# Can Large Language Models Truly Forget Scientific Claims? A Systematic Evaluation of Claim Level Unlearning

**Snigdha Paul** [1] **Manasi Patwardhan** [1] **Arman Cohan** [2]

## Abstract

Language models (LMs) are pre-trained on static scientific corpora, while scientific knowledge is dynamic and self-correcting. Scientific claims internalized by these models, can be retracted, falsified or revised later by the downstream studies, risking the propagation of incorrect information in scientific research. This motivates the need for LMs to selectively suppress obsolete scientific claims. Machine unlearning can be a natural choice as a solution, preserving overall model utility. Prior work focuses on instance-level forgetting, but scientific claims pose unique challenges due to their structured, multi-faceted, and evolving nature. We introduce a new task of Scientific Claim Unlearning and construct a benchmark SciUnlearn. To evaluate the task, we further demonstrate that existing unlearning methods fail to remove claim-level knowledge, exhibiting surface-level suppression. These findings highlight that scientific claim unlearning is fundamentally more challenging than standard fact unlearning, motivating the need for methods tailored to structured knowledge removal. The benchmark and code are available at https://github.com/snigdhapaul2003/SciUnlearn/.

## 1. Introduction

Large Language Models (LLMs) are increasingly used as scientific assistants (Exler et al., 2026), supporting hypothesis generation (Li et al., 2025; Garikaparthi et al., 2025), literature reviews (Tang et al., 2025), and scientific discovery (Gottweis et al., 2025; Novikov et al., 2025; Lu et al., 2024). However, LLMs are trained on static corpora with fixed knowledge cutoffs, which conflict with the dynamic and self-correcting nature of scientific knowledge. LLMs can internalize scientific claims that are later invalidated by retractions, errors, fraud, or irreproducibility (Lesné et al., 2006), allowing false knowledge to persist in tasks like ideation (Alkan et al., 2025), claim verification (Pradeep et al., 2021; Liu et al., 2024a), more broadly harming in domains like healthcare. LLMs lack mechanisms to deprecate outdated claims, leading to epistemic inertia (Example in Appendix A) despite new contradictory evidences (Caliskan et al., 2016; Gonen & Goldberg, 2019). Another good example is models treating the fabricated disease "Bixonimania" as real (Stokel-Walker, 2026). Additionally, some scientific knowledge later becomes harmful, ethically sensitive, or dual-use (Urbina et al., 2022; Pannu et al., 2025), raising regulatory concerns such as the right to be forgotten (Zhang et al., 2023), yet such knowledge is extremely difficult to remove once embedded in LLMs.

These challenges motivate scientific claim unlearning, where models remove outdated claims. Although unlearning has been studied for personal data, copyrighted content and safety-critical knowledge (Shi et al., 2024; Jiang et al., 2025; Fore et al., 2024), fine-grained unlearning of scientific claims remains largely unexplored. Existing work does not address claim-level unlearning or provide systematic benchmarks (Yang, 2025).

Here, we define the task of scientific claim unlearning (Section 3) and introduce a benchmark with structured forget and retain sets to evaluate claim-level forgetting while preserving valid knowledge (Section 4). We adapt baseline unlearning algorithms to this setting (Section 5) and analyze the result and trade-offs between effective forgetting and scientific knowledge retention (Section 6).

## 2. Related Work

Scientific claim extraction aims to structure knowledge from unstructured scientific text (Park & Blake, 2012; Wei, 2023). More recently, LLM-based pipeline combines extraction, evidence retrieval, and verification (Uprety et al., 2025). Progress has been driven by large annotated datasets across domains (Achakulvisut et al., 2019; Wadden et al., 2020; Diggelmann et al., 2020), and work on large-scale claim tracking and temporal evolution (Wang et al., 2025a; Pra-

[1]TCS Research, India [2]Yale University, United states. Correspondence to: Snigdha Paul <snigdha.paul@tcs.com>.

*AI4Research workshop at $43^{rd}$ International Conference on Machine Learning*, Seoul, South Korea. Copyright 2026 by the author(s).

manick et al., 2026). However, existing research primarily focuses on identifying and organizing claims, with less attention to unlearning claims once they become invalid.

Machine unlearning aims to remove the influence of specific training data while preserving model utility, approximating retraining without the removed data (Bourtoule et al., 2019), and has been extended to LLMs that may memorize harmful content (Liu et al., 2024b). Prior work covers privacy-driven removal, copyrighted knowledge, and evaluates unlearning using benchmark datasets such as TOFU (Maini et al., 2024), MUSE (Shi et al., 2024), WMDP (Li et al., 2024) and RWKU (Jin et al., 2024), establishing the forget–retain paradigm. Algorithmic approaches span gradient-based updates (Jin et al., 2025), preference optimization (Zhang et al., 2024), parameter-efficient methods (Liu et al., 2025; Abitante et al., 2026), and regularization or distillation techniques (Siddiqui et al., 2025; Wang et al., 2025b). However, evidence shows that removed knowledge can persist in representations and re-emerge across contexts (Jia et al., 2025; Rybak et al., 2026). So the focus of this work is to systematically evaluate claim level unlearning, where claims are interconnected, evolving and difficult to unlearn.

## 3. Scientific Claim Unlearning Task Definition

Let $\mathcal{M}_\theta$ denote a pretrained language model with parameters $\theta$. We assume availability of dataset $\mathcal{D} = \{\mathcal{F}_1, \mathcal{F}_2, \mathcal{R}_{ext}, \mathcal{R}_{int}\}$, where $\mathcal{F}_1 \cap \mathcal{F}_2 = \emptyset$, are forget sets, consists of distinct types of paraphrased question answer pairs $(q, a)$ derived from a shared set of underlying scientific claims $\mathcal{S}$, extracted from scientific papers $\mathcal{P}$ that are included in the pre-training corpus of $\mathcal{M}_\theta$. This enables evaluation of *claim level unlearning*: when a model is unlearned on $\mathcal{F}_1$, it is expected to also forget $\mathcal{F}_2$, and vice-versa, thereby testing whether the underlying claim has been removed rather than only specific superficial forms. The retain sets are designed to validate the preservation of non-target knowledge such that, $(\mathcal{R}_{ext} \cup \mathcal{R}_{int}) \cap (\mathcal{F}_1 \cup \mathcal{F}_2) = \emptyset$. The external retain set $\mathcal{R}_{ext}$ consists of cited prior or supporting knowledge from papers referenced by the anchor papers $\mathcal{P}$, while the internal retain set $\mathcal{R}_{int}$ contains background or contextual information drawn from the same anchor papers but unrelated to the target claims.

Given the base model $\mathcal{M}_\theta$ and a forget set $\mathcal{F}_i$ where $i = 1$ or 2, an unlearning algorithm $\mathcal{A}$ updates the model parameters as $\theta' = \theta + \mathcal{A}(\theta, \mathcal{F}_i, \mathcal{R})$ and $\mathcal{R} \subseteq \{\mathcal{R}_{ext}, \mathcal{R}_{int}\}$. Utilization of $\mathcal{R}$ is optional depending on the algorithm $\mathcal{A}$. Building upon the standard formulation of LLM unlearning (Liu et al., 2024b), the objective is defined by the Equation 1.

$$\min_{\theta'} \underbrace{\mathbb{E}_{(q,a)\in\mathcal{F}_i}\left[\ell_f(a|q;\theta')\right]}_{\text{Forget}} + \lambda \underbrace{\mathbb{E}_{(q,a)\in\mathcal{R}}\left[\ell_r(a|q;\theta')\right]}_{\text{Retain}} \quad (1)$$

where $\ell_f(a|q;\theta') = \log p_{\theta'}(a|q)$ denotes the forgetting loss, and $\ell_r(a|q;\theta') = -\log p_{\theta'}(a|q)$ denotes the retention loss. The parameter $\lambda \geq 0$ balances forgetting and retention. To evaluate whether unlearning generalizes, performance is measured on a disjoint forget set as $\mathbb{E}_{(q,a)\in\mathcal{F}_j}[\ell_f(a \mid q;\theta')]$, $j \neq i$., where $\mathcal{F}_i \cap \mathcal{F}_j = \emptyset$. This formulation evaluates claim-level forgetting beyond suppressing specific QA pairs, highlighting the need for more advanced scientific claim unlearning algorithms.

## 4. SciUnlearn Dataset

To study scientific claim unlearning, we build a benchmark dataset $\mathcal{D}$ from the Dolma corpus ($\mathcal{C}$) (Soldaini et al., 2024), which includes 38.8M papers from Semantic Scholar (Kinney et al., 2023). Dolma's scale, recency, and use in training OLMO models ensure the targeted claims are likely embedded in pretrained models, enabling meaningful evaluation. We select a subset of papers from $\mathcal{C}$ and target unlearning of their claims, assuming that success on arbitrary papers generalizes to explicitly falsified or retracted scientific findings. The full pipeline is shown in Figure 1.

We sample 100 recent (2018–2022) open-access computer science papers from the Dolma corpus using the Semantic Scholar API (Kinney et al., 2023), forming $\mathcal{C}_f \subset \mathcal{C}$. For each paper $p \in \mathcal{C}_f$, we extract 1–5 paraphrased claims $\mathcal{S}(p)$ using GPT-5[1] from the abstract, introduction, and conclusion, yielding a total of 219 claims (Prompt in Appendix E.1). We use an *LLM as a judge* to evaluate the extracted claims from all forget and retain set papers based on accuracy, independence, and clarity on a 0–5 scale, using GPT-4o[1] (prompt in Appendix F). As shown in Appendix C Figure 4, score distributions are skewed toward higher values, indicating high-quality claim extraction. To validate LLM judgments, we correlate LLM scores with expert annotations on claims from 2×10 randomly selected papers (Appendix G). Two computer science researchers evaluated claims using the same criteria and scale. Weighted Cohen's kappa shows moderate agreement for accuracy ($\kappa_{\text{acc}} = 0.535$) and good agreement for independence ($\kappa_{\text{ind}} = 0.674$) and clarity ($\kappa_{\text{clr}} = 0.734$), supporting LLM-based evaluation as a scalable proxy.

For each claim $s \in \mathcal{S}(p)$, we generate eight $(q, a)$ pairs—two each for MCQ, true/false(T/F), fill-in-the-blank(FB), and assertion–reason(AR), yielding 1,752 QA pairs (prompt in Appendix E.2). QA pairs are validated using OLMO-3-7B-Instruct ($\mathcal{M}$) with self-consistency checks (jin et al., 2024; Xie et al., 2025). Responses are evaluated using exact match, ROUGE (Lin, 2004), and semantic similarity (Reimers & Gurevych, 2019), with a threshold of 0.7

[1]Models are accessed via API with temperature = 1.0, top-p = 0.9, and a maximum input length of 40,000 characters.

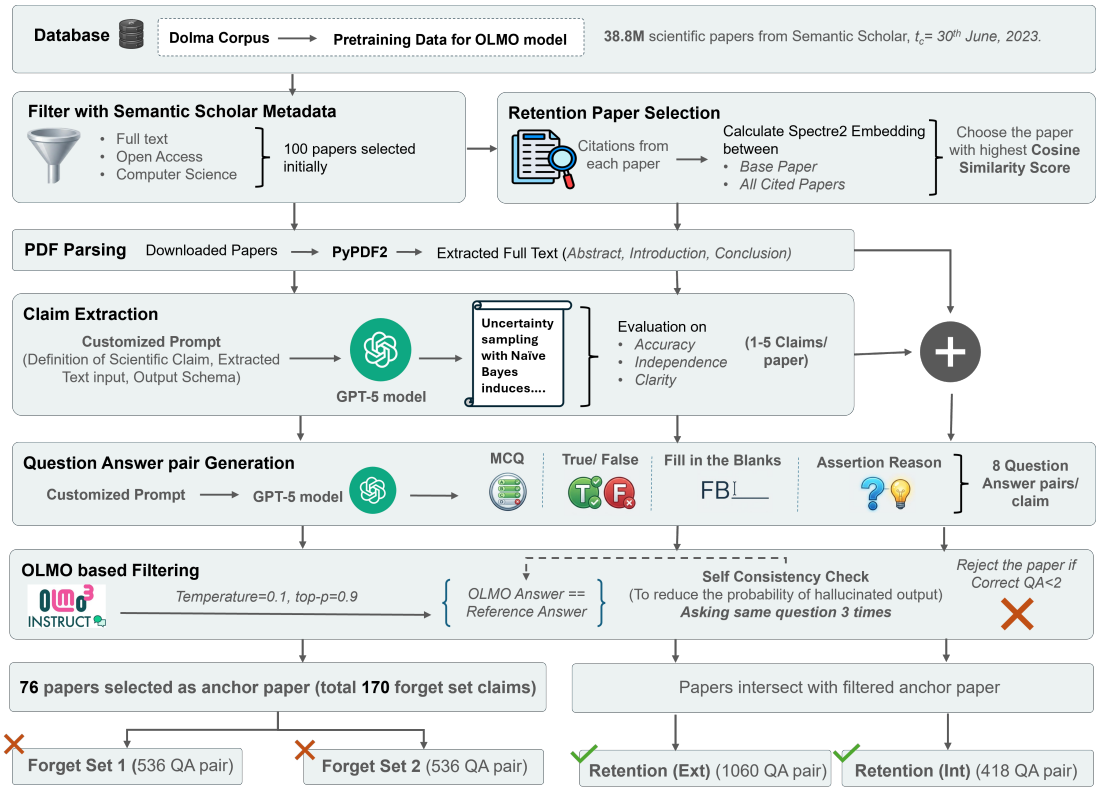



*Figure 1.* SciUnlearn Dataset generation pipeline.

for non-exact matches. Claims with fewer than two valid question types are discarded. This results in 1,072 QA pairs from 170 claims across 76 anchor papers ($\mathcal{P}$). QA pairs are splitted into $\mathcal{F}_1$ and $\mathcal{F}_2$, enabling cross-generalization evaluation of unlearning.

*Table 1.* Overview of the SciUnlearn dataset

| Statistic | F1 | F2 | $R_{ext}$ | $R_{int}$ |
|---|---|---|---|---|
| Claims | 170 | 170 | 171 | – |
| QA Pairs | 536 | 536 | 1060 | 418 |

The retain sets serve complementary roles: $\mathcal{R}_{ext}$ tests whether unlearning remains localized by preserving related knowledge from cited papers, while $\mathcal{R}_{int}$ ensures contextual knowledge within the same paper is retained. For $\mathcal{R}_{ext}$, we select for each anchor paper the most semantically similar cited paper using SPECTER2 embeddings (Singh et al., 2023), extracting claims and QA pairs to obtain 1,060 QA pairs from the anchor papers. $\mathcal{R}_{int}$ is built from the same anchor papers using non–claim-specific, contextual QA pairs, yielding 418 from an initial 608 after filtering. We ensure no overlap between forget and retain sets (Appendix I). Figure 2 shows an example, with full QA details provided in Appendix H. Final statistics of SciUnlearn dataset are presented in Table 1. More information and other analysis on the Dataset is provided in Appendix B and Appendix D respectively.

## 5. Experimental Set-up

We use OLMo-3-7B-Instruct, which supports full-parameter and LoRA training (rank 8) on a single 80GB A100 GPU while enabling evaluation of diverse unlearning methods. We benchmark representative approaches spanning gradient-based and preference-based methods, including Gradient Difference (GD) (Neel et al., 2020), Negative Preference Optimization (NPO) (Zhang et al., 2024), NPO with retain objective (NPO+RT) (Bronec & Helcl, 2025), SimNPO (Fan et al., 2024), and SimNPO with a retain set. All methods are applied to the pretrained model using forget sets $\mathcal{F}_1$ or $\mathcal{F}_2$. GD performs ascent on forget data and descent on retain data, NPO suppresses forget samples via preference optimization, NPO+RT improves retention, and SimNPO directly penalizes forget samples without a reference model. Implementation details and hyperparameters are provided in Appendix J and Appendix K. Task-specific performance is evaluated using Exact Match (EM) and ROUGE-L F1 (R-F1) (Lin, 2004), capturing strict and paraphrase-tolerant correctness, respectively. General knowledge retention is assessed on MMLU, ARC-Challenge (AC), and HellaSwag (HS) (Hendrycks et al., 2020; Clark et al., 2018; Zellers et al., 2019), ensuring unlearning does not degrade overall reasoning and language understanding.

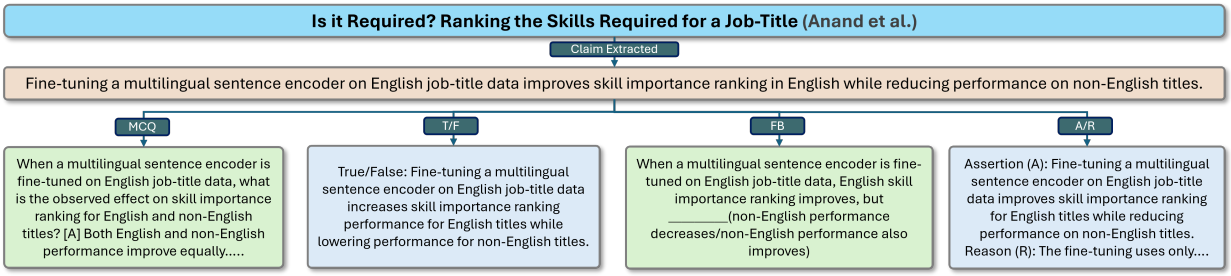



*Figure 2.* Example claim and QA in SciUnlearn dataset.

