Unlearning as Ablation: Toward a Falsifiable Benchmark for Generative Scientific Discovery

Robert Yang
S6 Research
San Jose, CA 95129
robert@s6research.org

Abstract

Bold claims about AI's role in science—from "AGI will cure all diseases" to promises of radically accelerated discovery—raise a central epistemic question: do large language models (LLMs) truly *generate* new knowledge, or do they merely remix memorized fragments? We propose **unlearning-as-ablation** as a falsifiable probe of constructive scientific discovery. The idea is to systematically removes a target result together with its *forget-closure* (supporting lemmas, paraphrases, and multi-hop entailments) and then evaluate whether the model can re-derive the result from only permitted axioms and tools. Success would indicate generative capability beyond recall; failure would expose current limits. Unlike prevailing motivations for unlearning—privacy, copyright, or safety—our framing repositions it as an *epistemic probe* for AI-for-Science. We outline a minimal pilot in mathematics and algorithms to illustrate feasibility, and sketch how the same approach could later be extended to domains such as physics or chemistry. This is a position paper: our contribution is conceptual and methodological, not empirical. We aim to stimulate discussion on how principled ablation tests could help distinguish models that reconstruct knowledge from those that merely retrieve it, and how such probes might guide the next generation of AI-for-Science benchmarks.

1 Introduction

Recent breakthroughs in foundation models have fueled bold claims—from predictions that "AGI will cure all diseases" to assertions that scientific progress will soon accelerate far beyond historical rates. These visions reflect real excitement, but they obscure a fundamental epistemic question: **do large language models (LLMs) genuinely generate new knowledge, or do they merely remix what was already present in their training data?**

This distinction matters deeply for AI-for-Science. Without a falsifiable test of constructive knowledge generation, claims of "discovery" remain philosophically ambiguous and scientifically ungrounded. If AI systems are to be trusted as collaborators in science, we must know whether they can *derive* new results from principles, rather than retrieve or interpolate memorized fragments.

We propose a new perspective: **unlearning-as-ablation**. The idea is straightforward. Select a scientific result T (e.g., a theorem or algorithm), identify its entire *forget-closure* $\mathcal{F}(T)$—all lemmas, paraphrases, aliases, and multi-hop entailments that lead to T—and perform strong unlearning over $\mathcal{F}(T)$. Afterward, provide the model only with permitted axioms and tools, and test whether it can re-derive T in a verifiable form. Success constitutes positive evidence of constructive generation, whereas failure or leakage exposes the boundaries of current capabilities.

This framing departs from prevailing motivations for unlearning. Surveys emphasize privacy, copyright, and safety as primary rationales [Xu et al., 2023, 2024], with evaluation focused on removal fidelity rather than generative ability. Recent work highlights the difficulty of faithfully removing

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: .

multi-hop or entangled knowledge [Choi et al., 2024, Wang et al., 2025, Shah et al., 2025], while other studies show that forgotten content can often be "relearned" through small finetunes or prompting [Shah et al., 2025]. In the safety and compliance setting, these phenomena are treated as risks. In our setting, they define the frontier: as unlearning methods improve in addressing leakage and robustness, the resulting ablations become more faithful, and the corresponding rediscovery benchmarks more stringent. In this way, progress in unlearning directly strengthens our ability to test whether models are capable of constructive scientific generation.

By reframing unlearning as an *experimental probe*, we aim to bridge AI-for-Science and safety communities. The result is a concrete, falsifiable methodology for testing the limits of LLMs: whether they are capable of genuine discovery, or whether their advances remain bounded by retrieval and interpolation. As a position paper, our contribution is primarily conceptual: we propose a methodological framework and outline pilot domains, leaving systematic empirical validation to future work.

2 Background: Unlearning Today

The study of unlearning in machine learning and large language models (LLMs) has grown rapidly in recent years, motivated largely by *external constraints* such as law, safety, or ethics rather than by epistemic goals. We briefly review the dominant rationales, common methodologies, and key evaluation challenges.

