ containing samples that should be unlearned;
- the **retain set** $\mathcal{D}_r = \mathcal{D} \setminus \mathcal{D}_f$ containing samples whose knowledge should be preserved.

The desired outcome is a new model $f_{\hat{\theta}}$ such that

$$f_{\hat{\theta}}(q) \not\approx a, \quad \forall (q, a) \in \mathcal{D}_f,$$

$$f_{\hat{\theta}}(q) \approx a, \quad \forall (q, a) \in \mathcal{D}_r.$$

In other words, the model should selectively erase information about the forget set while maintaining its performance on the retain set.

A.2 Task Formulation

The unlearning process can be expressed as a selective optimization problem that updates model parameters from θ to $\hat{\theta}$ by jointly enforcing two objectives: (i) degrade performance on the forget set, and (ii) preserve performance on the retain set. This trade-off can be formalized as

$$\min_{\theta} \Big[- \mathbb{E}_{(q,a)\in\mathcal{D}_f} L(a|q;\theta) + \lambda \, \mathbb{E}_{(q,a)\in\mathcal{D}_r} L(a|q;\theta) \Big], \tag{1}$$

where L is a loss function such as the negative log-likelihood, and λ controls the balance between forgetting and retention. This unified objective underlies most gradient-based MU methods and defines the optimization setting used in our experiments.

A.3 Unlearning Algorithms

Gradient Ascent (GA) (Jang et al., 2022). A basic approach that reverses standard training by maximizing the loss on the forget set $\mathcal{D}_f$. It can be written as

$$\mathcal{L}^{\text{GA}}(\theta) = -\mathcal{L}_{\text{NLL}}(\mathcal{D}_f, \theta), \tag{2}$$

where

$$\mathcal{L}_{\text{NLL}}(\mathcal{D}, \theta) = \tfrac{1}{|\mathcal{D}|} \sum_{(q,a)\in\mathcal{D}} -\log P_\theta(a|q).$$

Gradient Difference (GD) (Liu et al., 2022). An extension of GA that also preserves knowledge in the retain set. It increases loss on $\mathcal{D}_f$ while decreasing it on $\mathcal{D}_r$:

$$\mathcal{L}^{\text{GD}}(\theta) = -\mathcal{L}_{\text{NLL}}(\mathcal{D}_f, \theta) + \lambda\,\mathcal{L}_{\text{NLL}}(\mathcal{D}_r, \theta). \tag{3}$$

Negative Preference Optimization (NPO) (Zhang et al., 2024). This method frames unlearning as a preference optimization problem, penalizing the model for assigning a higher likelihood to the forgotten answers relative to a reference model θ_{ref}:

$$\mathcal{L}^{\text{NPO}}(\theta) = \tfrac{2}{\beta}\,\mathbb{E}_{(q,a)\in\mathcal{D}_f}\left[\log\left(1+\left(\frac{P_\theta(a|q)}{P_{\theta_{\text{ref}}}(a|q)}\right)^{\beta}\right)\right], \tag{4}$$

where $\beta = 1$ is used as recommended in the original formulation.

Each algorithm represents a different trade-off between erasing specific knowledge and preserving the rest, forming the basis for our comparative study in Section 4.

B Additional Experiments

B.1 Full-parameter SFT vs. LoRA SFT

We compare LoRA-based SFT (used throughout the paper) against full-parameter SFT to rule out the possibility that LoRA artificially facilitates unlearning. In both settings, the unlearning step uses LoRA. Tables 3 and 4 report ROUGE-L before and after unlearning at learning rates 2×10^{-5}, 4×10^{-5}, and 5×10^{-6} for the city-forget split, and at 2×10^{-5} for multi-domain data.

On the city-forget split (Table 3), full-parameter SFT shows qualitatively similar trends to LoRA SFT: higher learning rates produce stronger forgetting. However, forgetting is consistently less effective. For example, NPO at $lr = 2\times10^{-5}$ achieves a forget ROUGE-L of 0.997 on popular facts under full-parameter SFT, compared to 0.364 under LoRA SFT. On multi-domain data (Table 4), all three algorithms applied to the full-parameter SFT model result in near-complete output degradation, whereas LoRA SFT retains near-baseline retention throughout (see Table 1 and Figure 3). *Conclusion.* LoRA SFT followed by LoRA unlearning is strictly more stable than full-parameter SFT: it achieves stronger forget-set reduction at the same learning rates and avoids the catastrophic multi-domain collapse observed in the full-parameter setting.

B.2 Forget size effect.