*Table 2.* Results: LoRA unlearning with $\mathcal{F}_1$ and $\mathcal{F}_2$, ↓: lower is better (forgetting), and ↑: higher is better (retention and generalization), **Best** and **Second Best**, separately done for both Forget Sets

| Method | Forget Set 1 ↓ | | Forget Set 2 ↓ | | Retain (Ext) ↑ | | Retain (Int) ↑ | | General Benchmarks | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | R-F1 | EM | R-F1 | EM | R-F1 | EM | R-F1 | EM | MMLU | Arc Challenge | HellaSwag |
| Base Model | 97.54 | 94.21 | 93.92 | 89.17 | 95.40 | 90.75 | 92.44 | 83.49 | 58.75 | 54.35 | 76.02 |
| GD ($\mathcal{F}_1$) | **71.70** | **58.20** | 93.51 | 88.24 | **93.66** | **84.15** | **90.53** | 72.00 | 56.44 | **50.08** | **70.13** |
| NPO ($\mathcal{F}_1$) | **83.77** | 73.50 | **87.26** | **76.11** | 83.75 | 72.83 | 87.12 | 77.03 | 39.38 | 41.21 | 69.98 |
| NPO+RT ($\mathcal{F}_1$) | 88.81 | 81.52 | 93.67 | 89.55 | **92.44** | **86.13** | **90.66** | **82.05** | **58.28** | **49.91** | 69.89 |
| SimNPO ($\mathcal{F}_1$) | 86.00 | 75.37 | **87.75** | **79.47** | 86.38 | 76.98 | 89.57 | **79.66** | 43.68 | 40.35 | **71.37** |
| SimNPO+RT ($\mathcal{F}_1$) | 83.98 | **69.96** | 94.29 | 89.73 | 92.05 | 82.92 | 89.04 | 77.27 | **58.32** | 48.54 | 69.83 |
| GD ($\mathcal{F}_2$) | **91.65** | **78.35** | 89.60 | 83.58 | **93.27** | 84.52 | 91.02 | 77.03 | **57.54** | 49.57 | 70.93 |
| NPO ($\mathcal{F}_2$) | 94.18 | **81.15** | **73.87** | **54.47** | 84.58 | 67.92 | 87.26 | 77.51 | 27.73 | 41.55 | 70.36 |
| NPO+RT ($\mathcal{F}_2$) | 96.14 | 89.92 | 86.78 | 82.27 | 91.47 | **85.18** | **91.60** | **81.81** | 57.39 | **50.51** | **73.02** |
| SimNPO ($\mathcal{F}_2$) | **92.66** | 81.34 | **78.53** | **62.87** | 86.74 | 72.54 | 90.65 | 81.57 | 37.91 | 41.55 | 71.55 |
| SimNPO+RT ($\mathcal{F}_2$) | 94.99 | 89.55 | 92.31 | 86.75 | **93.23** | **86.32** | **92.33** | **82.77** | **58.25** | **50.34** | **73.44** |

## 6. Results

Tables 2 presents the performance of different unlearning methods in the LoRA setting, whereas Table 7 in Appendix M presents results for full-parameter unlearning.

The results show that forgetting is largely localized to the unlearned forget set, with minimal transfer to the other set. When unlearning is applied to $\mathcal{F}_1$ or $\mathcal{F}_2$ (Tables 2, methods such as GD, NPO, and SimNPO significantly reduce performance on the targeted set leaving the counterpart largely unaffected, even with retain-set fine-tuning. Qualitative analysis (Appendix N) indicates that simpler QA formats (e.g., T/F and AR) and higher numbers of paraphrased QAs, and non deeply rooted papers amplify forgetting. However, none of the methods cause sharp performance drops, largely due to LoRA-based fine-tuning, consistent with prior findings (jin et al., 2024). Overall, forgetting remains subset-specific, suggesting current methods suppress surface-level patterns rather than underlying claim-level knowledge.

Retention differs across methods: GD, NPO+RT, and SimNPO+RT preserve strong performance on both external and internal retain sets, while NPO and SimNPO show larger retention drops due to the absence of an explicit retain objective. This indicates that retain-aware methods better protect non-target knowledge; qualitative analysis further suggests retention sensitivity is driven more by QA fragility than by the specific forget set (Appendix N.4). General benchmarks are affected unevenly: NPO and SimNPO without retention cause notable MMLU degradation, while GD, NPO+RT, and SimNPO+RT largely preserve performance.

**Can LLMs truly forget Scientific Claims?** The results show that existing methods do not achieve true claim-level forgetting: they suppress performance on specific training instances. Only some papers transfer unlearning to the complementary forget set, that also superficially, typically only those papers that are not deeply rooted in the literature. This phenomena is quantified in the Appendix N.5.

## 7. Conclusion

This work introduces Scientific Claim Unlearning, along with the SciUnlearn benchmark, and evaluates existing unlearning methods on this task. Results show that current approaches fail to achieve true claim-level forgetting, exhibiting asymmetric forgetting that reflects surface-level suppression rather than removal of semantic knowledge. While SciUnlearn focuses on computer science claims, it opens directions for extending to domains with explicitly retracted claims and for exploring representation-based and mechanistic unlearning methods. Overall, our findings highlight the need for new algorithms and evaluation protocols that target structured, conceptual scientific knowledge.

## Impact Statement

The failure of AI models to unlearn retracted or falsified scientific claims has particularly serious implications, as science advances through iterative correction and refinement of knowledge. When models are pre-trained on datasets containing invalidated or retracted findings, they risk preserving and propagating outdated information, thereby effectively freezing scientific understanding and compromising the integrity of the knowledge encoded within their parameters. As a consequence, discredited theories may acquire artificial persistence and continued influence, potentially misdirecting AI-assisted research and hindering reliable scientific discovery. To the best of our knowledge, this is among the first efforts to systematically highlight this problem and to conduct preliminary investigations into the effectiveness of existing machine unlearning approaches in this context. Our findings indicate that current methods often achieve only superficial, surface-level suppression of such knowledge, leaving deeper representations largely intact. In conclusion, these observations underscore a critical and unresolved challenge of developing robust, principled approaches for deeply and reliably unlearning falsified or retracted scientific claims that are intricately embedded within the large-scale corpora used to pre-train modern AI systems.

## References


Abitante, J. V. B., Pasquali, J., Garcia, L. F., de Oliveira, E., Paula, T., Barros, R. C., and Kupssinskü, L. S. Quantization-robust llm unlearning via low-rank adaptation. *ArXiv*, abs/2602.13151, 2026. URL https://api.semanticscholar.org/CorpusID:285607004.

Achakulvisut, T., Bhagavatula, C., Acuna, D., and Kording, K. Claim extraction in biomedical publications using deep discourse model and transfer learning. *arXiv preprint arXiv:1907.00962*, 2019.

Alkan, A. K., Sourav, S., Jabłońska, M., Astarita, S., Chakrabarty, R., Garuda, N., Khetarpal, P., Pi'oro, M., Tanoglidis, D., Iyer, K. G., Polimera, M. S., Smith, M. J., Ghosal, T., Huertas-Company, M., Kruk, S., Schawinski, K., and Ciucua, I. A survey on hypothesis generation for scientific discovery in the era of large language models. *ArXiv*, abs/2504.05496, 2025. URL https://api.semanticscholar.org/CorpusID:277628169.

Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., He, H., Leahy, C., McDonell, K., Phang, J., Pieler, M. M., Prashanth, U. S., Purohit, S., Reynolds, L., Tow, J., Wang, B., and Weinbach, S. Gpt-neox-20b: An open-source autoregressive language model. *ArXiv*, abs/2204.06745, 2022. URL https://api.semanticscholar.org/CorpusID:248177957.

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., and Papernot, N. Machine unlearning. *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 141–159, 2019. URL https://api.semanticscholar.org/CorpusID:208909851.

Bronec, J. and Helcl, J. Atyaephyra at SemEval-2025 task 4: Low-rank negative preference optimization. In Rosenthal, S., Rosá, A., Ghosh, D., and Zampieri, M. (eds.), *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pp. 1415–1422, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-273-2. URL https://aclanthology.org/2025.semeval-1.187/.

Caliskan, A., Bryson, J. J., and Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356:183 – 186, 2016. URL https://api.semanticscholar.org/CorpusID:23163324.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457, 2018. URL https://api.semanticscholar.org/CorpusID:3922816.

Diggelmann, T., Boyd-Graber, J., Bulian, J., Ciaramita, M., and Leippold, M. Climate-fever: A dataset for verification of real-world climate claims, 2020.

Exler, D., Raimann, M., Münker, M. F., Rosin, M., Gómez, J. E. U., Niemeyer, C. M., Reischl, M., and Rettenberger, L. Llm-based scientific assistants for knowledge extraction: Which design choices matter? *Advanced Intelligent Discovery*, 2026. URL https://api.semanticscholar.org/CorpusID:287682689.

Fan, C., Liu, J., Lin, L., Jia, J., Zhang, R., Mei, S., and Liu, S. Simplicity prevails: Rethinking negative preference optimization for llm unlearning. *ArXiv*, abs/2410.07163, 2024. URL https://api.semanticscholar.org/CorpusID:273233165.

Fore, M., Singh, S., Lee, C., Pandey, A., Anastasopoulos, A., and Stamoulis, D. Unlearning climate misinformation in large language models. In Stammbach, D., Ni, J., Schimanski, T., Dutia, K., Singh, A., Bingler, J., Christiaen, C., Kushwaha, N., Muccione, V., A. Vaghefi, S., and Leippold, M. (eds.), *Proceedings of the 1st Workshop on Natural Language Processing*

*Meets Climate Change (ClimateNLP 2024)*, pp. 178–192, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.climatenlp-1.14. URL https://aclanthology.org/2024.climatenlp-1.14/.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., and Leahy, C. The pile: An 800gb dataset of diverse text for language modeling. *ArXiv*, abs/2101.00027, 2020. URL https://api.semanticscholar.org/CorpusID:230435736.

Garikaparthi, A., Patwardhan, M., Vig, L., and Cohan, A. IRIS: Interactive research ideation system for accelerating scientific discovery. In Mishra, P., Muresan, S., and Yu, T. (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 592–603, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-253-4. doi: 10.18653/v1/2025.acl-demo.57. URL https://aclanthology.org/2025.acl-demo.57/.

Gonen, H. and Goldberg, Y. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *North American Chapter of the Association for Computational Linguistics*, 2019. URL https://api.semanticscholar.org/CorpusID:73729169.

Gottweis, J., Weng, W.-H., Daryin, A., Tu, T., Palepu, A., Sirkovic, P., Myaskovsky, A., Weissenberger, F., Rong, K., Tanno, R., Saab, K., Popovici, D. G., Blum, J., Zhang, F., Chou, K., Hassidim, A., Gokturk, B., Vahdat, A., Kohli, P., Matias, Y., Carroll, A., Kulkarni, K., Tomaev, N., Guan, Y., Dhillon, V., Vaishnav, E. D., Lee, B., Costa, T. R. D., Penad'es, J. R., Peltz, G., Xu, Y., Pawlosky, A., Karthikesalingam, A., and Natarajan, V. Towards an ai co-scientist. *ArXiv*, abs/2502.18864, 2025. URL https://api.semanticscholar.org/CorpusID:276617649.

Groeneveld, D., Beltagy, I., Walsh, E., Bhagia, A., Kinney, R., Tafjord, O., Jha, A., Ivison, H., Magnusson, I., Wang, Y., Arora, S., Atkinson, D., Authur, R., Chandu, K., Cohan, A., Dumas, J., Elazar, Y., Gu, Y., Hessel, J., Khot, T., Merrill, W., Morrison, J., Muennighoff, N., Naik, A., Nam, C., Peters, M., Pyatkin, V., Ravichander, A., Schwenk, D., Shah, S., Smith, W., Strubell, E., Subramani, N., Wortsman, M., Dasigi, P., Lambert, N., Richardson, K., Zettlemoyer, L., Dodge, J., Lo, K., Soldaini, L., Smith, N., and Hajishirzi, H. OLMo: Accelerating the science of language models. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15789–15809, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.841. URL https://aclanthology.org/2024.acl-long.841/.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D. X., and Steinhardt, J. Measuring massive multitask language understanding. *ArXiv*, abs/2009.03300, 2020. URL https://api.semanticscholar.org/CorpusID:221516475.

Jia, H., Li, T., Guan, J., and Chandrasekaran, V. The erasure illusion: Stress-testing the generalization of llm forgetting evaluation. *ArXiv*, abs/2512.19025, 2025. URL https://api.semanticscholar.org/CorpusID:284077884.

Jiang, X., Dong, Y., Fang, Z., Ma, Y., Wang, T., Cao, R., Li, B., Jin, Z., Jiao, W., Li, Y., and Li, G. Large language model unlearning for source code. *ArXiv*, abs/2506.17125, 2025. URL https://api.semanticscholar.org/CorpusID:279464101.

Jin, X., Bu, Z., Vinzamuri, B., Ramakrishna, A., Chang, K.-W., Cevher, V., and Hong, M. Unlearning as multitask optimization: A normalized gradient difference approach with an adaptive learning rate. In Chiruzzo, L., Ritter, A., and Wang, L. (eds.), *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)*, pp. 11278–11294, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.563. URL https://aclanthology.org/2025.naacl-long.563/.

jin, Z., Cao, P., Wang, C., He, Z., Yuan, H., Li, J., Chen, Y., Liu, K., and Zhao, J. Rwku: Benchmarking real-world knowledge unlearning for large language models. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 98213–98263. Curran Associates, Inc., 2024. doi: 10.52202/079017-3117.

Jin, Z., Cao, P., Wang, C., He, Z., Yuan, H., Li, J., Chen, Y., Liu, K., and Zhao, J. Rwku: Benchmarking real-world knowledge unlearning for large language models, 2024.

Kinney, R. M., Anastasiades, C., Authur, R., Beltagy, I., Bragg, J., Buraczynski, A., Cachola, I., Candra, S., Chandrasekhar, Y., Cohan, A., Crawford, M., Downey, D., Dunkelberger, J., Etzioni, O., Evans, R., Feldman, S.,

Gorney, J., Graham, D. W., Hu, F., Huff, R., King, D., Kohlmeier, S., Kuehl, B., Langan, M., Lin, D., Liu, H., Lo, K., Lochner, J., MacMillan, K., Murray, T. C., Newell, C., Rao, S., Rohatgi, S., Sayre, P., Shen, S. Z., Singh, A., Soldaini, L., Subramanian, S., Tanaka, A., Wade, A. D., Wagner, L. M., Wang, L. L., Wilhelm, C., Wu, C., Yang, J., Zamarron, A., van Zuylen, M., and Weld, D. S. The semantic scholar open data platform. *ArXiv*, abs/2301.10140, 2023. URL https://api.semanticscholar.org/CorpusID:256194545.

Lesné, S. E., Koh, M. T., Kotilinek, L. A., Kayed, R., Glabe, C. G., Yang, A. J., Gallagher, M., and Ashe, K. H. A specific amyloid-beta protein assembly in the brain impairs memory. *Nature*, 440 7082:352–7, 2006. URL https://api.semanticscholar.org/CorpusID:43841629.

Li, L., Xu, W., Guo, J., Zhao, R., Li, X., Yuan, Y., Zhang, B., Jiang, Y., Xin, Y., Dang, R., Rong, Y., Zhao, D., Feng, T., and Bing, L. Chain of ideas: Revolutionizing research via novel idea development with LLM agents. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 8971–9004, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.477. URL https://aclanthology.org/2025.findings-emnlp.477/.

Li, N., Pan, A., Gopal, A., Yue, S., Berrios, D., Gatti, A., Li, J., Dombrowski, A.-K., Goel, S., Phan, L., Mukobi, G., Helm-Burger, N., Lababidi, R. R., Justen, L., Liu, A., Chen, M. K., Barrass, I., Zhang, O., Zhu, X., Tamirisa, R., Bharathi, B., Khoja, A., Herbert-Voss, A., Breuer, C. B., Zou, A., Mazeika, M., Wang, Z., Oswal, P., Liu, W., Hunt, A. A., Tienken-Harder, J., Shih, K. Y., Talley, K., Guan, J., Kaplan, R., Steneker, I., Campbell, D., Jokubaitis, B., Levinson, A. A., Wang, J., Qian, W., Karmakar, K. K., Basart, S., Fitz, S., Levine, M. V., Kumaraguru, P., Tupakula, U. K., Varadharajan, V., Shoshitaishvili, Y., Ba, J., Esvelt, K. M., Wang, A., and Hendrycks, D. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *ArXiv*, abs/2403.03218, 2024. URL https://api.semanticscholar.org/CorpusID:268247897.

Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Liu, H., Soroush, A., Nestor, J. G., Park, E., Idnay, B. R., Fang, Y., Pan, J., Liao, S., Bernard, M., Peng, Y., and Weng, C. Retrieval augmented scientific claim verification. *JAMIA Open*, 7, 2024a. URL https://api.semanticscholar.org/CorpusID:268272963.

Liu, S., Yao, Y., Jia, J., Casper, S., Baracaldo, N., Hase, P., Xu, X., Yao, Y., Liu, C., Li, H., Varshney, K. R., Bansal, M., Koyejo, S., and Liu, Y. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, 7:181 – 194, 2024b. URL https://api.semanticscholar.org/CorpusID:267657624.

Liu, Y., Chen, H., Huang, W., Ni, Y., and Imani, M. Lune: Efficient llm unlearning via lora fine-tuning with negative examples. *ArXiv*, abs/2512.07375, 2025. URL https://api.semanticscholar.org/CorpusID:283692989.

Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., and Ha, D. The ai scientist: Towards fully automated open-ended scientific discovery. *ArXiv*, abs/2408.06292, 2024. URL https://api.semanticscholar.org/CorpusID:271854887.

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

Neel, S., Roth, A., and Sharifi-Malvajerdi, S. Descent-to-delete: Gradient-based methods for machine unlearning. *ArXiv*, abs/2007.02923, 2020. URL https://api.semanticscholar.org/CorpusID:220364296.