2.1 Motivations for Unlearning

Three primary motivations recur across surveys and frameworks:

(1) Privacy and compliance. Regulations such as the General Data Protection Regulation (GDPR) enshrine a "right to be forgotten," requiring that models support the removal of sensitive or personally identifiable data. Surveys on digital forgetting in LLMs emphasize compliance with privacy law as a central driver of research in this area [Xu et al., 2024].

(2) Copyright and intellectual property. LLMs trained on large web scrapes may inadvertently memorize copyrighted text, code, or images. Several works argue that machine unlearning is necessary to respect intellectual property claims and to support takedown requests from rights-holders [Karamolegkou et al., 2023, Dou et al., 2025, Xu et al., 2024, Ren et al., 2025, Yao et al., 2024].

(3) Safety and dual-use knowledge. A third line of work focuses on removing *hazardous* content: for example, step-by-step instructions for synthesizing explosives or pathogens. Recent benchmarks such as WMDP [Li et al., 2024] evaluate whether unlearning can reduce dual-use risks while maintaining general utility.

2.2 Methodological Approaches

Most unlearning methods adapt techniques from model editing or fine-tuning. Examples include:

- **Gradient-ascent or anti-training:** adjusting model parameters to maximize loss on target examples, thereby forgetting them.
- **Representation-level interventions:** e.g., Amnesic Probing [Elazar et al., 2021] removes specific linguistic features from hidden states.
- **Retrieval suppression:** steering methods that block particular outputs without removing underlying representations.

While diverse, these approaches generally aim at *removal fidelity*: ensuring that specific facts or behaviors no longer appear in model outputs.

2.3 Evaluation Challenges

Evaluation is a persistent bottleneck. Several recent studies emphasize that:

- **Entangled knowledge is difficult to erase.** Multi-hop unlearning benchmarks show that even if intermediate nodes are removed, models can often reconstruct targets via alternative reasoning chains [Choi et al., 2024, Wang et al., 2025, Shah et al., 2025].
- **Suppression vs. removal.** SoK papers stress the importance of distinguishing true parameter-level removal from surface-level suppression, where models appear to forget but can be prompted to recall [Ren et al., 2025].
- **Relearning and robustness.** Empirical work demonstrates that forgotten content can often be "jogged" back into use with minimal finetuning or prompting [Lee et al., 2025].

2.4 Gap for AI-for-Science

Notably, none of the above rationales frame unlearning as a tool for *scientific epistemology*. Unlearning has been motivated by compliance and safety, not by the question of whether a model can *reconstruct* forgotten knowledge from first principles. This gap opens an opportunity: by treating unlearning as *ablation*, we can design falsifiable experiments to probe whether LLMs possess constructive generative capabilities, a perspective particularly urgent for AI-for-Science. Moreover, the progress of unlearning research directly determines the strength of such benchmarks: the more thorough and faithful the unlearning, the harder the rediscovery task becomes, and the more reliable the test of whether models can generate knowledge rather than recall it.

3 Proposal: Unlearning-as-Ablation

We propose to repurpose unlearning from its conventional role in privacy or safety into an *experimental ablation method* for probing constructive knowledge generation. The central idea is to remove not only a target result T, but also all of the *supporting knowledge that directly enables it*, and then ask the model to re-derive T from only axioms and tools that remain accessible. If the model succeeds under these conditions, we gain falsifiable evidence that it is not merely retrieving memorized fragments but genuinely generating knowledge.

3.1 Defining the Forget-Closure

The first step is to formally define the **forget-closure** $\mathcal{F}(T)$ of a target T. This closure includes:

- All direct statements of T (canonical forms, proofs, code).
- Paraphrases and rephrasings that preserve semantic equivalence.
- Intermediate lemmas or building blocks that entail T.
- Multi-hop reasoning chains where T can be reconstructed indirectly [Choi et al., 2024, Wang et al., 2025, Shah et al., 2025].
- Same-answer sets where multiple formulations yield equivalent outputs.