Both $N = 1\%$ (Figure 5) and $N = 10\%$ (Figure 6) follow the same qualitative trends as $N = 5\%$ (Figure 3). At $N = 1\%$ the unlearning signal is weaker, so the forget set is harder to erase: ROUGE-L decreases more slowly and typically requires higher learning rates or additional epochs to match the $N = 5\%$ effect, while retention is comparatively stable. At $N = 10\%$ the signal is stronger, so forgetting progresses faster than at $N = 5\%$ and at lower learning rates, with collateral degradation on the retain set emerging earlier, especially for the Pretrained model. These tendencies hold for both rare and popular subsets.

B.3 LLM as a Judge.

Table 5 reports Gemma-7B results combining ROUGE-L with an LLM-as-a-Judge evaluation (DeepSeek v1) that scores each answer on *Accuracy* (factual alignment with the reference) and *Fluency* (linguistic naturalness). LLaMA-3.1-8B results are in Table 1 in the main paper.

Judge Accuracy tracks ROUGE-L closely: configurations with lower forget ROUGE are consistently rated as less accurate, and the rare/popular asymmetry in automatic metrics is reflected in model-based judgments. Fluency remains stable across algorithms, confirming that observed drops are driven by factual removal rather than language degradation.

B.4 General LLM Benchmarks Evaluation.

Figure 7 shows that all three unlearning algorithms (GradAscent, GradDiff, NPO) preserve general model capabilities throughout the unlearning process. Across both rare and popular forget sets and all tested learning rates, deviations from the pretrained or SFT baseline on MMLU and HellaSwag remain within 3%. Neither metric exhibits a consistent downward trend as the learning rate increases, confirming that the forgetting signal does

Table 2: Examples of DUET question–answer pairs with popularity annotations.

Question	Answer	Pop. score	Label
What is the founded by Welch's?	Thomas Bramwell Welch	44	rare
What is the instance of Viltolarsen?	antisense oligonucleotide	47	rare
What is the subclass of cowpunk?	punk music	47	rare
What is the capital of Poland?	Warsaw	2435	popular
What is the highest judicial authority of Australia?	High Court of Australia	2443	popular
What is the country of Cayenne?	France	2445	popular

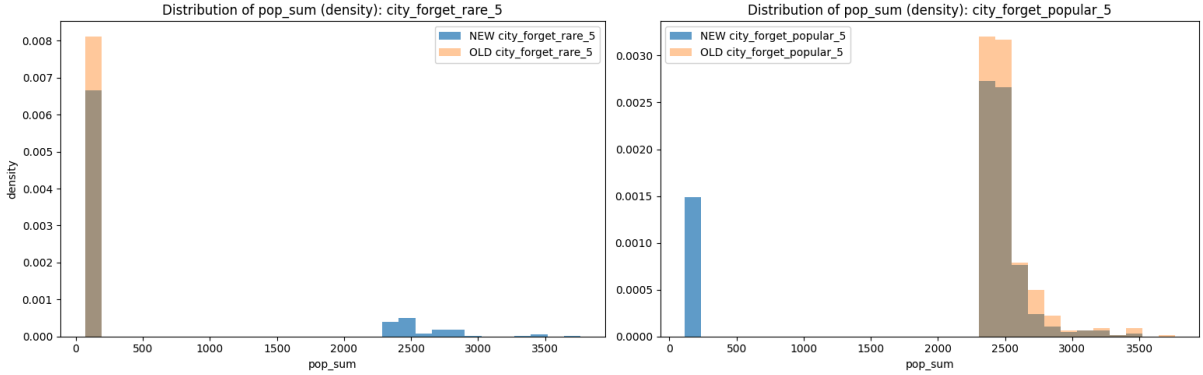


Figure 4: Distributions of popularity scores (*pop_sum*) for the *city* domain in the new and old benchmark splits. The left panel shows the *rare forget 5%* subset and the right panel the *popular forget 5%* subset. Of 482 samples, 87 differ from the previous version, yielding an overall overlap of 81.95%.

Table 3: Full-parameter SFT with LoRA unlearning on the city-forget split. ROUGE-L is reported before and after unlearning.

Algorithm	LR	Popularity	ROUGE-L Before	ROUGE-L After
GradAscent	2e-5	popular	0.998	0.439
GradAscent	4e-5	popular	0.998	0.008
GradAscent	5e-6	popular	0.998	0.998
GradAscent	2e-5	rare	0.994	0.991
GradAscent	4e-5	rare	0.994	0.369
GradAscent	5e-6	rare	0.994	0.993
GradDiff	2e-5	popular	0.998	0.821
GradDiff	4e-5	popular	0.998	0.044
GradDiff	5e-6	popular	0.998	0.998
GradDiff	2e-5	rare	0.994	0.993
GradDiff	4e-5	rare	0.994	0.942
GradDiff	5e-6	rare	0.994	0.993
NPO	2e-5	popular	0.998	0.997
NPO	4e-5	popular	0.998	0.214
NPO	5e-6	popular	0.998	0.998
NPO	2e-5	rare	0.994	0.987
NPO	4e-5	rare	0.994	0.312
NPO	5e-6	rare	0.994	0.994

not propagate to general knowledge. Pretrained and SFT variants behave similarly on these benchmarks, which stands in contrast to their divergent behavior on the task-specific retain set, further supporting the conclusion that capability degradation during unlearning is localized to the domain being unlearned rather than being a general phenomenon.