Novikov, A., Vũ, N., Eisenberger, M., Dupont, E., Huang, P.-S., Wagner, A. Z., Shirobokov, S., Kozlovskii, B., Ruiz, F. J. R., Mehrabian, A., Kumar, M. P., See, A., Chaudhuri, S., Holland, G., Davies, A., Nowozin, S., Kohli, P., and Balog, M. AlphaEvolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

Pannu, J., Bloomfield, D., MacKnight, R., Hanke, M. S., Zhu, A., Gomes, G., Cicero, A., and Inglesby, T. V. Dual-use capabilities of concern of biological ai models. *PLOS Computational Biology*, 21, 2025. URL https://api.semanticscholar.org/CorpusID:278427349.

Park, D. H. and Blake, C. Identifying comparative claim sentences in full-text scientific articles. In Van Den Bosch, A. and Shatkay, H. (eds.), *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, pp. 1–9, Jeju Island, Korea, July 2012. Association for Computational

Linguistics. URL https://aclanthology.org/W12-4301/.

Pradeep, R., Ma, X., Nogueira, R., and Lin, J. Scientific claim verification with VerT5erini. In Holderness, E., Jimeno Yepes, A., Lavelli, A., Minard, A.-L., Pustejovsky, J., and Rinaldi, F. (eds.), *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, pp. 94–103, online, April 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.louhi-1.11/.

Pramanick, A., Hou, Y., Mohammad, S. M., and Gurevych, I. Claimflow: Tracing the evolution of scientific claims in nlp, 2026. URL https://api.semanticscholar.org/CorpusID:286579708.

Reimers, N. and Gurevych, I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Inui, K., Jiang, J., Ng, V., and Wan, X. (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Rybak, P., Batorski, P., Swoboda, P., and Spurek, P. Rebel: Hidden knowledge recovery via evolutionary-based evaluation loop, 2026. URL https://arxiv.org/abs/2602.06248.

Sarukkai, V., Gupta, A., Hong, J., Gharbi, M., and Fatahalian, K. In-context distillation with self-consistency cascades: A simple, training-free way to reduce llm agent costs. *ArXiv*, abs/2512.02543, 2025. URL https://api.semanticscholar.org/CorpusID:283457846.

Shi, W., Lee, J., Huang, Y., Malladi, S., Zhao, J., Holtzman, A., Liu, D., Zettlemoyer, L. S., Smith, N. A., and Zhang, C. Muse: Machine unlearning six-way evaluation for language models. *ArXiv*, abs/2407.06460, 2024. URL https://api.semanticscholar.org/CorpusID:271064299.

Siddiqui, S. A., Weller, A., Krueger, D., Dziugaite, G. K., Mozer, M. C., and Triantafillou, E. From dormant to deleted: Tamper-resistant unlearning through weight-space regularization. *ArXiv*, abs/2505.22310, 2025. URL https://api.semanticscholar.org/CorpusID:278959245.

Singh, A., D'Arcy, M., Cohan, A., Downey, D., and Feldman, S. SciRepEval: A multi-format benchmark for scientific document representations. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5548–5566, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.338. URL https://aclanthology.org/2023.emnlp-main.338/.

Soldaini, L., Kinney, R., Bhagia, A., Schwenk, D., Atkinson, D., Authur, R., Bogin, B., Chandu, K., Dumas, J., Elazar, Y., Hofmann, V., Jha, A., Kumar, S., Lucy, L., Lyu, X., Lambert, N., Magnusson, I., Morrison, J., Muennighoff, N., Naik, A., Nam, C., Peters, M., Ravichander, A., Richardson, K., Shen, Z., Strubell, E., Subramani, N., Tafjord, O., Walsh, E., Zettlemoyer, L., Smith, N., Hajishirzi, H., Beltagy, I., Groeneveld, D., Dodge, J., and Lo, K. Dolma: an open corpus of three trillion tokens for language model pretraining research. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15725–15788, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.840. URL https://aclanthology.org/2024.acl-long.840/.

Stokel-Walker, C. Scientists invented a fake disease. AI told people it was real. *Nature*, 652:559–561, April 2026. doi: 10.1038/d41586-026-01100-y.

Tang, X., Duan, X., and Cai, Z. Large language models for automated literature review: An evaluation of reference generation, abstract writing, and review composition. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1602–1617, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.83. URL https://aclanthology.org/2025.emnlp-main.83/.

Uprety, S., Buyuklieva, B., and Tiwari, P. Using large language models for hypotheses and claims extraction from scientific literature. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, pp. 1645–1648, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713316. doi: 10.1145/3701716.3717752. URL https://doi.org/10.1145/3701716.3717752.

Urbina, F., Lentzos, F., Invernizzi, C., and Ekins, S. Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*, 4:189 – 191, 2022. URL https://api.semanticscholar.org/CorpusID:247302391.

Wadden, D., Lin, S., Lo, K., Wang, L. L., van Zuylen, M., Cohan, A., and Hajishirzi, H. Fact or fiction: Verifying scientific claims. In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7534–7550, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.609. URL https://aclanthology.org/2020.emnlp-main.609/.

Wang, Y., Feng, S., Tsvetkov, Y., and Hajishirzi, H. Sciencemeter: Tracking scientific knowledge updates in language models. *arXiv preprint arXiv:2505.24302*, 2025a.

Wang, Y., Niu, Z., Ji, H., He, G., Gao, H., and Hua, G. Robust mllm unlearning via visual knowledge distillation, 2025b. URL https://api.semanticscholar.org/CorpusID:283883633.

Wei, X. Claimdistiller: Scientific claim extraction with supervised contrastive learning. In *EEKE/AII@JCDL*, 2023. URL https://api.semanticscholar.org/CorpusID:259359584.

Xie, L., Teng, X., Ke, S., Wen, H., and Wang, S. Reveal and release: Iterative llm unlearning with self-generated data. In *Conference on Empirical Methods in Natural Language Processing*, 2025. URL https://api.semanticscholar.org/CorpusID:281394642.

Yang, R. Unlearning as ablation: Toward a falsifiable benchmark for generative scientific discovery. *ArXiv*, abs/2508.17681, 2025. URL https://api.semanticscholar.org/CorpusID:280711595.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. HellaSwag: Can a machine really finish your sentence? In Korhonen, A., Traum, D., and Màrquez, L. (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472/.

Zhang, D., Finckenberg-Broman, P., Hoang, T., Pan, S., Xing, Z., Staples, M., and Xu, X. Right to be forgotten in the era of large language models: implications, challenges, and solutions. *AI and Ethics*, 5:2445 – 2454, 2023. URL https://api.semanticscholar.org/CorpusID:259501864.

Zhang, R., Lin, L., Bai, Y., and Mei, S. Negative preference optimization: From catastrophic collapse to effective unlearning. *ArXiv*, abs/2404.05868, 2024. URL https://api.semanticscholar.org/CorpusID:269009619.

## A. Example of Retracted and Falsified Claims

This example illustrates a critical failure mode in modern language models: the persistence of outdated, retracted, or falsified scientific claims. Despite subsequent corrections in the scientific record, models such as OLMO-3-7B-Instruct (Groeneveld et al., 2024) may continue to generate responses that reflect superseded knowledge. As shown in Figure 3, the model produces confident outputs that do not adequately account for post-publication revisions or retractions.

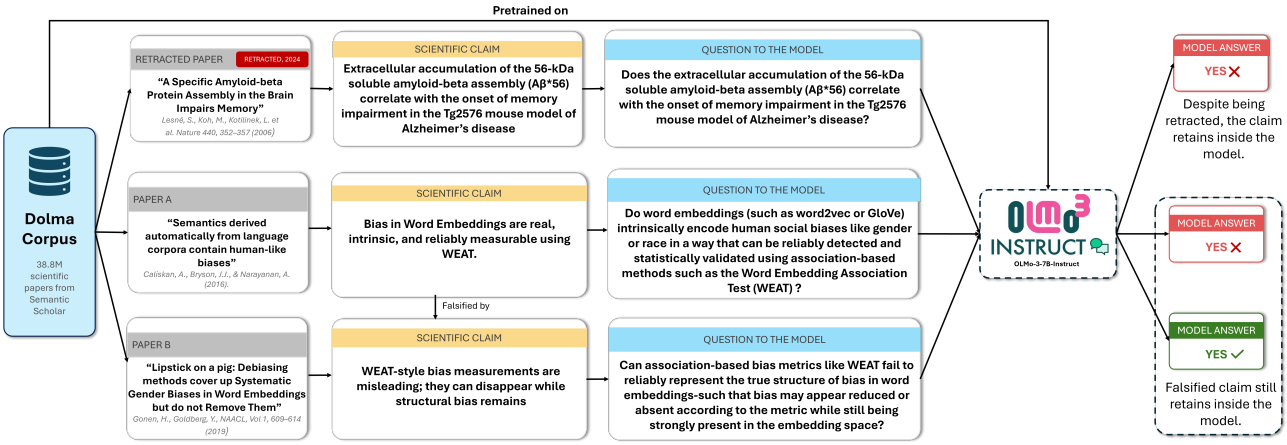



*Figure 3.* Failure case demonstrating the persistence of invalidated scientific knowledge in language models (OLMO-3-7B-Instruct (Groeneveld et al., 2024)).

## B. Additional Information about Dataset Creation

**Why OLMO model and DOLMA corpus is chosen?** Dolma is chosen for its scale, recency, and its direct use in training OLMO models, ensuring that evaluated claims are likely present in the pretraining data leading to meaningful unlearning evaluation. In contrast to models trained on The Pile (Gao et al., 2020), such as GPT-Neo ((Black et al., 2022)) and GPT-J, Dolma and OLMO provide a more suitable foundation for scientific reasoning. GPT-Neo and GPT-J are early base language models, and their authors explicitly note that they are not intended for direct use without task-specific fine-tuning (Black et al., 2022).

**How self consistency check is done while asking OLMO?** Self consistency check is done to mitigate stochasticity and hallucination (Sarukkai et al., 2025), where each question is queried thrice independently, retaining the pair only if the correct answer is produced in all trials.

**How QA pairs from the Forget set is splitted into Forget Set 1 and Forget Set 2?** We partition these QA pairs into two non-overlapping subsets, $\mathcal{F}_1$ and $\mathcal{F}_2$. We split the two QA pairs belong to each question type for a claim, which are paraphrased versions of each other, into the forget sets $\mathcal{F}_1$ and $\mathcal{F}_2$. This setup enables cross-generalization evaluation, such that unlearning on one forget set should induce forgetting on the other, and vice versa.

## C. LLM Evaluation

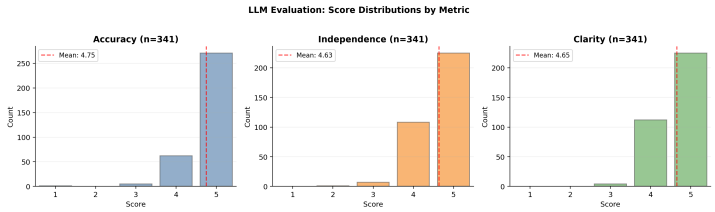
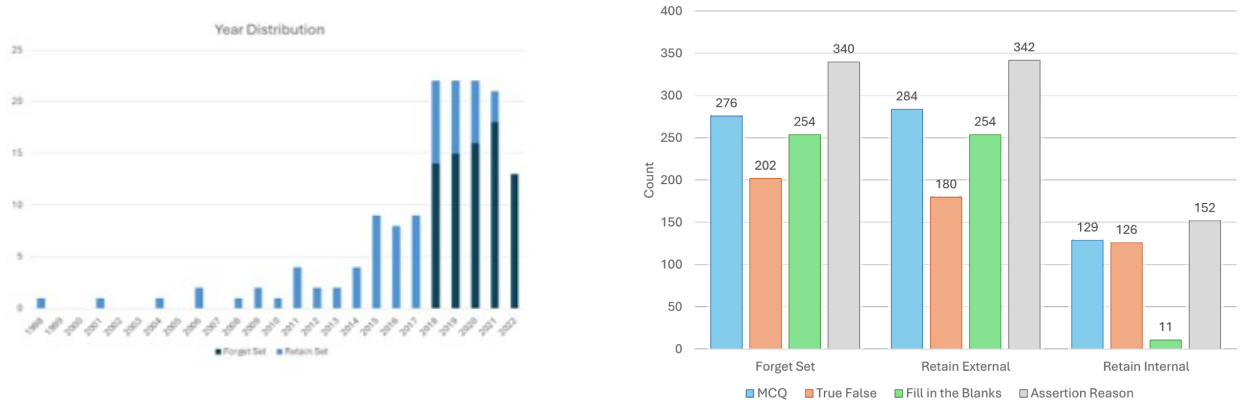



*Figure 4.* LLM as a judge score distribution

Figure 4 demonstrates the score distributions, where in each case the mean is skewed towards the higher end of the scale, indicating good quality of the extracted claims.

## D. Dataset Statistics




*(a)* Year-wise distribution of questions

*(b)* Question type distribution

*Figure 5.* SciUnlearn dataset statistics. (a) Year-wise distribution of questions. (b) Distribution across question types.

Figure 5a shows the year-wise distribution of questions, which is skewed towards more recent papers for both the forget and retain sets. Figure 5a presents the question-type distribution, illustrating the diversity and relative frequency of different question categories present in the forget and retain sets. The total cost of dataset creation is approximately $36.57 over 2534 GPT-5 API calls.

## E. Prompts Used for Dataset Construction

This section provides the prompts used for scientific claim extraction and question-answer pair generation.

### E.1. Claim Extraction Prompt

The following prompt is used to extract scientific claims from the abstract, introduction, and conclusion sections of each paper:

**Claim Extraction Prompt**

**Role:** You are a careful scientific reader.

**Definition.** A ”scientific claim” is a concise, declarative statement describing a finding, result, or methodological fact. A valid claim:

- stands alone as a factual statement,
- is specific and empirically or theoretically verifiable,
- reflects scope and conditions where necessary,
- does not refer to the paper, authors, or presentation (e.g., no ”this paper”, ”we”, ”the authors”).

**Task.** Read the provided paper text and generate the core scientific claims (do not quote).

**Output Rules.**

- Write claims as factual statements, not descriptions of the paper.
- Avoid phrases such as: ”the paper proposes”, ”this work shows”, ”the authors introduce”, ”we demonstrate”.

- Do not include numerical results; focus on conceptual insights or methods.
- Each claim must be exactly one concise sentence.
- Prefer a single consolidated claim when possible.
- If multiple claims exist, return at most 5 claims.

**Output Format (JSON).**

```
{
  "claims": [
    { "text": "one claim" },
    { "text": "one claim" }
  ]
}
```

If no valid scientific claim is present, return:

```
{ "claims": [] }
```

### E.2. Question-Answer Generation Prompt for Forget Set and Retain Set External

The following prompt is used to generate question-answer pairs for each extracted claim:

**Question-Answer Generation Prompt for Forget Set and Retain Set External**

**Role:** You are an expert exam-item writer.

**Definition.** A "scientific claim" is a concise, declarative statement that represents a fact-like proposition, is specific and verifiable (empirically or theoretically), and includes necessary scope or conditions without referring to the source or using presentation language.

**Task.** You will receive one scientific claim (a single sentence). Generate assessment items that evaluate the entire content of the claim.

**Output Format (JSON).** Return only valid JSON in the following schema:

```
{
  "mcq": [
    { "question": "Choose the correct option (A-D): ... [A] ... [B] ... [C] ... [D]
    ↪  ...", "answer": "A" },
    { "question": "Choose the correct option (A-D): ... [A] ... [B] ... [C] ... [D]
    ↪  ...", "answer": "C" }
  ],
  "true_false": [
    { "question": "...", "answer": "True" },
    { "question": "...", "answer": "False" }
  ],
  "fill_blank": [
    { "question": "... _________(option1/option2)", "answer": "option1" },
    { "question": "... _________(option1/option2)", "answer": "option1" }
  ],
  "assertion_reason": [
    { "question": "Assertion (A): ... Reason (R): ...", "answer": "A is True, R is
    ↪  True, and R explains A" },
    { "question": "Assertion (A): ... Reason (R): ...", "answer": "A is True, R is
    ↪  False" }
  ]
}
```

**Strict Rules.**

- Coverage: Each item must assess the entire claim, not a partial detail.
- Quantity: Generate exactly 2 items per category.
- Keys: Each item must contain only "question" and "answer".

- MCQ: Include exactly four options labeled [A]–[D]; answer must be a single letter.
- True/False: Answer must be exactly "True" or "False".
- Fill-in-the-blank:
  - Use exactly one blank: _____.
  - Include exactly two inline options in the format: `(option1/option2)`.
  - One option must be correct; the other must be a plausible distractor.
- Assertion–Reason: Use one assertion and one reason; answer must follow predefined formats.
- Do not create multi-part or multi-sentence questions.
- Use precise academic language; do not introduce external information.

**Critical Constraints.**

- Treat the claim as an atomic unit; do not refer to any paper, model, or author.
- Avoid phrases such as "the paper states" or "the authors propose".
- Each question must be self-contained and independently understandable.
- Restate the claim within each question to allow standalone interpretation.
- Do not assume access to the original claim.

**Example (Fill-in-the-Blank).**

```
Traditional RETE-based pattern matching is _________(insufficient/abstract) for
↪  frequently updated working memories.
The framework translates OWL 2 RL ontologies into _________(Horn-clause/graphical)
↪  rules.
```

### E.3. Internal Retain Set Question-Answer Generation Prompt

**Internal Retain Set QA Generation Prompt**

**Role:** You are a careful scientific reader and expert exam-item writer.

**Task.** Read the provided paper text and generate:

- a compact `paper_base` describing the foundational identity of the paper,
- paper-level assessment questions that test broad conceptual understanding.

**Important.** You will also be provided with a set of extracted scientific claims. These are given only as context and must **not** be directly tested, paraphrased, or verified in the generated questions.

**Question Design Requirements.** Generated questions must focus on:

- the main concept, model, or system,
- the core task or problem,
- the method family or paradigm,
- the input-output behavior,
- the overall framework or setup,
- the role of the central concept.