By removing the entire $\mathcal{F}(T)$, we close off not only surface forms but also indirect reasoning paths that would otherwise allow reconstruction through entanglement.

3.2 Performing Strong Unlearning

The second step is to apply **removal-oriented unlearning** across $\mathcal{F}(T)$. Unlike suppression methods that steer generation away from target outputs, removal aims to eliminate relevant information from the parameterization itself. Candidate techniques include gradient-ascent unlearning, targeted finetuning, or optimization-based methods evaluated in recent surveys [Ren et al., 2025]. To confirm removal, we propose adopting multi-faceted audits:

- Leakage checks on paraphrase, multi-hop, and same-answer sets.
- Counterfactual activation probes (inspired by Amnesic Probing) to test whether T-related features still reside in hidden states [Elazar et al., 2021].
- Robustness tests against "jogging" attacks, where small finetunes or prompting can restore forgotten knowledge [Lee et al., 2025].

These checks ensure that the unlearning process produces a genuine epistemic blank slate with respect to $\mathcal{F}(T)$.

3.3 Re-Derivation as a Falsifiable Test

Finally, we design a **re-derivation trial**. After unlearning, the model is provided with:

1. A set of axioms, primitives, or base tools that are *not* part of $\mathcal{F}(T)$.
2. A prompt or environment that permits constructive reasoning (e.g., a proof assistant or a test-driven code synthesis framework).

The task is to derive T in a form that can be verified by an external oracle: for example, a formal proof accepted by Lean or Isabelle, or a program passing a hidden test suite. Importantly, success is only counted if T is re-derived *without leakage from* $\mathcal{F}(T)$.

This yields a falsifiable criterion: if the model can re-derive T despite rigorous unlearning of all prerequisite paths, we have positive evidence for constructive generation. If it cannot, or if leakage audits reveal dependence on residual memory, then the claim of "scientific discovery" remains unsubstantiated.

3.4 Why This Matters

This approach connects progress in unlearning directly to progress in measuring scientific discovery. In the safety and compliance literature, challenges such as entanglement, multi-hop reasoning, and relearning are treated as failure modes because they undermine removal fidelity [Dou et al., 2025, Choi et al., 2024, Wang et al., 2025, Shah et al., 2025]. In our framing, they set the difficulty of the benchmark: the more effectively unlearning methods address these challenges, the more thoroughly the target knowledge is ablated, and the more demanding the rediscovery task becomes. Thus, advances in unlearning translate into sharper tests of whether LLMs truly possess constructive generative capability. Rather than turning flaws into benefits, we highlight that solving these long-standing problems in unlearning is what enables rigorous epistemic evaluation in AI-for-Science.

4 Minimal Pilot Study

While the long-term vision is to apply unlearning-as-ablation to scientific hypotheses in physics, chemistry, or biology, we propose beginning with domains where **verification is automatic and unambiguous**. This allows us to isolate the epistemic question—can a model *re-derive* knowledge once its closure has been forgotten?—without relying on subjective human judgment.

4.1 Mathematics: Formal Proofs

Mathematics provides an ideal testbed because results can be verified by proof assistants such as *Lean* or *Isabelle*. A minimal pilot could proceed as follows:

1. Select a mid-tier theorem (e.g., in number theory or combinatorics) that has a clear dependency structure.
2. Construct its forget-closure $\mathcal{F}(T)$, including canonical statements, paraphrased variants, and prerequisite lemmas.
3. Apply strong unlearning over $\mathcal{F}(T)$.
4. Task the model with re-proving T using only base axioms and allowed rules of inference.

Success is defined as producing a proof accepted by the proof assistant. Failure or leakage (e.g., shortcut recall of a forgotten lemma) falsifies the claim of rediscovery.