Table 4: Full-parameter SFT with LoRA unlearning on the multi-domain split at $lr = 2 \times 10^{-5}$. All algorithms show near-complete output collapse, in contrast to LoRA SFT (Table 1).

Algorithm	LR	Popularity	ROUGE-L Before	ROUGE-L After
GradAscent	2e-5	popular	0.988	0.000
GradAscent	2e-5	rare	0.964	0.000
GradDiff	2e-5	popular	0.988	0.001
GradDiff	2e-5	rare	0.964	0.001
NPO	2e-5	popular	0.988	0.144
NPO	2e-5	rare	0.964	0.046

B.5 Intrinsic memorization analysis

We analyze intrinsic memorization signals to understand how rare and popular facts are internally represented and how unlearning affects them. We focus on two complementary diagnostics: (i) token-level probability and rank shifts of the gold answer, and (ii) hidden-state similarity across model layers. All analyses contrast pretrained and SFT models and explicitly separate rare and popular facts.

Token probability and rank shifts. We measure changes in the conditional probability $\log P_\theta(a \mid q)$ of the gold answer token and its rank in the output distribution. Table 6 reports results for both GradAscent (GA) and NPO.

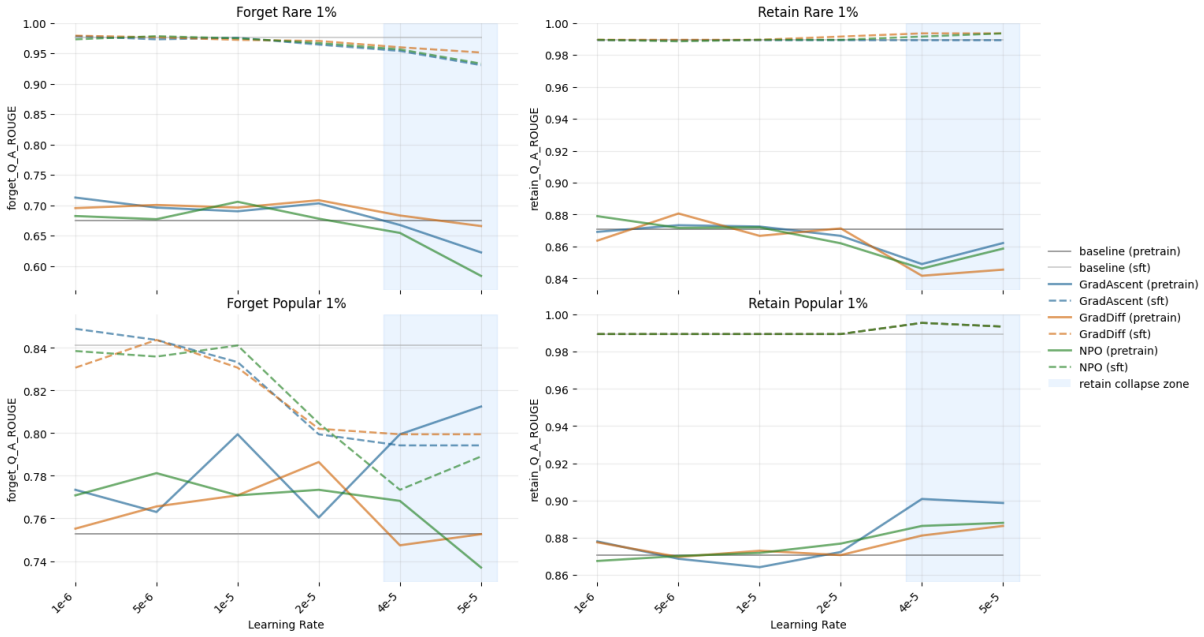


Figure 5: City, $N = 1\%$. Top: rare; bottom: popular. Left: forget (`city_forget_{rare,popular}_1`); right: retain (`city_fast_retain_500`). ROUGE is reported across learning rates. Pretrain is shown as a solid line with circles; SFT as a dashed line with stars. Baselines are solid horizontal lines.