All questions must be:

- standalone and self-contained,

- concept-level and foundational,
- independent of any explicit paper or author references.

**Do Not Generate.**

- claim-level or claim-verification questions,
- numerical or result-specific questions,
- paraphrases of provided claims,
- references to authors, paper, or presentation,
- trivial metadata-based questions.

**Output Format (JSON).** Return only valid JSON in the following format:

```
{
  "qa_by_base": {
    "mcq": [
      { "question": "Choose the correct option (A-D): ... [A] ... [B] ... [C] ...
      ↪  [D] ...", "answer": "A" },
      { "question": "Choose the correct option (A-D): ... [A] ... [B] ... [C] ...
      ↪  [D] ...", "answer": "C" }
    ],
    "true_false": [
      { "question": "...", "answer": "True" },
      { "question": "...", "answer": "False" }
    ],
    "fill_blank": [
      { "question": "... _______ ...", "answer": "..." },
      { "question": "... _______ ...", "answer": "..." }
    ],
    "assertion_reason": [
      { "question": "Assertion (A): ... Reason (R): ...", "answer": "A is True, R is
      ↪  True, and R explains A" },
      { "question": "Assertion (A): ... Reason (R): ...", "answer": "A is True, R is
      ↪  False" }
    ]
  }
}
```

**Strict Rules.**

- Generate exactly 2 questions per category.
- MCQ must include exactly four options labeled [A], [B], [C], [D].
- True/False answers must be exactly ”True” or ”False”.
- Fill-in-the-blank must contain exactly one blank represented as _____.
- Assertion–Reason answers must follow standard formats.

**Critical Constraints.**

- Do not refer to any paper, author, or source.
- Do not include phrases like ”the paper states” or ”the authors propose”.
- Each question must be independently understandable.
- Avoid overlap with provided scientific claims.

## F. Claim Evaluation Prompt

To assess the quality of the extracted scientific claims, the authors employ a structured evaluation strategy that measures multiple dimensions of claim quality. Specifically, each claim is evaluated along three key aspects: accuracy, independence, and clarity. These dimensions are designed to capture whether a claim faithfully represents the underlying scientific content, is self-contained, and is expressed in a precise and interpretable manner.

To ensure consistency and scalability, we use a prompt-based evaluation strategy with a large language model as well as with human evaluation. The prompt provides detailed scoring rubrics for each dimension and enforces a structured output format for reliable aggregation and analysis.

**Claim Evaluation Prompt**

**Role:** You are an expert scientific-claim evaluator.

**Task.** You will receive a paper title, a scientific claim, and the corresponding paper text. Your task is to evaluate the claim along three dimensions:

- Accuracy (0–5)
- Independence (0–5)
- Clarity (0–5)

**Definitions.**

- **Accuracy:** Faithfulness to the paper's findings and scope, without contradiction or exaggeration.
- **Independence:** Whether the claim is self-contained and does not rely on hidden assumptions.
- **Clarity:** Whether the claim is precise, unambiguous, and clearly interpretable.

**Scoring Rubric.**
**Accuracy**

- 5: Fully accurate; explicitly supported with correct scope and qualifiers.
- 4: Mostly accurate; minor simplifications without changing meaning.
- 3: Partially accurate; some omissions or mild overgeneralization.
- 2: Weakly accurate; misstates key aspects such as method, population, or condition.
- 1: Minimally accurate; substantially misrepresents the finding.
- 0: Inaccurate; unsupported, contradicted, or fabricated.

**Independence**

- 5: Fully independent; completely self-contained.
- 4: Mostly independent; minor implicit assumptions.
- 3: Moderately independent; combines related ideas but still interpretable.
- 2: Weakly independent; multiple claims that should be separated.
- 1: Minimally independent; relies heavily on unstated assumptions.
- 0: Not independent; cannot be evaluated in isolation.

**Clarity**

- 5: Very clear; precise and unambiguous.
- 4: Clear; minor ambiguity.
- 3: Moderately clear; somewhat vague.

- 2: Unclear; difficult to interpret.
- 1: Very unclear; confusing.
- 0: Incoherent; unintelligible.

**Output Format (JSON).**

```
{
  "accuracy_score": 0,
  "accuracy_reason": "...",
  "independence_score": 0,
  "independence_reason": "...",
  "clarity_score": 0,
  "clarity_reason": "...",
  "overall_assessment": "..."
}
```

**Strict Requirements.**

- Output must be valid JSON only.
- Scores must be integers in the range 0 to 5.
- Output must never be empty.
- If evidence is insufficient, assign conservative scores and justify uncertainty.

## G. Sample Claims and Evaluation Scores

We present representative examples of scientific claims extracted from papers(title mentioned in bold) in Table 3 and Table 4 along with their corresponding LLM-based and human evaluation scores.

*Table 3.* Comparison of LLM and Human Evaluation Scores in Forget Set. Metrics: Accuracy (A), Independence (I), and Clarity (C).

| Claim | LLM (A/I/C) | Human (A/I/C) |
|---|---|---|
| **Federated Deep Reinforcement Learning for Open RAN Slicing in 6G Networks** | | |
| Federated deep reinforcement learning, in which MVNOs locally train DDPG-based bandwidth allocation policies and share model updates for weighted aggregation at a non-RT RIC, produces privacy-preserving and generalizable RAN slicing policies that maintain URLLC delay and eMBB throughput SLAs under heterogeneous non-i.i.d. conditions while enabling low-latency online inference at the near-RT RIC. | 5/4/4 | 5/4/4 |
| **Web-Based Volunteer Distributed Computing for Handling Time-Critical Urgent Workloads** | | |
| Browser-based volunteer distributed computing using WebWorkers and a centralized cross-origin task manager can viably execute embarrassingly parallel urgent workloads without client-side software installation across heterogeneous platforms. | 4/5/5 | 4/5/5 |
| In web-based volunteer computing, a small minority of long dwell-time sessions complete most tasks while many short sessions complete none, making performance dependent on value sessions and imposing overhead from non-value sessions. | 5/5/5 | 4/5/5 |
| Checkpointing partial results, batching tasks, and asynchronously prefetching new tasks increase the amount of useful work completed during volatile browsing sessions at the cost of higher server-side load. | 5/5/5 | 5/4/5 |

*Continued on next page*

| Claim | LLM (A/I/C) | Human (A/I/C) |
|---|---|---|
| Tuning task granularity is essential because smaller tasks increase the likelihood of value sessions but also amplify task management and communication overheads. | 5/5/4 | 5/5/5 |
| For data-bound workloads, `XMLHttpRequest` communication can outperform WebSockets, whereas compute-bound kernels are relatively insensitive to protocol choice and their browser-executed performance relative to native code depends on algorithm characteristics and browser engine optimizations. | 4/5/5 | 4/4/5 |
| **A topological analysis of cointegrated data: a Z24 Bridge case study** | | |
| Time-delay embeddings of Z24 Bridge natural frequency time series exhibit a toroidal topology driven by cyclic temperature effects, whereas embeddings of cointegrated residuals exhibit an open-ball topology characteristic of approximately Gaussian white-noise processes. | 5/5/5 | 4/5/5 |
| The nonlinear dependence of the second natural frequency on temperature leads linear cointegration to leave residual topological structure, while Gaussian process-based nonlinear cointegration trained on data that include the nonlinear regime more effectively removes these effects. | 5/5/5 | 4/5/5 |
| Persistent homology comparisons using Wasserstein distances on Vietoris-Rips complexes of embedded series provide a quantitative topology-based metric for assessing the effectiveness of cointegration in detrending environmental and operational variability. | 5/5/4 | 5/4/4 |
| **Predicting Companies' ESG Ratings from News Articles Using Multivariate Timeseries Analysis** | | |
| Annual corporate ESG ratings can be inferred from news by modeling monthly multivariate timeseries that integrate weakly supervised ESG-relevance detection, sentiment ratios for relevant and irrelevant articles, and semantic topic cluster frequencies. | 4/4/4 | 4/4/5 |
| Formulating ESG rating prediction as a regression task over these timeseries with deep convolutional architectures yields superior predictive performance compared to classification and to index-aggregation baselines based on document-level transformer outputs. | 4/4/4 | 4/4/4 |
| Weak supervision using sentence-embedding similarity to ESG definitions effectively bootstraps a DistilBERT-based classifier to generate scalable ESG-relevance signals without manual keyword selection. | 5/5/5 | 5/5/5 |
| A large-scale news corpus aligned to thousands of US companies and annual third-party ESG scores over multiple years is compiled and publicly released to support training and evaluation of ESG prediction models. | 5/5/5 | 4/5/5 |
| **Bayesian optimization with partially specified queries** | | |
| Bayesian optimization with partially specified queries formalizes optimization where only a subset of input variables are specified and the remainder are randomly drawn, optimizing the conditional expectation over unspecified variables, and posterior-sampling algorithms for known and unknown input distributions—incorporating GP samples and UCB-style bonuses for empirical distribution estimates under an independence assumption—admit sublinear Bayesian cumulative regret for popular kernels without exponential dependence on dimensionality. | 5/4/4 | 5/4/4 |
| **Federated Deep Reinforcement Learning for Open RAN Slicing in 6G Networks** | | |

*Continued on next page*

| Claim | LLM (A/I/C) | Human (A/I/C) |
|---|---|---|
| Federated deep reinforcement learning, in which MVNOs locally train DDPG-based bandwidth allocation policies and share model updates for weighted aggregation at a non-RT RIC, produces privacy-preserving and generalizable RAN slicing policies that maintain URLLC delay and eMBB throughput SLAs under heterogeneous non-i.i.d conditions while enabling low-latency online inference at the near-RT RIC. | 5/4/4 | 5/4/5 |
| **Bayesian optimization with partially specified queries** | | |
| Bayesian optimization with partially specified queries formalizes optimization where only a subset of input variables are specified and the remainder are randomly drawn, optimizing the conditional expectation over unspecified variables, and posterior-sampling algorithms for known and unknown input distributions—incorporating GP samples and UCB-style bonuses for empirical distribution estimates under an independence assumption—admit sublinear Bayesian cumulative regret for popular kernels without exponential dependence on dimensionality. | 5/4/4 | 4/4/4 |
| **Deep-Learning-Assisted IoT-Based RIS for Cooperative Communications** | | |
| Training deep neural networks on channel state information to compute RIS element phase shifts and to classify received symbols enables IoT-coordinated, low-complexity multi-RIS cooperative communications without a direct path, with reliability governed by relay placement under Rayleigh fading and path loss. | 4/4/4 | 4/4/5 |
| **Deep Learning Based Customer Preferences Analysis in Industry 4.0 Environment** | | |
| A deep learning approach that partitions restaurant sales time series by day-of-week and holiday type, models ordinary days with stacked LSTMs, and uses MLPs to infer total holiday demand from pre-holiday signals and allocate it across holiday days effectively captures weekly seasonality and event-driven surges in nonlinear, intermittent demand while remaining computationally lightweight for real-time deployment. | 3/3/3 | 4/3/2 |
| **Detection of agglomerate fog based on a shallow convolutional neural network** | | |
| A shallow convolutional neural network with progressively smaller convolution kernels and pooling followed by two fully connected layers effectively detects agglomerate fog in images by capturing multi-scale global and local features while reducing overfitting and computational cost relative to deeper architectures. | 5/5/5 | 5/5/4 |
| Partitioning an image into sub-images and aggregating sub-classification decisions via voting improves robustness of agglomerate fog detection compared with whole-image classification by mitigating the influence of fog-free regions. | 4/4/4 | 4/5/4 |
| Applying dropout to the fully connected layers enhances generalization performance for agglomerate fog detection compared to training without dropout. | 3/4/5 | 4/4/5 |
| Training with synthetic fog generated using the atmospheric scattering model produces data sufficient to learn detectors that transfer to real-world images. | 4/4/5 | 4/4/5 |

*Table 4.* Comparison of LLM and Human Evaluation Scores in Retain Set. Metrics: Accuracy (A), Independence (I), and Clarity (C).

| Claim | LLM (A/I/C) | Human (A/I/C) |
|---|---|---|

*Continued on next page*

| Claim | LLM (A/I/C) | Human (A/I/C) |
|---|---|---|
| **V-BOINC: The Virtualization of BOINC** | | |
| Running BOINC applications inside lightweight virtual machines that embed a BOINC client enables execution on heterogeneous volunteer hosts without per-architecture porting, while providing sandbox isolation and system-level checkpointing and recovery. | 5/5/5 | 5/5/4 |
| In web-based volunteer computing, a small minority of long dwell time sessions complete most tasks while many short sessions complete none, making performance dependent on value sessions and imposing overhead from non-value sessions. | 5/5/5 | 4/5/5 |
| Mountable dependency disks carrying application software stacks allow BOINC applications with complex dependencies to run without modifying host environments and without inflating the base virtual machine image. | 5/5/5 | 4/5/5 |
| Using headless VirtualBox with bridged networking and guest control APIs satisfies the need for unique guest addressing, remote management, and VM snapshotting while maintaining user transparency on Linux and macOS. | 5/5/5 | 5/4/4 |
| Performance differences relative to native BOINC are primarily due to virtualization-imposed resource constraints, with negligible additional overhead introduced by the V-BOINC control layer. | 5/5/5 | 5/5/5 |
| Periodic virtual machine snapshots preserve job progress across host interruptions with typically low storage and time overheads for CPU-bound workloads, increasing for memory- or disk-intensive tasks. | 5/5/5 | 5/4/5 |
| **On the application of topological data analysis: a Z24 Bridge case study** | | |
| A point-count-normalized sum of Wasserstein distances between persistence barcodes provides a sampling-invariant metric for comparing manifold topology across data partitions. | 5/4/4 | 5/4/4 |
| **Semi-supervised Text Mining for Monitoring the News About the ESG Performance of Companies** | | |
| A semi-supervised text mining approach that uses word-embedding–expanded ESG lexicons, valence-aware sentiment scoring, and rigorous filtering to build time-varying entity-level news indicators produces signals that, in a Dutch-language news corpus of European companies, anticipate ESG rating downgrades and enhance portfolio screening. | 4/4/4 | 4/4/5 |
| **Distributionally Robust Bayesian Quadrature Optimization** | | |
| Maximizing the expected integrand under the worst-case distribution within a $\chi^2$-divergence ball around the empirical context distribution defines a distributionally robust Bayesian quadrature objective that reduces to a variance-regularized risk in a high-variance regime, admits efficient inner minimization via KKT-based bisection, and can be optimized by posterior sampling with sublinear Bayesian regret under standard kernel assumptions. | 4/4/3 | 4/4/4 |
| **Dynamic SDN-Based Radio Access Network Slicing With Deep RL for URLLC and eMBB Services** | | |
| The global two-level resource block allocation for URLLC and eMBB with inter-gNodeB borrowing, OFDMA exclusivity, per-user resource limits, eMBB minimum rates, and URLLC delay bounds constitutes a non-linear binary optimization problem that is NP-hard via reduction from the 0–1 knapsack problem. | 5/5/5 | 5/5/4 |
| A two time-scale SDN-based RAN slicing mechanism that allocates resource blocks to gNodeBs using a single-agent EXP3 bandit and performs per-user scheduling and cooperative inter-gNodeB resource sharing using multi-agent deep Q-learning dynamically adapts to traffic and channel variations while limiting controller–gNodeB signaling. | 5/5/5 | 5/4/5 |

*Continued on next page*

| Claim | LLM (A/I/C) | Human (A/I/C) |
|---|---|---|
| Packet delays for URLLC and eMBB traffic are modeled with an M/M/1 queuing system, and minimum-rate and maximum-delay constraints are enforced within the allocation and learning processes. | 5/5/5 | 4/5/5 |
| Support for multiple 5G NR OFDM numerologies is achieved by selecting numerology-specific trained models at the gNodeB based on channel state information. | 4/5/4 | 4/5/5 |
| **Deep-Learning-aided Detection for Reconfigurable Intelligent Surfaces** | | |
| A deep learning detector with fully connected layers and tanh activations can jointly infer channel fading and RIS phase shifts directly from received signals without pilot signaling, enabling symbol detection in passive RIS-assisted links with lower inference complexity than maximum-likelihood and MMSE detectors and robust error performance under imperfect CSI and mismatches in channel models or RIS element counts. | 4/4/4 | 4/4/5 |
| **Daily Sales Forecasting for Grapes by Support Vector Machine** | | |
| Support vector regression with radial basis function kernels, trained with cross-validated hyperparameters on limited historical sales augmented by quantified weather, weekday/weekend, and price information, yields more accurate daily demand forecasts for table grapes than artificial neural networks and decision trees in small-sample, nonlinear retail settings. | 5/4/4 | 5/5/4 |
| **Fast Fog Detection for De-Fogging of Road Driving Images** | | |
| Temporal averaging of the value-to-saturation ratio computed within a vanishing-point–centered region of interest provides a fast, training-free indicator of fog presence and density in road driving images, and using this indicator to modulate transmission while detecting sky via a blue-channel threshold and processing sky (gamma) and non-sky (de-fogging) separately reduces over-enhancement and sky color distortion. | 4/3/3 | 3/4/3 |

## H. Sample Question-Answer Generation and Filtering

This section presents qualitative examples illustrating the end-to-end pipeline, including extracted scientific claims, generated question-answer pairs, and their evaluation. For each paper, we list the extracted claims, followed by representative question-answer pairs derived from a selected claim, verbatim statement from where the claim is taken, along with the reference answers, model predictions, and evaluation metrics.

**Forget Paper Title: Is it Required? Ranking the Skills Required for a Job-Title**

**Extracted Claim:**

- Fine-tuning a multilingual sentence encoder on English job-title data improves skill importance ranking in English while reducing performance on non-English titles.

**Verbatim Claim:**

- The model also performs fairly on other languages (as seen in the table 3), but by fine tuning we lose the multi- lingual capabilities of the ranking system.