4.2 Algorithms: Verified Implementations

Algorithms provide another tractable domain, where correctness can be checked against hidden test suites. For example:

1. Forget the Knuth–Morris–Pratt (KMP) string matching algorithm, along with all prerequisite explanations, code templates, and paraphrases.
2. After unlearning, ask the model to derive an efficient string-matching procedure from first principles (e.g., reasoning about prefix functions).
3. Validate correctness using adversarial test cases and runtime complexity checks.

As in mathematics, the evaluation is binary: either the model reconstructs a working implementation, or it does not.

4.3 Evaluation Metrics

To assess the outcome of such pilots, we propose three classes of metrics:

- **Success rate.** Fraction of trials where the model re-derives T in a verifiable form (proof acceptance, program passes test suite).
- **Leakage audits.** Performance on paraphrase, multi-hop, and same-answer sets drawn from $\mathcal{F}(T)$, ensuring the model is not recalling forgotten material [Choi et al., 2024, Wang et al., 2025, Shah et al., 2025].
- **Utility retention.** Accuracy on unrelated benchmarks (e.g., a subset of MMLU) to confirm that unlearning did not degrade general capability [Ren et al., 2025, Yao et al., 2024].

4.4 Why a Minimal Pilot is Valuable

Even small-scale pilots can decisively answer whether LLMs exhibit generative capability under ablation. If a model successfully re-derives a theorem or algorithm after strong unlearning of its closure, we obtain falsifiable evidence that it constructs knowledge rather than merely retrieving it. Conversely, if models fail under such controlled conditions, this highlights a concrete epistemic limit of current systems. Either outcome offers high-value insight for AI-for-Science, where claims of accelerated discovery remain both enticing and contested.

5 Implications for AI-for-Science

The proposed unlearning-as-ablation framework has direct consequences for how we understand the promise and limits of AI-for-Science.

5.1 Epistemic Clarity in Scientific Discovery

The central value of this approach is that it provides a *falsifiable test* of discovery. Today, when an LLM proposes a hypothesis, proves a theorem, or writes an algorithm, it remains unclear whether this is a product of genuine reasoning or of subtle retrieval from training data. By first *removing* all accessible pathways to a result and then testing for *re-derivation*, we create a clean epistemic separation: success implies constructive generation, while failure implies dependence on stored fragments. This reframing allows the AI-for-Science community to move beyond speculation about "discovery" and instead ground claims in falsifiable evidence.

5.2 Turning Failure Modes into Probes

Unlearning research has traditionally cast entanglement, multi-hop reasoning, and relearning as obstacles [Choi et al., 2024, Wang et al., 2025, Shah et al., 2025]. In our setting, these challenges become useful stress tests. If a model cannot succeed once closure paths are blocked, it indicates that the relevant knowledge was never truly generative. If it can succeed, it demonstrates robustness and constructive capacity. Either way, phenomena previously treated as evaluation headaches become diagnostic instruments for probing the depth of model reasoning.

5.3 Broader AI-for-Science Roadmap

Although we highlight mathematics and algorithms as tractable pilot domains, the methodology generalizes. In physics, one could remove an established equation and test whether the model can

re-derive it from fundamental laws. In chemistry, one could unlearn a well-known synthesis route and test whether the model can rediscover it from reaction rules. In biology, one could unlearn a canonical protein interaction and test for re-derivation from structural principles. These extensions would demand careful closure construction and domain-specific verification, but they illustrate how the same ablation logic scales to real scientific practice.

5.4 Redefining the Boundary of AI Progress

Finally, this framework speaks directly to the theme of this workshop: the reach and limits of AI in scientific discovery. If unlearning-as-ablation pilots reveal that models can re-derive knowledge under strong ablation, this strengthens the case that AI can generate truly novel insights. If they reveal consistent failures, it delineates a boundary condition: LLMs may accelerate retrieval, interpolation, and synthesis, but fall short of independent knowledge generation. In both outcomes, the methodology provides a principled way to map the contours of what AI can and cannot do for science.