Train type	Algo	Pop.	ROUGE-L Forget ↓	ROUGE-L Retain ↑	Judge Acc. Forget ↓	Judge Acc. Retain ↑	Judge Flu. Forget ↑	Judge Flu. Retain ↑
Pretrain	w/o	rare	0.515	0.445	0.212	0.562	0.973	0.990
SFT	w/o	rare	0.973	0.923	0.770	0.903	0.952	0.915
Pretrain	GA	rare	0.267 (−0.25)	0.148 (−0.30)	0.369 (+0.16)	0.269 (−0.29)	0.546 (−0.43)	0.830 (−0.16)
SFT	GA	rare	0.337 (−0.64)	0.869 (−0.05)	0.316 (−0.45)	0.801 (−0.10)	0.765 (−0.19)	0.820 (−0.10)
Pretrain	GD	rare	0.421 (−0.09)	0.633 (+0.19)	0.149 (−0.06)	0.645 (+0.08)	0.620 (−0.35)	0.877 (−0.11)
SFT	GD	rare	0.342 (−0.63)	0.893 (−0.03)	0.325 (−0.44)	0.843 (−0.06)	0.823 (−0.13)	0.839 (−0.08)
Pretrain	NPO	rare	0.472 (−0.04)	0.643 (+0.20)	0.291 (+0.08)	0.635 (+0.07)	0.749 (−0.22)	0.909 (−0.08)
SFT	NPO	rare	0.298 (−0.68)	0.873 (−0.05)	0.318 (−0.45)	0.818 (−0.08)	0.810 (−0.14)	0.822 (−0.09)
Pretrain	w/o	pop	0.698	0.445	0.561	0.562	0.992	0.990
SFT	w/o	pop	0.994	0.923	0.907	0.903	0.932	0.915
Pretrain	GA	pop	0.032 (−0.67)	0.110 (−0.34)	0.148 (−0.41)	0.224 (−0.34)	0.747 (−0.24)	0.838 (−0.15)
SFT	GA	pop	0.000 (−0.99)	0.315 (−0.61)	0.190 (−0.72)	0.522 (−0.38)	0.297 (−0.64)	0.606 (−0.31)
Pretrain	GD	pop	0.141 (−0.56)	0.242 (−0.20)	0.208 (−0.35)	0.408 (−0.15)	0.880 (−0.11)	0.853 (−0.14)
SFT	GD	pop	0.965 (−0.03)	0.914 (−0.01)	0.886 (−0.02)	0.864 (−0.04)	0.806 (−0.13)	0.853 (−0.06)
Pretrain	NPO	pop	0.429 (−0.27)	0.533 (+0.09)	0.522 (−0.04)	0.574 (+0.01)	0.892 (−0.10)	0.893 (−0.10)
SFT	NPO	pop	0.961 (−0.03)	0.931 (+0.01)	0.887 (−0.02)	0.877 (−0.03)	0.893 (−0.04)	0.827 (−0.09)

Table 5: Gemma-7B at $N = 5\%$, $lr = 2 \times 10^{-5}$. ROUGE-L measures lexical forgetting and retention. Judge Accuracy measures factual alignment; Judge Fluency measures linguistic naturalness (DeepSeek v1). Parentheses show change from the matching *w/o* baseline.

Across both algorithms, unlearning popular facts leads to larger absolute probability shifts than rare facts. However, rank dynamics reveal a qualitatively different behavior: unlearning rare facts induces substantially larger rank changes than popular ones. This effect is most pronounced in the pretrained model, where forgetting rare facts shifts the average rank from single digits to above 200, while popular facts exhibit only mild rank perturbations. In SFT models, probability shifts remain non-trivial, but rank changes become more controlled and less extreme.

Table 6: Token-level probability and rank shifts under unlearning for GradAscent (GA) and NPO. Values are averaged over samples; Δ denotes unlearned minus base.

Alg.	Set	Phase	Pop.	n	$\Delta \log P$	$\text{rank}_{\text{base}}$	rank_{unl}	Δrank
GA	forget	pretrain	popular	482	3.88	34.2	41.8	7.5
GA	forget	pretrain	rare	482	2.07	8.8	230.4	221.6
GA	forget	sft	popular	482	10.94	59.9	55.2	-4.6
GA	forget	sft	rare	482	2.72	111.3	49.7	-61.6
NPO	forget	pretrain	popular	482	0.32	34.2	35.3	1.1
NPO	forget	pretrain	rare	482	0.99	8.8	17.6	8.9
NPO	forget	sft	popular	482	7.99	59.9	56.0	-3.9
NPO	forget	sft	rare	482	2.26	111.3	43.3	-68.0

Rank-based diagnostics reveal a strong asymmetry between rare and popular facts that is not captured by changes in probability alone. Rare facts are significantly more fragile under unlearning, particularly in pretrained models, while SFT stabilizes rank behavior.