**Sample QA Pairs (from the above claim):**

**MCQ Example 1** **Question:** Choose the correct option (A–D): When a multilingual sentence encoder is fine-tuned on English job-title data, what is the observed effect on skill importance ranking for English and non-English titles? [A] Both English and non-English performance improve equally [B] English skill importance ranking improves, but performance on non-English titles decreases [C] English performance decreases while non-English improves [D] There is no measurable

change in either English or non-English performance
**Reference Answer:** B
**OLMo Answer:** B
**Metrics:** Exact Match = 1.0, ROUGE = 1.0, Semantic Similarity = 1.0

**MCQ Example 2** **Question:** Choose the correct option (A–D): Fine-tuning a multilingual sentence encoder using only English job-title data leads to which outcome in skill importance ranking across languages? [A] Improvement limited to non-English titles with no change in English [B] Degradation for both English and non-English titles [C] Improved English results coupled with reduced performance on non-English titles [D] No trade-off; both English and non-English results improve
**Reference Answer:** C
**OLMo Answer:** C
**Metrics:** Exact Match = 1.0, ROUGE = 1.0, Semantic Similarity = 1.0

**True/False Example 1** **Question:** Fine-tuning a multilingual sentence encoder on English job-title data increases skill importance ranking performance for English titles while lowering performance for non-English titles.
**Reference Answer:** True
**OLMo Answer:** True
**Metrics:** Exact Match = 1.0, ROUGE = 1.0, Semantic Similarity = 1.0

**True/False Example 2** **Question:** Fine-tuning a multilingual sentence encoder on English job-title data improves skill importance ranking performance for both English and non-English titles. **Reference Answer:** False
**OLMo Answer:** False
**Metrics:** Exact Match = 1.0, ROUGE = 1.0, Semantic Similarity = 1.0

**Fill in the Blanks Example 1** **Question:** When a multilingual sentence encoder is fine-tuned on English job-title data, English skill importance ranking improves, but _____(non-English performance decreases/non-English performance also improves)
**Reference Answer:** non-English performance decreases
**OLMo Answer:** non-English performance decreases
**Metrics:** Exact Match = 1.0, ROUGE = 1.0, Semantic Similarity = 1.0

**Fill in the Blanks Example 2** **Question:** Fine-tuning a multilingual sentence encoder using English job-title data produces a trade-off: better English skill importance ranking and _____(worse non-English performance/unchanged non-English performance)
**Reference Answer:** worse non-English performance
**OLMo Answer:** worse non-English performance
**Metrics:** Exact Match = 1.0, ROUGE = 1.0, Semantic Similarity = 1.0

**Assertion Reason Example 1** **Question:** Assertion (A): Fine-tuning a multilingual sentence encoder on English job-title data improves skill importance ranking for English titles while reducing performance on non-English titles. Reason (R): The fine-tuning uses only English job-title examples, not multilingual data.
**Reference Answer:** A is True, R is True, and R explains A.
**OLMo Answer:** A is True, R is True, and R explains A.
**Metrics:** Exact Match = 1.0, ROUGE = 1.0, Semantic Similarity = 1.0

**Assertion Reason Example 2** **Question:** Fine-tuning a multilingual sentence encoder on English job-title data yields improved English skill importance ranking alongside reduced performance on non-English titles. Reason (R): Because the model is perfectly multilingual, English-only fine-tuning cannot hurt non-English performance.
**Reference Answer:** A is True, R is False
**OLMo Answer:** A is True, R is False
**Metrics:** Exact Match = 1.0, ROUGE = 1.0, Semantic Similarity = 1.0

**Retain Paper Title: JobBERT: Understanding Job Titles through Skills**

**Extracted Claim:**

- Semantic representations of job titles learned by a BERT-based encoder trained with distant supervision to predict co-occurring skills from vacancy texts, using token-level gating and negative sampling, enable taxonomy-agnostic nearest-neighbor normalization without manual title labels and outperform generic sentence encoders on this task.

**Sample QA Pairs (from the above claim):**

**MCQ Example 1** **Question:** Choose the correct option (A–D): Which statement correctly describes how semantic representations of job titles are learned and what they achieve? [A] Representations are learned by a BERT-based encoder trained with distant supervision to predict co-occurring skills from vacancy texts, using token-level gating and negative sampling; they enable taxonomy-agnostic nearest-neighbor normalization without manual title labels and outperform generic sentence encoders on this task. [B] Representations are learned by an LSTM-based encoder trained with manual title labels, without token-level gating or negative sampling; they enable taxonomy-specific normalization via rule-based matching and do not outperform generic sentence encoders on this task. [C] Representations are learned by a BERT-based encoder trained to predict job titles from skills, using token-level gating but no negative sampling; they require manual title labels for normalization and show performance comparable to generic sentence encoders. [D] Representations are obtained by unsupervised averaging of word embeddings from vacancy texts; they support taxonomy-agnostic cluster-based normalization but only match the performance of generic sentence encoders on this task.
**Reference Answer:** A
**OLMo Answer:** A
**Metrics:** Exact Match = 1.0, ROUGE = 1.0, Semantic Similarity = 1.0

**MCQ Example 2** **Question:** Choose the correct option (A–D): What configuration yields job-title representations that enable taxonomy-agnostic nearest-neighbor normalization without manual title labels and surpass generic sentence encoders on that task? [A] Unsupervised averaging of token embeddings from vacancy texts with cluster-based normalization and no negative sampling. [B] A BERT-based encoder trained to reconstruct titles from taxonomies using manual labels, with attention but no negative sampling. [C] A BERT-based encoder trained with distant supervision to predict co-occurring skills from vacancy texts, using token-level gating and negative sampling. [D] A generic sentence encoder fine-tuned on unrelated corpora with taxonomy-specific nearest-centroid mapping.
**Reference Answer:** C
**OLMo Answer:** C
**Metrics:** Exact Match = 1.0, ROUGE = 1.0, Semantic Similarity = 1.0

**True/False Example 1** **Question:** Semantic representations of job titles learned by a BERT-based encoder trained with distant supervision to predict co-occurring skills from vacancy texts, using token-level gating and negative sampling, enable taxonomy-agnostic nearest-neighbor normalization without manual title labels and outperform generic sentence encoders on this task.
**Reference Answer:** True
**OLMo Answer:** True
**Metrics:** Exact Match = 1.0, ROUGE = 1.0, Semantic Similarity = 1.0

**True/False Example 2** **Question:** Semantic representations of job titles learned by a BERT-based encoder trained with distant supervision to predict co-occurring skills from vacancy texts, without token-level gating or negative sampling, require manual title labels for taxonomy-specific nearest-neighbor normalization and do not outperform generic sentence encoders on this task. **Reference Answer:** False
**OLMo Answer:** False
**Metrics:** Exact Match = 1.0, ROUGE = 1.0, Semantic Similarity = 1.0

**Fill in the Blanks Example 1** **Question:** The learned job-title representations support _____(taxonomy-agnostic/taxonomy-specific) nearest-neighbor normalization without manual title labels and outperform generic sentence

encoders, given they are produced by a BERT-based encoder trained with distant supervision to predict co-occurring skills from vacancy texts using token-level gating and negative sampling.
**Reference Answer:** taxonomy-agnostic
**OLMo Answer:** taxonomy-agnostic
**Metrics:** Exact Match = 1.0, ROUGE = 1.0, Semantic Similarity = 1.0

**Fill in the Blanks Example 2** **Question:** Semantic representations of job titles produced by a _____(BERT-based/LSTM-based) encoder trained with distant supervision to predict co-occurring skills from vacancy texts, with token-level gating and negative sampling, enable taxonomy-agnostic nearest-neighbor normalization without manual title labels and outperform generic sentence encoders on this task.
**Reference Answer:** worse non-English performance
**OLMo Answer:** worse non-English performance
**Metrics:** Exact Match = 1.0, ROUGE = 1.0, Semantic Similarity = 1.0

**Assertion Reason Example 1** **Question:** Assertion (A): Semantic representations of job titles learned by a BERT-based encoder trained with distant supervision to predict co-occurring skills from vacancy texts, using token-level gating and negative sampling, enable taxonomy-agnostic nearest-neighbor normalization without manual title labels and outperform generic sentence encoders on this task. Reason (R): Predicting co-occurring skills from vacancy texts under distant supervision aligns representations with job-title semantics, and token-level gating plus negative sampling increases discriminative power.
**Reference Answer:** A is True, R is True, and R explains A.
**OLMo Answer:** A is True, R is True, and R explains A.
**Metrics:** Exact Match = 1.0, ROUGE = 1.0, Semantic Similarity = 1.0

**Assertion Reason Example 2** **Question:** Assertion (A): Semantic representations of job titles learned by a BERT-based encoder trained with distant supervision to predict co-occurring skills from vacancy texts, using token-level gating and negative sampling, enable taxonomy-agnostic nearest-neighbor normalization without manual title labels and outperform generic sentence encoders on this task. Reason (R): Manual title labels and taxonomy-specific rules are required for training and normalization in this setting.
**Reference Answer:** A is True, R is False
**OLMo Answer:** A is True, R is False
**Metrics:** Exact Match = 1.0, ROUGE = 1.0, Semantic Similarity = 1.0

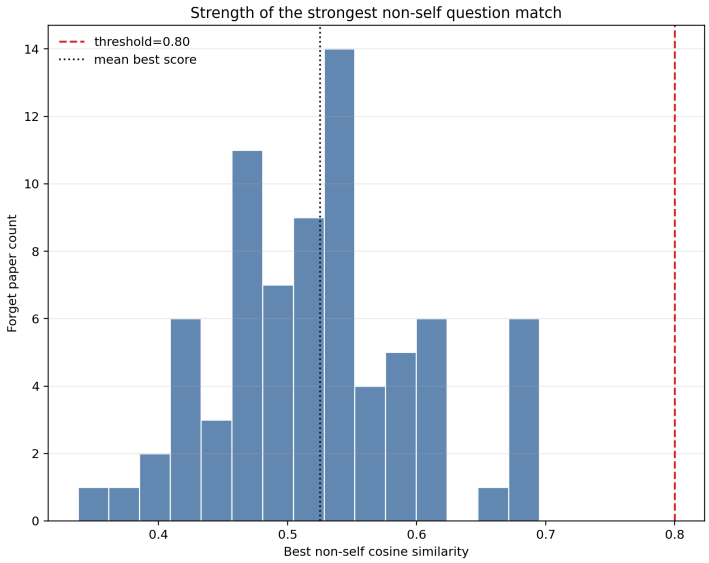



*Figure 6.* Distribution of the strongest non-self cosine similarity for each forget paper. The red dashed line indicates the overlap threshold ($\tau = 0.8$), and the black dotted line denotes the mean similarity.

# I. Distribution of Semantic Overlap Strength

We analyze the extent of semantic overlap between the forget and retain sets, the distribution of the strongest non-self question similarity for each forget paper is examined. Here, non-self denotes all retain papers excluding the retain paper that directly corresponds to the anchor paper itself. For each forget paper, the maximum cosine similarity between any of its questions and all questions from non-corresponding retain papers was computed. This yields a single "best non-self similarity" score per paper, capturing the strongest potential semantic overlap. Figure 6 shows the distribution of these scores across all forget papers. An empirical threshold of $\tau = 0.8$ is used to define significant semantic overlap.

The distribution is concentrated in the range of approximately $0.45$ to $0.65$, with a mean similarity around $0.53$. Importantly, no forget paper reaches or exceeds the threshold of $0.8$, indicating the absence of high-confidence semantic overlap between the forget and retain sets. The absence of high-similarity matches confirms that the retain set is effectively disjoint from the forget set at a semantic level. This validates the dataset construction process and ensures that any observed unlearning behavior cannot be attributed to direct leakage or duplication across the two sets.

# J. Implementation Details

## J.1. Gradient Difference (GD)

Gradient Difference modifies model parameters by performing gradient descent on the retain set while simultaneously performing gradient ascent on the forget set, thereby reducing the likelihood of generating target forget data (Neel et al., 2020). The update rule is defined in Equation 2.

$$\theta' = \theta - \eta\, w_r \nabla_\theta \mathcal{L}_{retain}(\theta) + \eta\, w_f \nabla_\theta \mathcal{L}_{forget}(\theta) \tag{2}$$

where $\theta$ represents model parameters, $\eta$ is the learning rate, $\mathcal{L}_{forget}$ denotes the loss on the forget set, and $\mathcal{L}_{retain}$ denotes the loss on a retain set to preserve useful knowledge.

## J.2. Negative Preference Optimization (NPO)

Negative Preference Optimization (Zhang et al., 2024; Bronec & Helcl, 2025) formulates unlearning as a preference learning problem, where the model is encouraged to prefer retained responses over those associated with the forget set. The objective is given in Equation 3.

$$\mathcal{L}_{\text{NPO}} = -\frac{2}{\beta}\, \mathbb{E}_{(x,y_f)} \left[ \log \sigma \left( \beta \, \log \frac{p_{\theta_0}(y_f|x)}{p_\theta(y_f|x)} \right) \right] \tag{3}$$

where $p_\theta(y|x)$ denotes the conditional likelihood of generating response $y$ given input $x$ under the model parameterized by $\theta$, and $p_{\theta_0}(y|x)$ denotes the corresponding likelihood under the reference (pretrained) model, which remains fixed during training. $y_f$ denotes responses corresponding to the forget set, $\sigma(\cdot)$ is the sigmoid function, $\beta$ is a scaling parameter controlling the sharpness of the preference.

## J.3. NPO with Retain Set Fine-Tuning

To improve stability and mitigate unintended degradation of useful knowledge, the NPO objective is augmented with an additional retain set optimization. While NPO suppresses the likelihood of forget samples relative to a reference model, it does not explicitly enforce the preservation of retained knowledge. Therefore, a complementary retain objective is introduced.

The retain loss is defined in Equation 4.

$$\mathcal{L}_{retain} = -\mathbb{E}_{(x,y_r)} \left[ \log p_\theta(y_r|x) \right] \tag{4}$$

The overall training objective is given in Equation 5.

$$\mathcal{L} = \lambda_{\text{NPO}} \mathcal{L}_{\text{NPO}} + \lambda_r \mathcal{L}_{retain} \tag{5}$$

where $\lambda_{\text{NPO}}$ and $\lambda_r$ control the relative contributions of the unlearning and retention objectives, respectively.

### J.4. SimNPO (Simplified Negative Preference Optimization)

SimNPO (Fan et al., 2024) is a simplified variant of preference-based unlearning that operates without a reference model or explicit retain objective. Unlike standard NPO, which compares the current model against a fixed reference, SimNPO directly penalizes the likelihood of generating forget samples, making it a fully self-contained and reference-free unlearning approach.

The objective is defined in Equation 6.

$$\mathcal{L}_{\text{SimNPO}} = -\frac{2}{\beta}\,\mathbb{E}_{(x,y_f)}\left[\log\sigma\left(-\beta\,\log p_\theta(y_f|x) - \gamma\right)\right], \tag{6}$$

where $p_\theta(y_f|x)$ denotes the average log-likelihood of the response $y_f$ given input $x$, computed over answer tokens only. Here, $\sigma(\cdot)$ is the sigmoid function, $\beta$ controls the sharpness of the penalty, and $\gamma$ is a margin term.

This formulation directly encourages the model to reduce the likelihood of forget samples without relying on external comparisons or paired preferences. By operating on the average log-probability over answer tokens, SimNPO focuses on suppressing the generation of target responses at a sequence level.

### J.5. SimNPO with Retain Set Fine-Tuning (SimNPO+RT)

To alleviate unintended forgetting and better preserve useful knowledge, SimNPO is extended with an explicit retain set objective. While SimNPO effectively suppresses forget samples by directly penalizing their likelihood, it does not impose any constraint to maintain performance on retained data. To address this limitation, a retain loss is incorporated alongside the SimNPO objective.

The retain loss is defined identically to Equation 4.Thus the combined SimNPO+RT objective is then given in Equation 7

$$\mathcal{L} = \lambda_{\text{SimNPO}}\mathcal{L}_{\text{SimNPO}} + \lambda_r\mathcal{L}_{\text{retain}}, \tag{7}$$

Here, $\lambda_{\text{SimNPO}}$ and $\lambda_r$ balance the contributions of the unlearning and retention objectives, respectively. By explicitly encouraging correct predictions on the retain set, SimNPO+RT improves training stability and helps maintain overall model utility while remaining free of reference-model comparisons.

## K. Hyperparameter Selection for Algorithm Implementation

*Table 5.* Hyperparameters for unlearning methods under forget sets $\mathcal{F}_1$ and $\mathcal{F}_2$. LR is Learning Rate; $\beta$ is preference strength in NPO; $\gamma$ scales retain regularization in SimNPOs; $w_f$ and $w_r$ is weights for forget and retain objectives.

| **Param** | **GD** | | **NPO** | | **NPO+RT** | | **SimNPO** | | **SimNPO+RT** | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\mathcal{F}_1$ | $\mathcal{F}_2$ | $\mathcal{F}_1$ | $\mathcal{F}_2$ | $\mathcal{F}_1$ | $\mathcal{F}_2$ | $\mathcal{F}_1$ | $\mathcal{F}_2$ | $\mathcal{F}_1$ | $\mathcal{F}_2$ |
| LR | $2e-5$ | $1e-5$ | $1e-5$ | $1e-5$ | $1e-5$ | $1e-5$ | $2e-5$ | $2e-5$ | $2e-5$ | $2e-5$ |
| Epochs | 50 | 10 | 10 | 10 | 30 | 15 | 15 | 15 | 30 | 10 |
| $\beta$ | – | – | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 |
| $\gamma$ | – | – | – | – | – | – | 0.0 | 0.0 | 0.0 | 0.0 |
| $w_f, w_r$ | (1.0, 3.0) | (1.2, 1.5) | (1.0, 0.0) | (1.0, 0.0) | (1.0, 4.0) | (1.2, 0.75) | (1.0, 4.0) | (1.0, 4.0) | (1.0, 4.0) | (1.0, 0.25) |

As shown in Table 5, hyperparameters are chosen specifically to balance the trade-off between forgetting and retention, ensuring that the unlearning process effectively removes targeted information while preserving performance on retained data. Values are selected to achieve sufficient forgetting without inducing over-forgetting, which would unnecessarily degrade retained knowledge, thereby maintaining an optimal equilibrium between removal efficacy and utility preservation. Besides, different algorithms require distinct hyperparameters due to inherent differences in their unlearning dynamics and sensitivity to the forgetting–retention trade-off.