5.5 Toward the Next Major Benchmark

A final implication is that unlearning-as-ablation offers a clear path toward the next generation of benchmarks for AI progress. Just as ImageNet catalyzed advances in computer vision by providing a well-defined task on which algorithms could be compared [Deng et al., 2009], a benchmark grounded in constructive re-derivation after unlearning could serve as a lodestar for AI-for-Science. Existing evaluations of knowledge regurgitation and short-form reasoning are increasingly saturated—as highlighted by works such as Humanity's Last Exam [Phan et al., 2025]—suggesting that the next frontier must measure whether models can move beyond retrieval and interpolation to genuine discovery. We believe that such an "unlearning-as-ablation" benchmark could become a distinguishing test of model strength, separating systems that can merely recall from those that can constructively generate new scientific knowledge.

Importantly, the strength of such a benchmark is coupled to the progress of unlearning research itself. As unlearning methods become more faithful and thorough, the corresponding benchmarks become more stringent: rediscovery requires deeper reasoning, and successful re-derivation provides stronger evidence of constructive capability. In this way, advances in unlearning directly drive advances in our ability to measure—and eventually to achieve—genuine AI scientific discovery.

6 Conclusion

We have proposed *unlearning-as-ablation* as a new lens on large language models, reframing unlearning from a tool of compliance and safety into a falsifiable probe of scientific discovery. By systematically removing a target result and its forget-closure, and then testing whether the model can re-derive the result from permitted axioms and tools, we obtain an experimental method to separate retrieval from constructive generation. This approach directly addresses one of the most pressing open questions in AI-for-Science: can AI systems truly generate new knowledge? Even minimal pilots in mathematics or algorithms provide decisive evidence either way, while extensions to physics, chemistry, and biology can delineate the boundaries of future AI scientific progress. Whether the outcome is success or failure, unlearning-as-ablation offers the community a principled framework to move beyond speculation and anchor claims of discovery in falsifiable tests.

7 Call to Action

We are seeking collaborators to help turn this conceptual framework into a practical evaluation pipeline. We aim to release an initial benchmark prototype in the coming months and ultimately work towards a full benchmark paper. Contributions from researchers in unlearning, model editing, verification, and evaluation design are warmly welcomed.

8 Acknowledgments

Special thanks to my parents for the fruitful discussions in December 2023 on LLMs for scientific discovery and agentic systems evaluation (then not known by this term). The insightful questions posed were indispensible in the ideation of this method.

9 Transparency on AI usage

Although not required for NeurIPS submission, for full transparency of the paper, we include this disclosure here. AI was used extensively throughout the paper for editing (suggesting terminology use and phrasing, plus paraphrasing of the author's writing to increase quality) fully adhering to NeurIPS 2025 guidelines. The authors are responsible for the entire content of the paper, including all text, figures, and references.

References

Common crawl. `https://commoncrawl.org/`. A free, openly accessible web crawl corpus.

Minseok Choi, ChaeHun Park, Dohyun Lee, and Jaegul Choo. Breaking chains: Unraveling the links in multi-hop knowledge unlearning. 10 2024. doi: 10.48550/arXiv.2410.13274.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems (gsm8k dataset). `https://arxiv.org/abs/2110.14168`, 2021. arXiv:2110.14168.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Guangyao Dou, Zheyuan Liu, Qing Lyu, Kaize Ding, and Eric Wong. Avoiding copyright infringement via large language model unlearning. pages 5176–5200, 01 2025. doi: 10.18653/v1/2025.findings-naacl.288.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 03 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00359. URL `https://doi.org/10.1162/tacl_a_00359`.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, et al. Towards an ai co-scientist. *arXiv preprint arXiv:2502.18864*, 2025.