Hidden-state similarity. We further analyze how unlearning modifies internal representations by computing hidden-state similarity between base and unlearned models across all layers. Table 7 summarizes cosine similarity and ℓ_2 distance statistics for both GA and NPO.

In pretrained models, hidden states remain highly similar even for facts targeted by unlearning, indicating limited internal adaptation. In contrast, SFT models exhibit well-localized changes in representation. When forgetting popular facts, internal states shift substantially, while rare facts remain

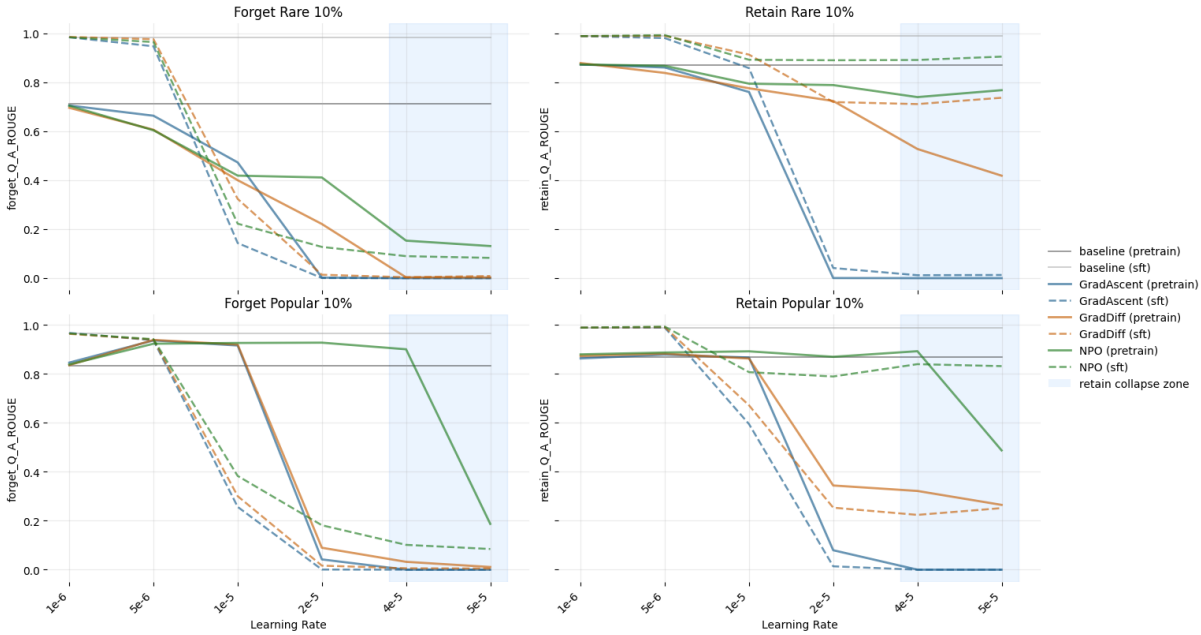


Figure 6: City, $N = 10\%$. Top: rare; bottom: popular. Left: forget (`city_forget_{rare,popular}_1`); right: retain (`city_fast_retain_500`). ROUGE is reported across learning rates. Pretrain is shown as a solid line with circles; SFT as a dashed line with stars. Baselines are solid horizontal lines.

Table 7: Hidden-state similarity between base and unlearned models for GradAscent (GA) and NPO. Metrics are averaged across layers and samples.

Alg.	Phase	Unlearn	Pop.	cosine	ℓ_2	cosine (std)	ℓ_2 (std)
GA	pretrain	pop	popular	0.982	3.35	0.006	0.61
GA	pretrain	rare	rare	0.981	2.70	0.008	0.66
GA	sft	pop	popular	0.958	5.47	0.016	1.01
GA	sft	rare	rare	0.966	3.81	0.011	0.81
NPO	pretrain	pop	popular	0.996	1.52	0.001	0.27
NPO	pretrain	rare	rare	0.990	1.98	0.003	0.36
NPO	sft	pop	popular	0.967	4.82	0.014	0.98
NPO	sft	rare	rare	0.973	3.40	0.009	0.72

largely intact; the reverse occurs when forgetting rare facts. This suggests that SFT enables more selective and controlled modification of internal representations.

Conclusion. SFT produces sharper and more localized forgetting: hidden states shift substantially for the targeted fact type while remaining stable for the other, whereas pretrained models show limited internal adaptation regardless of which facts are unlearned.

Taken together, token-level and hidden-state diagnostics confirm that rare and popular facts respond asymmetrically to unlearning at the representation level. SFT substantially improves the selectivity of forgetting: it concentrates representation changes on the targeted fact type while leaving the other largely intact, an effect absent in pretrained models.