## L. Examples of Unlearning Across Algorithms

To demonstrate the effectiveness of unlearning, we present representative examples from the forget set. All algorithms consistently produce incorrect outputs, indicating successful forgetting.

*Table 6.* Examples showing consistent incorrect responses across different unlearning algorithms.

**Example 1**

**Question:** For Z24 Bridge natural frequency data, time-delay embeddings of the raw series exhibit a toroidal topology driven by cyclic temperature effects, whereas embeddings of cointegrated residuals exhibit an open-ball topology characteristic of approximately Gaussian white-noise processes.

**Ground Truth:** True

**GD:** False **NPO:** False **NPO+RT:** False **SimNPO:** False **SimNPO+RT:** False

**Example 2**

**Question:** When the second natural frequency depends nonlinearly on temperature, linear cointegration leaves residual topological structure, whereas Gaussian process-based nonlinear cointegration trained on data that include the nonlinear regime more effectively removes these effects.

**Ground Truth:** True

**GD:** False **NPO:** False **NPO+RT:** False **SimNPO:** False **SimNPO+RT:** False

**Example 3**

**Question:** In a linear classifier with an additive patch trigger, the optimal untargeted adversarial perturbation has a _____(substantial projection/near-zero projection) onto the trigger direction, theoretically explaining the similarity between adversarial and triggered inputs.

**Ground Truth:** substantial

**GD:** near-zero **NPO:** near-zero **NPO+RT:** near-zero **SimNPO:** near-zero **SimNPO+RT:** near-zero

**Example 4**

**Question:** Choose the correct option (A–D): When predicted skill importance scores are multiplied by the inverse document frequency of each skill, what is the resulting effect on ranking? [A] It enhances the ranking of specialized skills by penalizing generic skills. [B] It enhances the ranking of generic skills by penalizing specialized skills. [C] It leaves the ranking unchanged regardless of skill specificity. [D] It randomizes the ranking without regard to skill frequency.

**Ground Truth:** A

**GD:** B **NPO:** B **NPO+RT:** B **SimNPO:** B **SimNPO+RT:** B

**Example 5**

**Question:** Fine-tuning a multilingual sentence encoder on English job-title data increases skill importance ranking performance for English titles while lowering performance for non-English titles.

**Ground Truth:** True

**GD:** False **NPO:** False **NPO+RT:** False **SimNPO:** False **SimNPO+RT:** False

## M. Full Parameter Unlearning

For each unlearning algorithm (full parameter unlearning) the result is shown in Table 7 for all experimental setting (i.e., training with Forget Set 1 and Forget Set 2), also training dynamics are visualized using three complementary plots arranged side-by-side. This unified visualization enables direct comparison of optimization behavior across loss components and accuracy trends.

For every experiments, the following three plots are jointly presented. Loss Decomposition that tracks the evolution of the unlearning objective and retain loss across training epochs. Accuracy trade-off that shows the change in forget accuracy and retain accuracy, capturing the balance between knowledge removal and preservation. Total training loss that illustrates overall optimization convergence over epochs.

Figures 7 to 16 illustrate the training behavior of full-parameter unlearning using the GD, NPO, NPO+RT, SimNPO and SimNPO+RT on Forget Sets $\mathcal{F}_1$ and $\mathcal{F}_2$, respectively. In GD, the loss decomposition plots show a steady increase in forget loss while the retain loss remains relatively stable, indicating that the unlearning objective effectively targets forget examples without inducing large optimization pressure on retained data. Similarly in NPO and SimNPO, the forget loss decreases, which aligns with the expected behaviour. The forget-versus-retain accuracy plots further support this observation: accuracy on forget samples consistently degrades over epochs, while

*Table 7.* Full Parameter Unlearning. R-F1(%): ROUGE-F1, EM(%): Exact Match, AC: Arc Challenge, HS: HellaSwag., ↓: lower is better (forgetting), and ↑: higher is better (retention and generalization), **Best** and **Second Best**

| Method | Forget Set 1 ↓ | | Forget Set 2 ↓ | | Retain (Ext) ↑ | | Retain (Int) ↑ | | General Benchmarks | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | R-F1 | EM | R-F1 | EM | R-F1 | EM | R-F1 | EM | MMLU | Arc Challenge | HellaSwag |
| Base Model | 97.54 | 94.21 | 93.92 | 89.17 | 95.40 | 90.75 | 92.44 | 83.49 | 58.75 | 54.35 | 76.02 |
| GD ($\mathcal{F}_1$) | **71.33** | **55.59** | 91.73 | 87.12 | **95.68** | **93.01** | **93.70** | **85.40** | 56.23 | 52.30 | 72.40 |
| NPO ($\mathcal{F}_1$) | 94.92 | 89.55 | 94.58 | 89.36 | 94.29 | 88.67 | **92.87** | 83.73 | 57.59 | 50.76 | 75.47 |
| NPO+RT ($\mathcal{F}_1$) | 81.41 | 74.81 | 93.98 | 90.11 | 91.33 | 83.20 | 90.70 | 75.83 | 57.71 | 52.30 | 75.00 |
| SimNPO ($\mathcal{F}_1$) | 96.26 | 91.23 | 93.94 | 89.36 | 94.18 | 88.77 | 92.84 | 82.77 | 58.24 | **52.64** | **75.94** |
| SimNPO+RT ($\mathcal{F}_1$) | **73.45** | **64.73** | 92.58 | 89.17 | **94.84** | 89.24 | 92.34 | 81.10 | **58.96** | 52.47 | 74.50 |
| GD ($\mathcal{F}_2$) | 95.57 | 92.35 | **61.59** | **53.91** | 94.61 | **90.84** | 92.79 | **84.44** | 56.05 | 49.65 | 64.63 |
| NPO ($\mathcal{F}_2$) | 96.78 | 91.79 | 87.01 | 78.54 | 92.89 | 86.03 | 91.74 | 82.29 | 57.45 | 51.87 | 75.66 |
| NPO+RT ($\mathcal{F}_2$) | 96.88 | 92.53 | 87.09 | 77.79 | 92.50 | 85.28 | 91.81 | 82.29 | 58.01 | 52.13 | **75.73** |
| SimNPO ($\mathcal{F}_2$) | 96.44 | 91.79 | 87.01 | 78.54 | 92.89 | 86.03 | 91.74 | 82.29 | 57.45 | 51.87 | 75.66 |
| SimNPO+RT ($\mathcal{F}_2$) | 96.09 | 92.35 | **65.90** | **46.08** | 92.11 | 83.11 | 92.37 | 82.05 | **59.14** | **52.73** | 72.45 |

retain accuracy remains largely preserved, demonstrating localized forgetting rather than indiscriminate performance collapse. The total training loss curves decrease smoothly without instability, suggesting that the optimization process converges reliably despite the competing objectives.

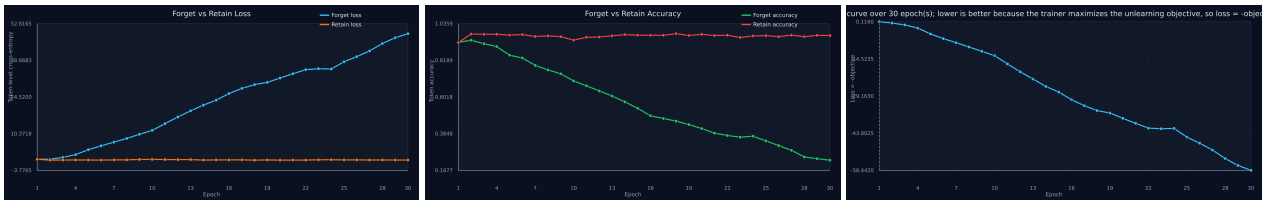



*Figure 7.* Training dynamics for full parameter unlearning with Gradient Difference Algorithm and Forget Set 1. From left to right: (1) loss decomposition, (2) forget vs retain accuracy, and (3) total training loss across epochs.

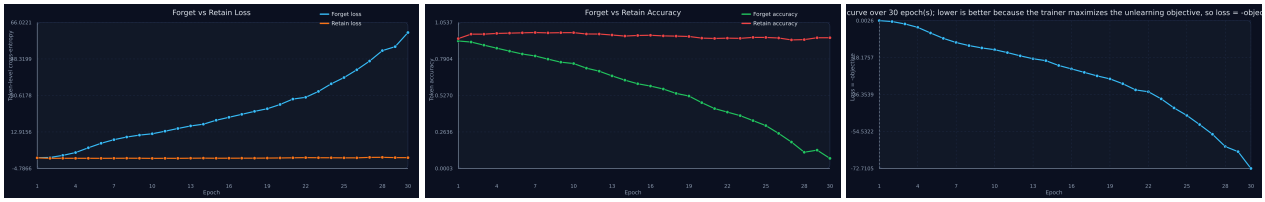



*Figure 8.* Training dynamics for full parameter unlearning with Gradient Difference Algorithm and Forget Set 2. From left to right: (1) loss decomposition, (2) forget vs retain accuracy, and (3) total training loss across epochs.

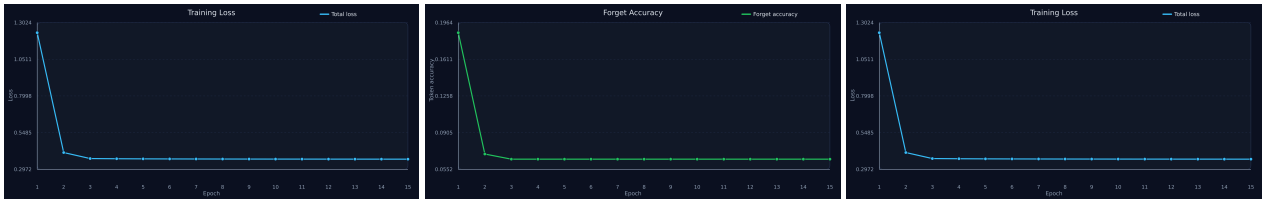



*Figure 9.* Training dynamics for full parameter unlearning with NPO Algorithm and Forget Set 1. From left to right: (1) loss decomposition, (2) forget vs retain accuracy, and (3) total training loss across epochs.

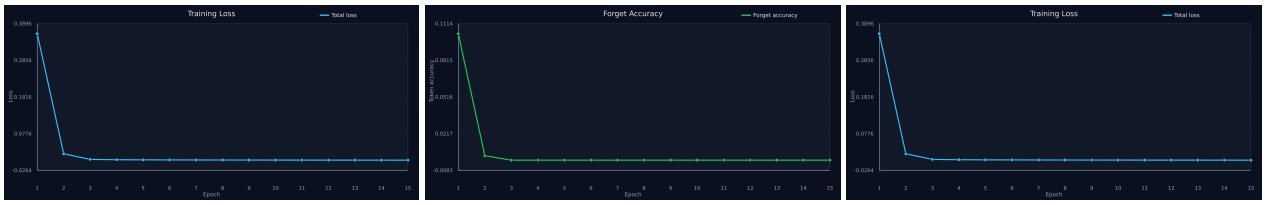



*Figure 10.* Training dynamics for full parameter unlearning with NPO Algorithm and Forget Set 2. From left to right: (1) loss decomposition, (2) forget vs retain accuracy, and (3) total training loss across epochs.

*Figure 11.* Training dynamics for full parameter unlearning with NPO+RT Algorithm and Forget Set 1. From left to right: (1) loss decomposition, (2) forget vs retain accuracy, and (3) total training loss across epochs.

*Figure 12.* Training dynamics for full parameter unlearning with NPO+RT Algorithm and Forget Set 2. From left to right: (1) loss decomposition, (2) forget vs retain accuracy, and (3) total training loss across epochs.

*Figure 13.* Training dynamics for full parameter unlearning with SimNPO Algorithm and Forget Set 1. From left to right: (1) loss decomposition, (2) forget vs retain accuracy, and (3) total training loss across epochs.

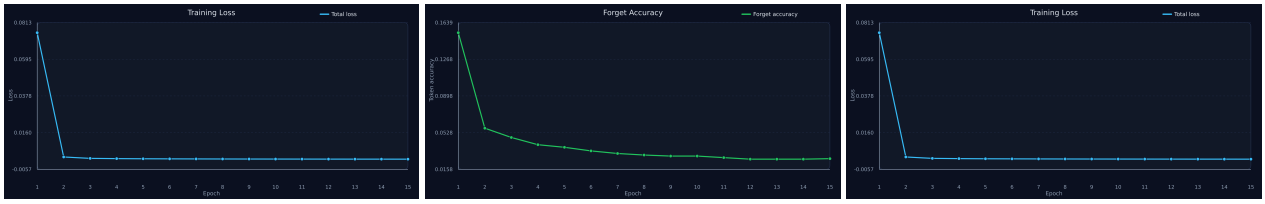



*Figure 14.* Training dynamics for full parameter unlearning with SimNPO Algorithm and Forget Set 2. From left to right: (1) loss decomposition, (2) forget vs retain accuracy, and (3) total training loss across epochs.

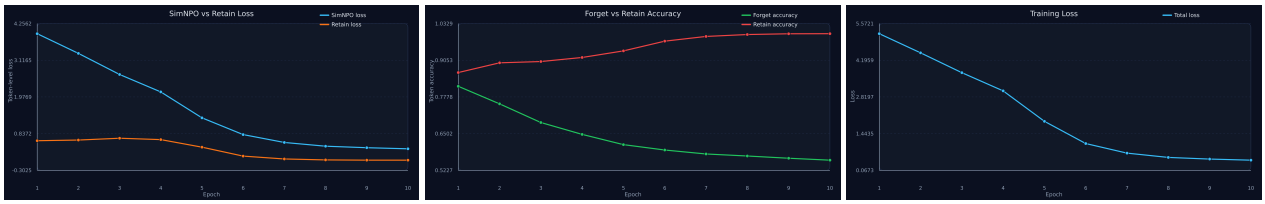



*Figure 15.* Training dynamics for full parameter unlearning with SimNPO+RT Algorithm and Forget Set 1. From left to right: (1) loss decomposition, (2) forget vs retain accuracy, and (3) total training loss across epochs.

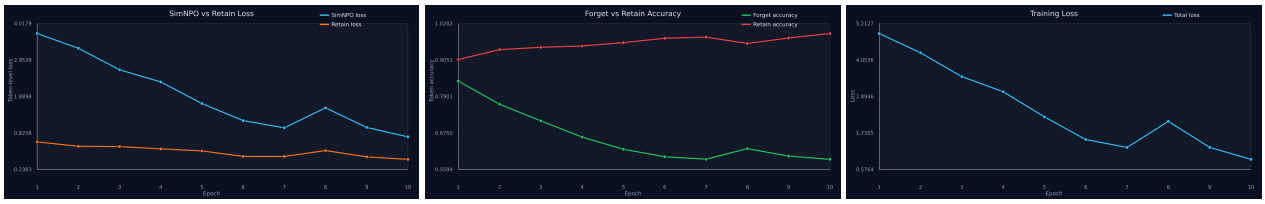



*Figure 16.* Training dynamics for full parameter unlearning with SimNPO+RT Algorithm and Forget Set 2. From left to right: (1) loss decomposition, (2) forget vs retain accuracy, and (3) total training loss across epochs.

## N. Qualitative Analysis of Unlearning

This section presents a detailed qualitative analysis of unlearning across Gradient Difference (GD), NPO+RT, and SimNPO+RT under both Forget Set 1 and Forget Set 2 training settings. The analysis combines quantitative error distributions with behavioral observations to understand how unlearning manifests at the level of scientific claims.