B.6 Distillation-based unlearning (UNDIAL)

To broaden validation beyond SOTA unlearning methods, we also ran a distillation-based forgetting algorithm, Self-Distillation with Adjusted Logits (UNDIAL) (Dong et al., 2025). Unlike gradient updates that directly modify model parameters to suppress target knowledge, UNDIAL performs controlled self-distillation at the token level, aiming to adjust the model's output distribution to reduce reliance on the targeted content while preserving general capabilities. We evaluated UNDIAL on DUET for both `popular` and `rare` forget splits, with pretrained and SFT checkpoints and across multiple learning rates (Table 8).

Overall, UNDIAL exhibits the same qualitative pattern we observe throughout the paper. For SFT models, retention remains very high across learning rates (retain ROUGE ≈ 0.987 to 1.000), while forgetting behaves more smoothly and predictably, especially on the popular split where forget ROUGE decreases as the learning rate increases. In contrast, pretrained models show substantially weaker retention (retain ROUGE $\approx 0.879 - 0.920$) and less stable forgetting dynamics across learning rates, with popular forgetting not improving monotonically. Importantly, the rare vs. popular asymmetry persists under self-distillation, suggesting that the popularity effect is not tied to a specific unlearning mechanism, but reflects a more general interaction

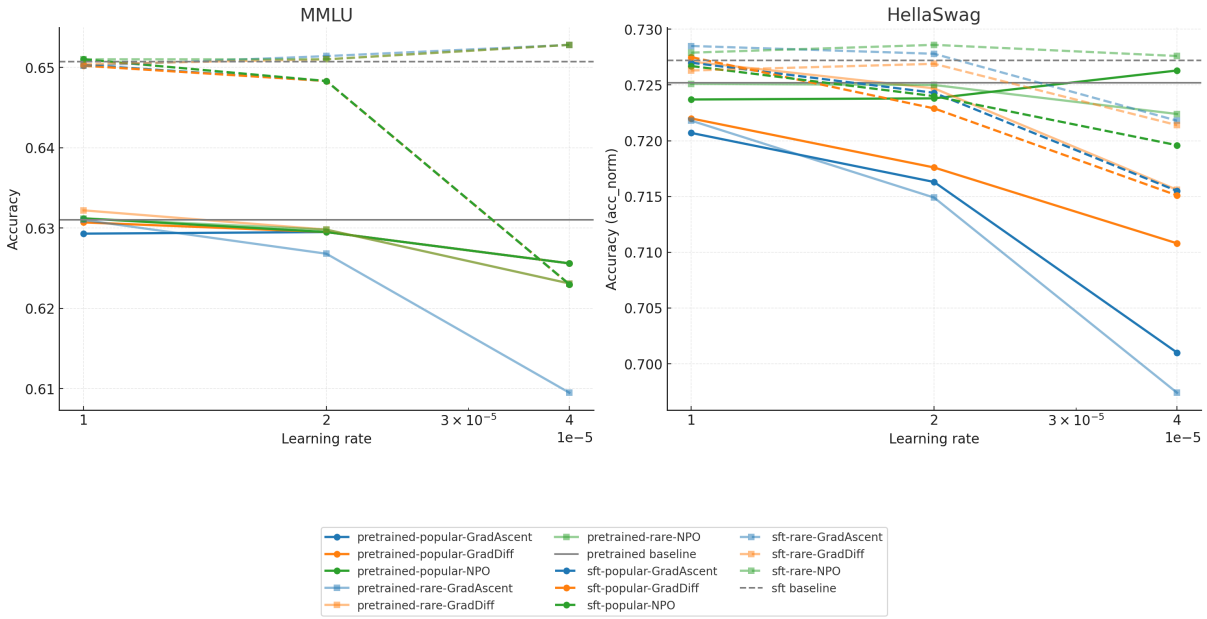


Figure 7: Evaluation on LLM benchmarks. Left: MMLU accuracy. Right: HellaSwag accuracy (acc_norm). Curves show unlearning algorithms (GradAscent, GradDiff, NPO) on rare and popular forget sets across learning rates; pretrained models are solid, SFT models dashed; horizontal lines indicate the corresponding baselines.

Table 8: UNDIAL (self-distillation) unlearning results on DUET across learning rates, evaluated on popular and rare forget splits.

Train type	Algorithm	Forget set	LR	Forget	Retain
SFT	UNDIAL	popular	1e-6	0.931	0.990
SFT	UNDIAL	popular	2e-5	0.865	0.998
SFT	UNDIAL	popular	5e-5	0.823	1.000
SFT	UNDIAL	rare	1e-6	0.988	0.990
SFT	UNDIAL	rare	2e-5	0.648	0.987
SFT	UNDIAL	rare	5e-5	0.679	0.994
pretrain	UNDIAL	popular	1e-6	0.771	0.885
pretrain	UNDIAL	popular	2e-5	0.753	0.890
pretrain	UNDIAL	popular	5e-5	0.780	0.920
pretrain	UNDIAL	rare	1e-6	0.694	0.879
pretrain	UNDIAL	rare	2e-5	0.562	0.884
pretrain	UNDIAL	rare	5e-5	0.608	0.922

between fact popularity and the model's training stage.