### N.1. Error Distribution Across Question Formats

*Table 8.* Unified error distribution across all methods and evaluation splits. Cells show errors / total (error rate %).

| Method | Split | MCQ | True/False | Fill in the Blanks | Assertion Reason |
|---|---|---|---|---|---|
| *GD trained on Forget Set 1* | | | | | |
| GD-FS1 | Forget Set 1 | 21/138 (15.2%) | 68/101 (67.3%) | 49/127 (38.6%) | 86/170 (50.6%) |
| | Forget Set 2 | 9/138 (6.5%) | 0/101 (0.0%) | 22/127 (17.3%) | 30/170 (17.6%) |
| | Retain_ext | 2/284 (0.7%) | 32/180 (17.8%) | 22/254 (8.7%) | 112/342 (32.8%) |
| | Retain_int | 0/129 (0.0%) | 14/126 (11.1%) | 9/11 (81.8%) | 94/152 (61.8%) |
| *GD trained on Forget Set 2* | | | | | |
| GD-FS2 | Forget Set 1 | 1/138 (0.7%) | 1/101 (1.0%) | 10/127 (7.9%) | 33/170 (19.4%) |
| | Forget Set 2 | 4/138 (2.9%) | 50/101 (49.5%) | 15/127 (11.8%) | 35/170 (20.6%) |
| | Retain_ext | 8/284 (2.8%) | 40/180 (22.2%) | 32/254 (12.6%) | 72/342 (21.1%) |
| | Retain_int | 3/129 (2.3%) | 7/126 (5.6%) | 11/11 (100.0%) | 63/152 (41.4%) |
| *NPO+RT trained on Forget Set 1* | | | | | |
| NPO+RT-FS1 | Forget Set 1 | 7/138 (5.1%) | 35/101 (34.7%) | 25/127 (19.7%) | 32/170 (18.8%) |
| | Forget Set 2 | 5/138 (3.6%) | 1/101 (1.0%) | 35/127 (27.6%) | 15/170 (8.8%) |
| | Retain_ext | 6/284 (2.1%) | 30/180 (16.7%) | 63/254 (24.8%) | 48/342 (14.0%) |
| | Retain_int | 3/129 (2.3%) | 13/126 (10.3%) | 11/11 (100.0%) | 48/152 (31.6%) |
| *NPO+RT trained on Forget Set 2* | | | | | |
| NPO+RT-FS2 | Forget Set 1 | 5/138 (3.6%) | 1/101 (1.0%) | 13/127 (10.2%) | 35/170 (20.6%) |
| | Forget Set 2 | 2/138 (1.4%) | 47/101 (46.5%) | 25/127 (19.7%) | 21/170 (12.4%) |
| | Retain_ext | 10/284 (3.5%) | 37/180 (20.6%) | 51/254 (20.1%) | 59/342 (17.3%) |
| | Retain_int | 3/129 (2.3%) | 7/126 (5.6%) | 10/11 (90.9%) | 56/152 (36.8%) |

| Method | Split | MCQ | True/False | Fill in the Blanks | Assertion Reason |
|---|---|---|---|---|---|
| *SimNPO+RT trained on Forget Set 1* | | | | | |
| SimNPO+RT-FS1 | Forget Set 1 | 13/138 (9.4%) | 38/101 (37.6%) | 33/127 (26.0%) | 77/170 (45.3%) |
| | Forget Set 2 | 7/138 (5.1%) | 0/101 (0.0%) | 33/127 (26.0%) | 15/170 (8.8%) |
| | Retain_ext | 5/284 (1.8%) | 26/180 (14.4%) | 66/254 (26.0%) | 84/342 (24.6%) |
| | Retain_int | 4/129 (3.1%) | 14/126 (11.1%) | 11/11 (100.0%) | 66/152 (43.4%) |
| *SimNPO+RT trained on Forget Set 2* | | | | | |
| SimNPO+RT-FS2 | Forget Set 1 | 6/138 (4.3%) | 4/101 (4.0%) | 20/127 (15.7%) | 26/170 (15.3%) |
| | Forget Set 2 | 3/138 (2.2%) | 12/101 (11.9%) | 30/127 (23.6%) | 26/170 (15.3%) |
| | Retain_ext | 12/284 (4.2%) | 13/180 (7.2%) | 54/254 (21.3%) | 66/342 (19.3%) |
| | Retain_int | 4/129 (3.1%) | 2/126 (1.6%) | 10/11 (90.9%) | 56/152 (36.8%) |

As shown is Table 8, across all methods, error rates are significantly higher on the forget set used during unlearning, indicating effective degradation on seen instances. However, when evaluated on a disjoint forget set, error rates drop sharply. This suggests that unlearning does not generalize well to semantically equivalent but differently phrased questions, and remains largely confined to the exact training instances.

Unlearning effectiveness varies across question formats. Structured formats such as True/False and Assertion/Reason exhibit higher error rates, indicating stronger forgetting. In contrast, MCQ and Fill-in-the-blank formats show relatively lower degradation, suggesting that these formats require deeper semantic understanding and are more resistant to unlearning. Overall, current methods appear to induce surface-level suppression rather than removing underlying claim-level knowledge.

Analysis on retain sets reveals a consistent pattern across methods. The highest degradation is observed for Fill-in-the-blank questions, followed by Assertion/Reason. MCQ questions are the most retained, while True/False lies in between. This indicates that retention is strongest for formats requiring recognition rather than generation.

A recurring failure mode is observed in Assertion–Reason questions, where the model struggles with explanation linkage. Although the model often correctly evaluates both the assertion and the reason individually, it fails to determine whether the reason logically explains the assertion.

- Reference: `A is True, R is True, and R explains A`
- Model: `A is True, R is True, but R does not explain A`

This behavior suggests that while factual knowledge is retained, the *relational structure* between concepts is degraded during unlearning.

## N.2. Effect of Claim Complexity

Unlearning disproportionately affects short, simple claims that require a single polarity decision. In contrast, longer and compositional claims involving mechanisms and dependencies are more robust.

### N.2.1. Effect of Claim Complexity on Unlearning

These longer claims, which involve multiple conditions, mechanisms, and dependencies, are more robust to unlearning. This suggests that while surface-level or atomic claims are easily suppressed, deeper compositional knowledge structures remain largely intact.

*Table 9.* Examples of complex claims that are largely retained after unlearning.

| Example | Observed Outcome | Interpretation |
|---|---|---|
| A ConvLSTM-augmented fully convolutional network trained on composited video frames can segment facial components in video sequences without manual pose or contour annotations. | Correct | Long, procedural claim with multiple linked conditions |
| A model-based approach uses UML state machine and class diagrams to automatically derive executable end-to-end test scripts for IoT systems. | Correct | Multi-step methodological claim with explicit construction |
| Few-shot image classification can be performed by representing objects as compositions of shared component embeddings and spatial activation patterns. | Correct | Long claim involving mechanism, constraints, and outcome |

*Table 10.* Examples of complex claims that are largely retained after unlearning.

| Example | Observed Outcome | Interpretation |
|---|---|---|
| A ConvLSTM-augmented fully convolutional network trained on composited video frames can segment facial components in video sequences without manual pose or contour annotations. | Correct | Long, procedural claim with multiple linked conditions |
| A model-based approach uses UML state machine and class diagrams to automatically derive executable end-to-end test scripts for IoT systems. | Correct | Multi-step methodological claim with explicit construction |
| Few-shot image classification can be performed by representing objects as compositions of shared component embeddings and spatial activation patterns. | Correct | Long claim involving mechanism, constraints, and outcome |

## N.3. Top Error-Prone Papers Across Algorithms

### N.3.1. GD (Trained on Forget Set 1)

*Table 11.* Top error-prone papers per method and training forget set.

| Paper | Wrong | Total |
|---|---|---|
| *GD trained on Forget Set 1* | | |
| How to measure uncertainty in uncertainty sampling for active learning | 11 | 32 |
| Deep Insights into Convolutional Networks for Video Recognition | 11 | 30 |
| Approximation of quantum control correction scheme using deep neural networks | 11 | 26 |
| Math-word embedding in math search and semantic extraction | 10 | 38 |
| Distributed deep learning platform for pedestrian detection on IT convergence environment | 10 | 32 |
| A hybrid model for EEG-based gender recognition | 10 | 30 |
| Hierarchical sparse Cholesky decomposition for spatio-temporal filtering | 10 | 30 |
| Is it Required? Ranking the Skills Required for a Job-Title | 8 | 26 |
| Disk Cluster Allocation Behavior in Windows and NTFS | 8 | 14 |
| Web-Based Volunteer Distributed Computing for Handling Time-Critical Urgent Workloads | 7 | 32 |
| *GD trained on Forget Set 2* | | |
| Deep Insights into Convolutional Networks for Video Recognition | 7 | 30 |
| Math-word embedding in math search and semantic extraction | 6 | 38 |
| How to measure uncertainty in uncertainty sampling for active learning | 6 | 32 |
| Reducing false wake-up in contention-based wake-up control of wireless LANs | 6 | 20 |
| Real-time face alignment: evaluation methods, training strategies and implementation optimization | 5 | 28 |
| ReliefE: feature ranking in high-dimensional spaces via manifold embeddings | 5 | 28 |
| Detection of agglomerate fog based on a shallow convolutional neural network | 5 | 20 |
| Robust Bayesian model selection for variable clustering with the Gaussian graphical model | 5 | 18 |
| GPU acceleration of splitting schemes applied to differential matrix equations | 4 | 16 |
| LSB steganography detection in monochromatic still images using artificial neural networks | 4 | 12 |
| *NPO+RT trained on Forget Set 1* | | |
| How to measure uncertainty in uncertainty sampling for active learning | 8 | 32 |
| Is it Required? Ranking the Skills Required for a Job-Title | 8 | 26 |
| Disk Cluster Allocation Behavior in Windows and NTFS | 8 | 14 |
| Math-word embedding in math search and semantic extraction | 7 | 38 |
| Deep Insights into Convolutional Networks for Video Recognition | 7 | 30 |
| ReliefE: feature ranking in high-dimensional spaces via manifold embeddings | 7 | 28 |
| Reducing false wake-up in contention-based wireless LANs | 6 | 20 |
| Distributed deep learning platform for pedestrian detection on IT convergence environment | 5 | 32 |
| LSB steganography detection in monochromatic still images using artificial neural networks | 5 | 12 |
| Hierarchical sparse Cholesky decomposition with applications to high-dimensional spatio-temporal filtering | 4 | 30 |
| *NPO+RT trained on Forget Set 2* | | |

*Continued on next page*

*(continued)*

| Paper | Wrong | Total |
|---|---|---|
| How to measure uncertainty in uncertainty sampling for active learning | 6 | 32 |
| Deep Insights into Convolutional Networks for Video Recognition | 6 | 30 |
| ReliefE: feature ranking in high-dimensional spaces via manifold embeddings | 6 | 28 |
| Reducing false wake-up in contention-based wake-up control of wireless LANs | 6 | 20 |
| Detection of agglomerate fog based on a shallow convolutional neural network | 5 | 20 |
| A hybrid model for EEG-based gender recognition | 4 | 30 |
| Hierarchical sparse Cholesky decomposition with applications to high-dimensional spatio-temporal filtering | 4 | 30 |
| Approximation of quantum control correction scheme using deep neural networks | 4 | 26 |
| Is it Required? Ranking the Skills Required for a Job-Title | 4 | 26 |
| A topological analysis of cointegrated data: a Z24 Bridge case study | 4 | 18 |
| *SimNPO+RT trained on Forget Set 1* | | |
| How to measure uncertainty in uncertainty sampling for active learning | 14 | 32 |
| Math-word embedding in math search and semantic extraction | 9 | 38 |
| Deep Insights into Convolutional Networks for Video Recognition | 9 | 30 |
| ReliefE: feature ranking in high-dimensional spaces via manifold embeddings | 7 | 28 |
| Disk Cluster Allocation Behavior in Windows and NTFS | 7 | 14 |
| Distributed deep learning platform for pedestrian detection on IT convergence environment | 6 | 32 |
| Hierarchical sparse Cholesky decomposition with applications to high-dimensional spatio-temporal filtering | 6 | 30 |
| Is it Required? Ranking the Skills Required for a Job-Title | 6 | 26 |
| A topological analysis of cointegrated data: A Z24 Bridge case study | 6 | 18 |
| Progressive Backdoor Erasing via connecting Backdoor and Adversarial Attacks | 6 | 14 |
| *SimNPO+RT trained on Forget Set 2* | | |
| Deep Insights into Convolutional Networks for Video Recognition | 7 | 30 |
| ReliefE: feature ranking in high-dimensional spaces via manifold embeddings | 6 | 28 |
| How to measure uncertainty in uncertainty sampling for active learning | 5 | 32 |
| Hierarchical sparse Cholesky decomposition with applications to high-dimensional spatio-temporal filtering | 4 | 30 |
| Is it Required? Ranking the Skills Required for a Job-Title | 4 | 26 |
| Reducing false wake-up in contention-based wake-up control of wireless LANs | 4 | 20 |
| Robust Bayesian model selection for variable clustering with the Gaussian graphical model | 4 | 18 |
| GPU acceleration of splitting schemes applied to differential matrix equations | 4 | 16 |
| Math-word embedding in math search and semantic extraction | 3 | 38 |
| Approximation of quantum control correction scheme using deep neural networks | 3 | 26 |

Across all algorithms, a consistent subset of papers repeatedly appears among the most error-prone:

- **Highly consistent across all settings:**
  - How to measure uncertainty in uncertainty sampling for active learning
  - Deep Insights into Convolutional Networks for Video Recognition
  - Math-word embedding in math search and semantic extraction
  - ReliefE: feature ranking in high-dimensional spaces via manifold embeddings
- **Frequently appearing:**
  - Is it Required? Ranking the Skills Required for a Job-Title
  - Hierarchical sparse Cholesky decomposition with applications to high-dimensional spatio-temporal filtering
- **Moderately frequent:**
  - Distributed deep learning platform for pedestrian detection on IT convergence environment
  - Disk Cluster Allocation Behavior in Windows and NTFS
  - Reducing false wake-up in contention-based wake-up control of wireless LANs

- Approximation of quantum control correction scheme using deep neural networks

These papers typically contain simple, atomic claims with a single decision boundary. Their repeated occurrence across methods indicates that errors arise from *intrinsic claim fragility* rather than algorithm-specific weaknesses.

#### N.3.2. Question Density Analysis of Error-Prone Papers

The 10 most error-prone papers exhibit a substantially higher question density, with an average of 27.6 questions per paper. All selected papers lie above the dataset median, and 9 out of 10 exceed the 80th percentile, indicating that they are significantly overrepresented in terms of question coverage.

*Table 12.* Question density statistics for the most error-prone papers.

| Paper | Questions | Percentile |
|---|---|---|
| How to measure uncertainty in uncertainty sampling for active learning | 32 | 98.7 |
| Deep Insights into Convolutional Networks for Video Recognition | 30 | 94.7 |
| Math-word embedding in math search and semantic extraction | 38 | 100.0 |
| ReliefE: feature ranking in high-dimensional spaces via manifold embeddings | 28 | 89.5 |
| Is it Required? Ranking the Skills Required for a Job-Title | 26 | 86.8 |
| Hierarchical sparse Cholesky decomposition with applications to high-dimensional spatio-temporal filtering | 30 | 94.7 |
| Distributed deep learning platform for pedestrian detection on IT convergence environment | 32 | 98.7 |
| Disk Cluster Allocation Behavior in Windows and NTFS | 14 | 61.8 |
| Reducing false wake-up in contention-based wake-up control of wireless LANs | 20 | 81.6 |
| Approximation of quantum control correction scheme using deep neural networks | 26 | 86.8 |

Among these, *Math-word embedding in math search and semantic extraction* is the most densely represented paper with 38 questions, corresponding to the 100th percentile. In contrast, *Disk Cluster Allocation Behavior in Windows and NTFS* has the lowest count among the selected set (14 questions), yet it still lies above the dataset median.

These findings indicate that highly represented papers are more likely to appear among the most error-prone cases, suggesting that both exposure frequency and claim density contribute to unlearning difficulty.

### N.4. Retain External Set: Error Analysis Across Algorithms

To analyze the impact of unlearning on retained knowledge, the most error-prone papers in the *retain_external* split are examined across different algorithms and training settings.

*Table 13.* Top error-prone papers on external retain set across methods and training forget sets.

| Paper | Wrong / Total |
|---|---|
| *GD trained on Forget Set 1* | |
| An Empirical Study of Recent Face Alignment Methods | 9 / 34 |
| Evolution of the social network of scientific collaborations | 8 / 32 |
| Recurrent Neural Networks with Top-k Gains for Session-based Recommendations | 7 / 20 |
| Preliminary Exploration of Formula Embedding for Mathematical Information Retrieval: can mathematical formulae be embedded like a natural language? | 7 / 28 |
| Dynamic SDN-Based Radio Access Network Slicing With Deep Reinforcement Learning for URLLC and eMBB Services | 5 / 24 |
| Adversarial Machine Learning for 5G Communications Security | 5 / 18 |
| Towards Massively Parallel Computations in Algebraic Geometry | 5 / 24 |
| Quantifying sources of error in McPAT and potential impacts on architectural studies | 5 / 28 |
| The Perception of Technical Debt in the Embedded Systems Domain: An Industrial Case Study | 5 / 24 |
| Efficient Non-greedy Optimization of Decision Trees | 4 / 26 |

*Continued on next page*

*(continued)*

| Paper | Wrong / Total |
|---|---|
| *GD trained on Forget Set 2* | |
| Group Sparse Priors for Covariance Estimation | 9 / 34 |
| Towards Massively Parallel Computations in Algebraic Geometry | 6 / 24 |
| Recurrent Neural Networks with Top-k Gains for Session-based Recommendations | 6 / 20 |
| Dynamic SDN-Based Radio Access Network Slicing With Deep Reinforcement Learning for URLLC and eMBB Services | 5 / 24 |
| Adversarial Machine Learning for 5G Communications Security | 5 / 18 |
| An Empirical Study of Recent Face Alignment Methods | 5 / 34 |
| Efficient Non-greedy Optimization of Decision Trees | 5 / 26 |
| Autonomous Mobility-on-Demand Systems for Future Urban Mobility | 5 / 32 |
| *NPO+RT trained on Forget Set 1* | |
| An Empirical Study of Recent Face Alignment Methods | 9 / 34 |
| Group Sparse Priors for Covariance Estimation | 6 / 34 |
| Dynamic SDN-Based Radio Access Network Slicing With Deep Reinforcement Learning for URLLC and eMBB Services | 5 / 24 |
| Ensemble Kalman Methods for High-Dimensional Hierarchical Dynamic Space-Time Models | 5 / 18 |
| Multi-step prediction for influenza outbreak by an adjusted long short-term memory | 5 / 16 |
| The Perception of Technical Debt in the Embedded Systems Domain: An Industrial Case Study | 5 / 24 |
| Evolution of the social network of scientific collaborations | 5 / 32 |
| Efficient detection of contagious outbreaks in massive metropolitan encounter networks | 5 / 20 |
| Using Recurrent Neural Networks to Optimize Dynamical Decoupling for Quantum Memory | 5 / 16 |
| Fuzzy Jaccard Index: A robust comparison of ordered lists | 4 / 22 |
| *NPO+RT trained on Forget Set 2* | |
| Dynamic SDN-Based Radio Access Network Slicing With Deep Reinforcement Learning for URLLC and eMBB Services | 7 / 24 |
| Group Sparse Priors for Covariance Estimation | 7 / 34 |
| An Empirical Study of Recent Face Alignment Methods | 6 / 34 |
| Recurrent Neural Networks with Top-k Gains for Session-based Recommendations | 6 / 20 |
| Efficient Non-greedy Optimization of Decision Trees | 5 / 26 |
| Towards Massively Parallel Computations in Algebraic Geometry | 5 / 24 |
| Efficient detection of contagious outbreaks in massive metropolitan encounter networks | 5 / 20 |
| V-BOINC virtualization framework | 4 / 26 |
| Self-Aware Personalized Federated Learning | 4 / 36 |
| *SimNPO+RT trained on Forget Set 1* | |
| Evolution of the social network of scientific collaborations | 12 / 32 |
| An Empirical Study of Recent Face Alignment Methods | 8 / 34 |
| Efficient Non-greedy Optimization of Decision Trees | 6 / 26 |
| Recurrent Neural Networks with Top-k Gains for Session-based Recommendations | 6 / 20 |
| Dynamic SDN-Based Radio Access Network Slicing With Deep Reinforcement Learning for URLLC and eMBB Services | 5 / 24 |
| Adversarial Machine Learning for 5G Communications Security | 5 / 18 |
| Fuzzy Jaccard Index: A robust comparison of ordered lists | 5 / 22 |
| Towards Massively Parallel Computations in Algebraic Geometry | 5 / 24 |
| Trust as indicator of robot functional and social acceptance. An experimental study on user conformation to iCub answers | 5 / 20 |
| Adaptive high-order splitting schemes for large-scale differential Riccati equations | 5 / 18 |
| *SimNPO+RT trained on Forget Set 2* | |
| Dynamic SDN-Based Radio Access Network Slicing With Deep Reinforcement Learning for URLLC and eMBB Services | 7 / 24 |