Conclusion. UNDIAL results are consistent with the gradient-based findings: SFT models produce more stable and effective forgetting, retain near-baseline knowledge on the retain set, and exhibit the same rare/popular asymmetry. The popularity effect thus generalizes across unlearning mechanisms with distinct optimization objectives.

B.7 Model effect.

For Gemma, we observe a consistent trend across forget sets (Figure 8): rare entities are easier to erase than popular ones, as shown by the smooth exponential decay of ROUGE-L with increasing learning rate. The SFT model, however, exhibits greater sensitivity to the learning rate, with performance dropping sharply from nearly 1.0 to 0 over a narrow range. NPO maintains stable retention across both model types, while GradAscent and GradDiff lead to considerable degradation, especially in the SFT variant (see Table 5 in Appendix B.3). Overall, the SFT model preserves rare-entity knowledge more effectively than the pretrained model, but their response curves differ substantially. This divergence may reflect architectural differences between Gemma and LLaMA, as well as weaker alignment between Wikipedia-based popularity and Gemma's pretraining corpus.

For Qwen 2.5 7B in the Figure 9, the effect observed is more similar to the trends seen in LLaMA -3.1-8B.

Experiments on all three models show general trends, thereby generalizing and reinforcing the conclusions in the main part of the article.

B.8 Multi-domain validation beyond cities.

While DUET is dominated by the *city* domain, this choice was intentional to isolate popularity effects without noise in domain-induced confounders. To verify that our conclusions are not domain-specific, we additionally conducted multi-domain experiments on a domain-balanced subset of DUET, using 500 forget and 500 retain examples across do-

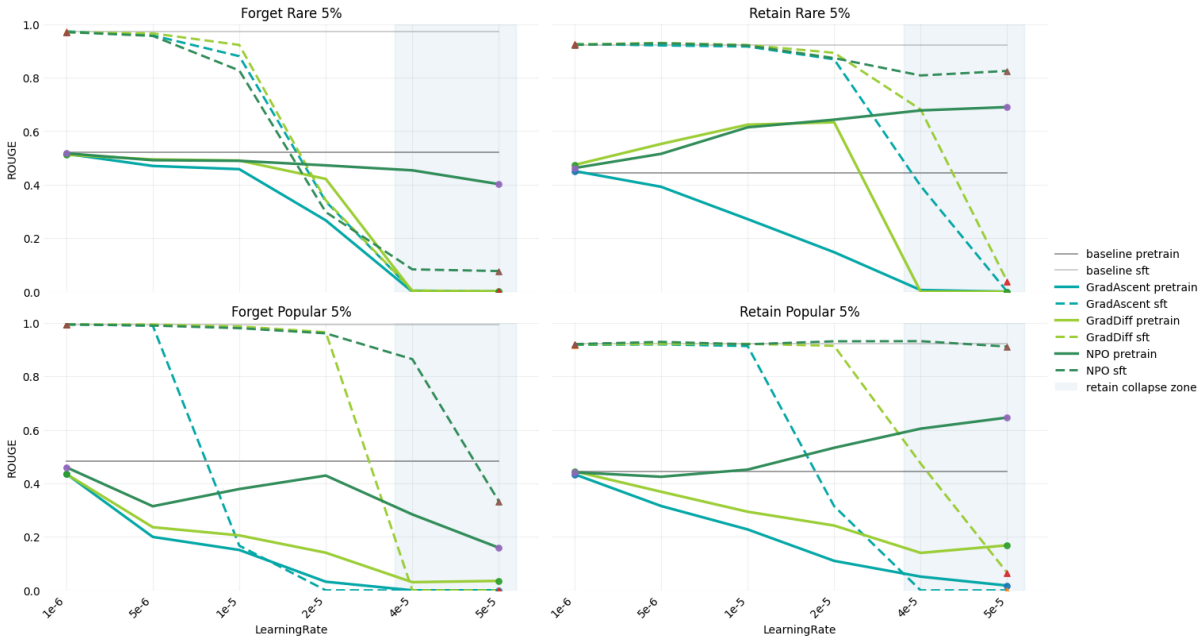


Figure 8: City split ($N = 5\%$) evaluated on Gemma 7B. Top: rare; bottom: popular. Left: forget sets (`city_forget_{rare,popular}_5`); right: retain set (`city_fast_retain_500`). ROUGE scores are shown across learning rates. Pretrained models are plotted as solid lines with circles, and SFT models are plotted as dashed lines with stars. Solid horizontal lines indicate baselines.

mains. Table 9 reports representative results at learning rate 2e-5.