*Continued on next page*

*(continued)*

| Paper | Wrong / Total |
|---|---|
| An Empirical Study of Recent Face Alignment Methods | 6 / 34 |
| Towards Massively Parallel Computations in Algebraic Geometry | 6 / 24 |
| Group Sparse Priors for Covariance Estimation | 6 / 34 |
| Efficient Non-greedy Optimization of Decision Trees | 5 / 26 |
| Fuzzy Jaccard Index: A robust comparison of ordered lists | 5 / 22 |
| Recurrent Neural Networks with Top-k Gains for Session-based Recommendations | 5 / 20 |
| Adaptive high-order splitting schemes for large-scale differential Riccati equations | 5 / 18 |
| Reflection Backdoor: A Natural Backdoor Attack on Deep Neural Networks | 4 / 6 |
| Adversarial Machine Learning for 5G Communications Security | 4 / 18 |

A strong consistency is observed across algorithms. Two papers appear in all configurations:

- An Empirical Study of Recent Face Alignment Methods
- Dynamic SDN-Based Radio Access Network Slicing With Deep Reinforcement Learning for URLLC and eMBB Services

Several papers appear in five out of six settings:

- Towards Massively Parallel Computations in Algebraic Geometry
- Recurrent Neural Networks with Top-k Gains for Session-based Recommendations
- Efficient Non-greedy Optimization of Decision Trees
- Adversarial Machine Learning for 5G Communications Security

The same papers recur across algorithms because they are the most fragile items in the dataset, so different training setups hit the same weak points. This looks more like question-design sensitivity than simple topic difficulty: broad topic recognition survives, but exact relation checking and exact phrase recall do not.

### N.5. Analysis of Cross-Forget-Set Spillover in Unlearning

For a subset of papers, unlearning with one forget set induces measurable forgetting in the *other* forget set as well—yet this spillover does not occur universally. Understanding why some papers are susceptible to this cross-forget-set effect while others are not is the central question explored in subsection.

Table 14 reports, for each unlearning algorithm and forget-set configuration, how many papers exhibit errors (i.e. measurable forgetting) in **both** forget sets simultaneously, in **only** the first forget set, and in **only** the second forget set.

*Table 14.* Per-paper error distribution across the two forget sets for each unlearning algorithm and forget-set configuration. *Both* denotes papers showing forgetting in both sets; *Only F1* and *Only F2* denote papers showing forgetting exclusively in Forget Set 1 or Forget Set 2, respectively.

| Algorithm / Config | Both Sets | Only F1 | Only F2 | Total Affected |
|---|---|---|---|---|
| GD F1 | 39 | 30 | 1 | 70 |
| GD F2 | 30 | 5 | 25 | 60 |
| NPO RT F1 | 29 | 20 | 7 | 56 |
| NPO RT F2 | 28 | 13 | 23 | 64 |
| SimNPO RT F1 | 32 | 31 | 3 | 66 |
| SimNPO RT F2 | 29 | 9 | 15 | 53 |

The pattern is consistent across all six algorithm–configuration combinations. This consistency suggests the susceptibility to spillover is a property of the *papers themselves* rather than an artefact of any particular unlearning algorithm.

### N.5.1. Proxy Measure: Citation-Based Embeddedness

The natural hypothesis is that papers which are forgotten by *both* forget sets are less deeply embedded in the model's pretraining corpus—their knowledge representations are less entangled with a narrow neighbourhood of related content. Directly measuring how deeply a paper is rooted in a pretrained model's weights is intractable, however, as it would require exhaustive probing across the full parameter space.

We therefore adopt two **proxy measures** derived from the Semantic Scholar citation graph.

1. **Normal citation count.** The number of papers that directly cite the paper under consideration. A highly-cited paper is more likely to have its findings reproduced, discussed, and cross-referenced in a large volume of text that would itself be present in a large pretraining corpus. Consequently, the paper's content is represented not just once but implicitly through thousands of derivative works.

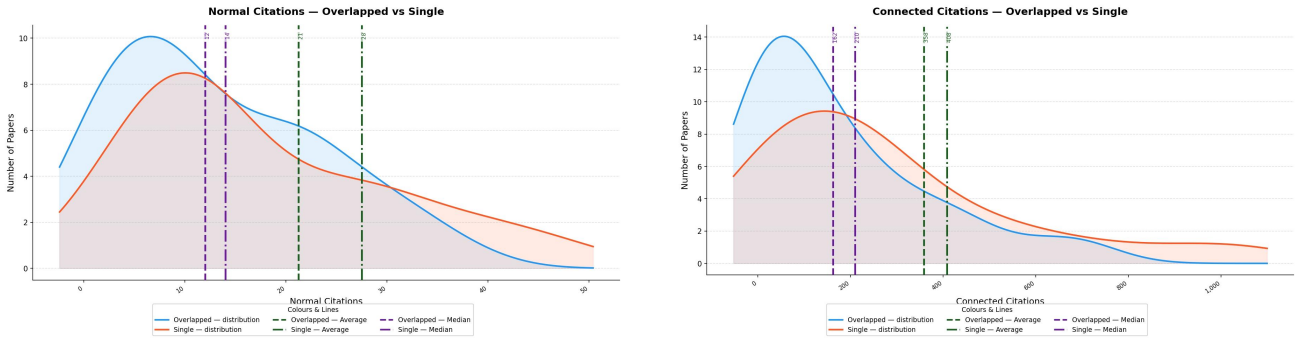
2. **Connected citation count.** Defined as the paper's own citation count *plus* the sum of citation counts of every paper that cites it:

$$C_{\text{connected}}(p) = C(p) + \sum_{q \in \text{citers}(p)} C(q). \tag{8}$$

   This second-order measure captures how influential the *citers* of a paper are. A paper cited by other heavily-cited papers propagates its conceptual footprint far more widely through the literature and thus through any corpus derived from it—than a paper cited only by obscure works.

Both measures serve as proxies for *corpus embeddedness*: the more a paper's ideas permeate the broader literature, the more redundant and distributed its representation is likely to be in the model's weights, and the harder it becomes to confine unlearning to a single, targeted forget set.

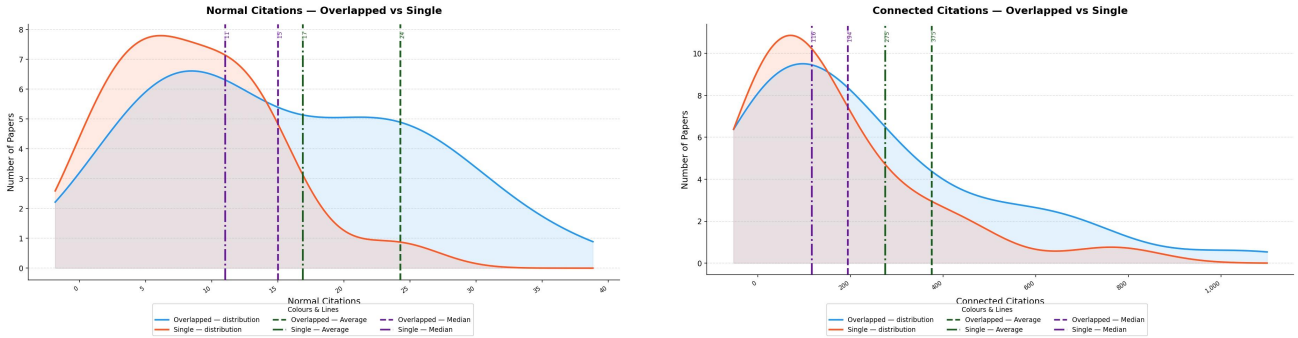
To examine whether citation counts differ systematically between the two groups, we plot, for each algorithm and forget-set configuration, the distribution of both normal and connected citation counts for papers affected *only* by the single active forget set (labelled Single) against papers affected by *both* forget sets (labelled Overlapped). Each distribution is rendered as a KDE-smoothed curve (outliers removed via the IQR fence method with $k = 1.5$ prior to plotting), with vertical lines indicating the group mean (dark green) and median (dark purple). Separate plots are produced for normal citation counts and connected citation counts for each of the six algorithm–configuration pairs, yielding twelve figures in total from Figure 17a to Figure 22b.




*(a)* KDE Plot of normal citations

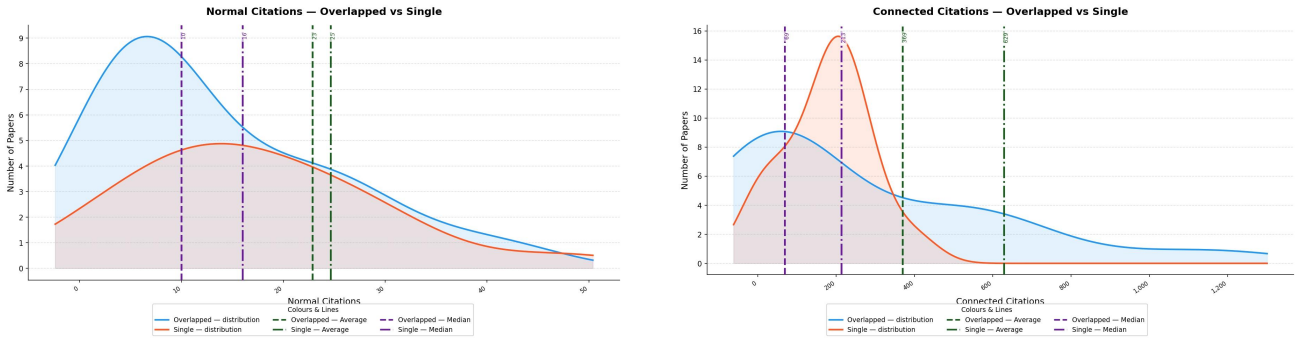
*(b)* KDE Plot of connected citations

*Figure 17.* KDE plot for GD algorithm when unlearned with Forget Set 1.



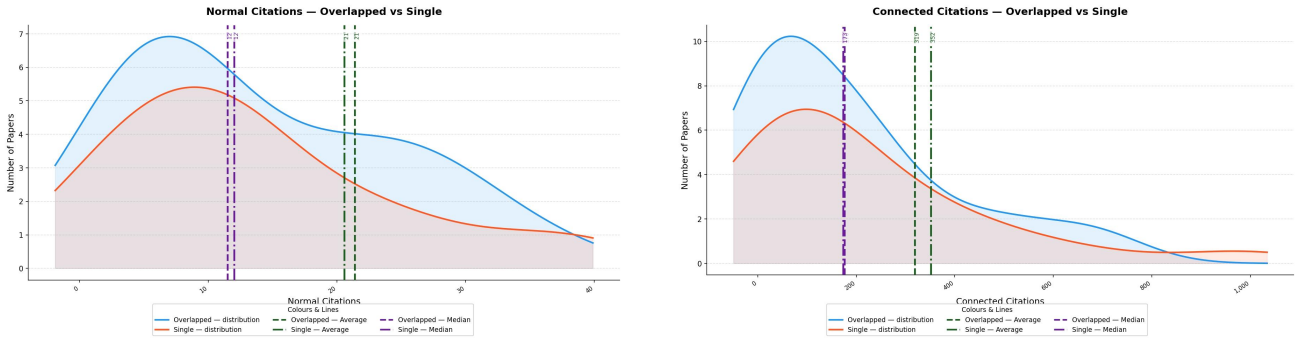

*(a)* KDE Plot of normal citations *(b)* KDE Plot of connected citations

*Figure 18.* KDE plot for GD algorithm when unlearned with Forget Set 2.



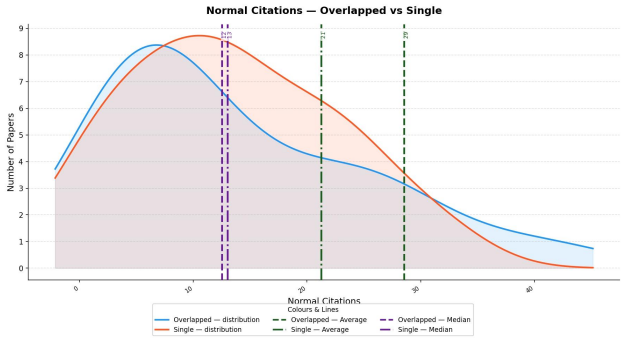

*(a)* KDE Plot of normal citations *(b)* KDE Plot of connected citations

*Figure 19.* KDE plot for NPO+RT algorithm when unlearned with Forget Set 1.

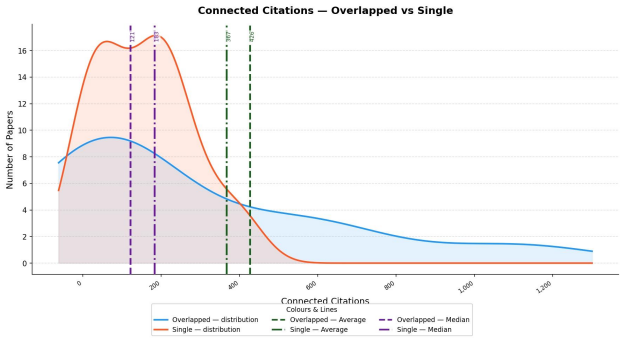



*(a)* KDE Plot of normal citations *(b)* (KDE Plot of connected citations

*Figure 20.* KDE plot for NPO+RT algorithm when unlearned with Forget Set 2.

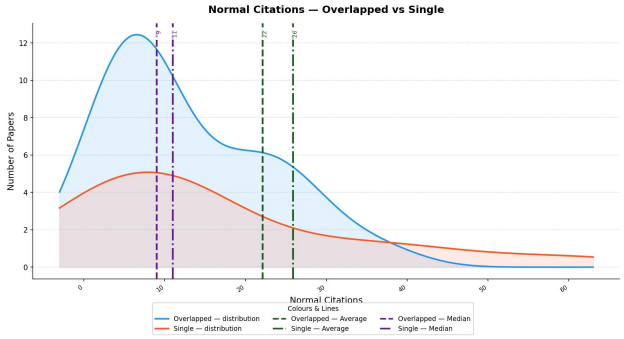



*(a)* KDE Plot of normal citations

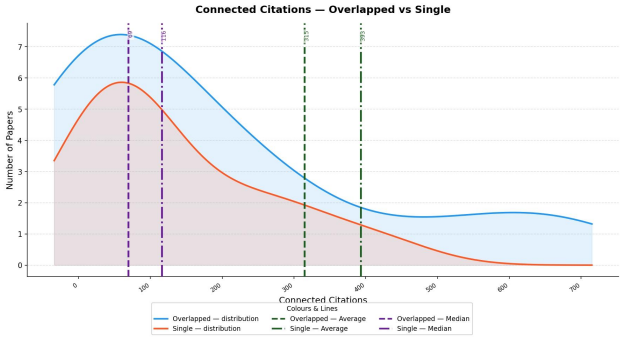



*(b)* KDE Plot of connected citations

*Figure 21.* KDE plot for SimNPO+RT algorithm when unlearned with Forget Set 1.




*(a)* KDE Plot of normal citations




*(b)* KDE Plot of connected citations

*Figure 22.* KDE plot for SimNPO+RT algorithm when unlearned with Forget Set 2.

Across the plots, a consistent trend emerges: **papers that are forgotten by only the forget set used during unlearning tend to exhibit higher citation counts**—both normal and connected—than papers that are forgotten by both the forget set. This holds across most of the algorithm - configuration pair where the median citation value of the non overlapped paper set is to the right of the overlapped paper set (Figure 17a, 17b, 19a, 19b, 20a, 21a, 21b, 22a, 22b).

The implication is consistent with our proxy hypothesis: papers that are more heavily cited—and whose citers are themselves heavily cited—leave a broader imprint on pretraining corpora. Their knowledge is represented not only directly but also indirectly through a large web of derivative literature. When unlearning is applied to one forget set, the gradient signal is strong enough to erode representations that overlap with the other forget set, because those representations are built from shared, widely-distributed textual evidence. By contrast, papers with low citation counts are more narrowly represented, and unlearning one forget set does not provide sufficient signal to disturb the model's encoding of the other.

This finding suggests that highly-cited papers may require targeted, multi-set unlearning strategies to achieve precise and contained forgetting.