Across all algorithms and training regimes, the same qualitative pattern persists. Forgetting popular facts consistently results in stronger degradation on the forget split compared to rare facts, while retention performance remains higher. This effect is especially clear in SFT models, where rare facts are better preserved during popular-forget training. In pretrained models, the gap between popular and rare forgetting is even larger, confirming that popularity-driven asymmetry generalizes beyond the city domain.

Table 9: Multi-domain unlearning results on DUET subset. Metrics are reported for learning rate $2e{-}5$.

Train type	Algorithm	Forget set	LR	Forget	Retain
SFT	GradAscent	popular	2e-5	0.461	0.736
SFT	GradAscent	rare	2e-5	0.651	0.729
SFT	GradDiff	popular	2e-5	0.477	0.743
SFT	GradDiff	rare	2e-5	0.654	0.754
SFT	NPO	popular	2e-5	0.482	0.741
SFT	NPO	rare	2e-5	0.668	0.731
pretrain	GradAscent	popular	2e-5	0.694	0.608
pretrain	GradAscent	rare	2e-5	0.527	0.485
pretrain	GradDiff	popular	2e-5	0.698	0.645
pretrain	GradDiff	rare	2e-5	0.594	0.588
pretrain	NPO	popular	2e-5	0.715	0.686
pretrain	NPO	rare	2e-5	0.602	0.578

Conclusion. The rare/popular asymmetry is not an artifact of the city-heavy DUET distribution. It holds consistently across domains and training regimes, supporting the claim that fact popularity is a general determinant of unlearning difficulty.

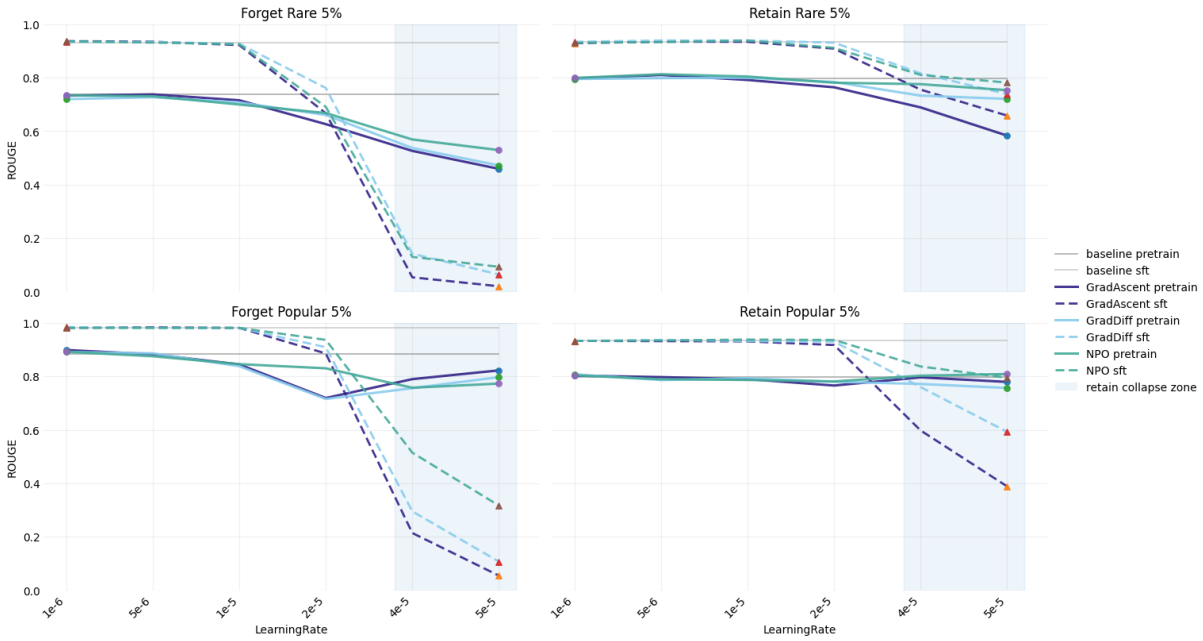


Figure 9: City split ($N = 5\%$) evaluated on Qwen-2.5 7B. Top: rare; bottom: popular. Left: forget sets (`city_forget_{rare,popular}_5`); right: retain set (`city_fast_retain_500`). ROUGE scores are shown across learning rates. Pretrained models are plotted as solid lines with circles, and SFT models are plotted as dashed lines with stars. Solid horizontal lines indicate baselines.