Dan Hendrycks et al. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Jiaxin Huang, Yizhou Wang, Xiang Lin, Yue Zhang, Chen Liang, Yixuan Li, and Pin-Yu Chen. Understanding and evaluating unlearning in large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 20386–20404. PMLR, 2024. URL `https://proceedings.mlr.press/v235/huang24u.html`.

Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. Copyright violations and large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7403–7412, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.458. URL `https://aclanthology.org/2023.emnlp-main.458/`.

Bruce W. Lee, Addie Foote, Alex Infanger, Leni Shor, Harish Kamath, Jacob Goldman-Wetzler, Bryce Woodworth, Alex Cloud, and Alexander Matt Turner. Distillation robustifies unlearning, 2025. URL `https://arxiv.org/abs/2506.06278`.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew Bo Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Ariel Herbert-Voss, Cort B Breuer, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam Alfred Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Ian Steneker, David Campbell, Brad Jokubaitis, Steven Basart, Stephen Fitz, Ponnurangam Kumaraguru, Kallol Krishna Karmakar, Uday Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 28525–28550. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/li24bc.html.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer (memit). https://arxiv.org/abs/2210.07229, 2022. arXiv:2210.07229.

Long Phan, Andrew Gatti, Zihao Han, et al. Humanity's last exam. 2025. URL https://arxiv.org/abs/2501.14249.

Jie Ren, Yue Xing, Yingqian Cui, Charu C Aggarwal, and Hui Lui. Sok: Machine unlearning for large language models. 06 2025. doi: arXiv.2506.09227.

Raj Sanjay Shah, Jing Huang, Keerthiram Murugesan, Nathalie Baracaldo, and Diyi Yang. The unlearning mirage: A dynamic framework for evaluating LLM unlearning. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=exW2SFJK4H.

Kenneth O. Stanley and Joel Lehman. *Why greatness cannot be planned: The myth of the objective*. MIT Press, 2015.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Tal Linzen, Grzegorz Chrupała, and Afra Alishahi, editors, *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL https://aclanthology.org/W18-5446/.

Changsheng Wang, Chongyu Fan, Yihua Zhang, Jinghan Jia, Dennis Wei, Parikshit Ram, Nathalie Baracaldo, and Sijia Liu. Reasoning model unlearning: Forgetting traces, not just answers, while preserving reasoning skills. 06 2025. doi: arXiv.2506.12963.

Heng Xu, Tianqing Zhu, Lefeng Zhang, Wanlei Zhou, and Philip S. Yu. Machine unlearning: A survey. *ACM Comput. Surv.*, 56(1), August 2023. ISSN 0360-0300. doi: 10.1145/3603620. URL https://doi.org/10.1145/3603620.

Jie Xu, Zihan Wu, Cong Wang, and Xiaohua Jia. Machine unlearning: Solutions and challenges. *IEEE Transactions on Emerging Topics in Computational Intelligence*, PP:1–19, 06 2024. doi: 10.1109/TETCI.2024.3379240.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, et al. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025.

Yuanshun Yao, Xiaojun Xu, and YangLiu. Large language model unlearning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 105425–105475. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/be52acf6bccf4a8c0a90fe2f5cfcead3-Paper-Conference.pdf.

Kankan Zhou, Eason Lai, Wei Bin Au Yeong, Kyriakos Mouratidis, and Jing Jiang. Rome: Evaluating pre-trained vision-language models on reasoning beyond visual common sense. https://arxiv.org/abs/2310.19301, 2023. arXiv:2310.19301.

A Workshop Discussion: Alternative Views and Benchmark Versioning

This appendix distills key objections and design refinements raised during workshop discussion. Two themes recurred: (i) rediscovery must be operationalized with *external verification* (proof assistants, hidden test suites), and (ii) the community benefits from a *static, comparable benchmark* whose validity does not depend on a moving "past cutoff date" heuristic, as is common in ad hoc evaluations of AI-scientist systems. Unlearning-as-ablation is intended to provide such a fixed substrate: models and frameworks can be compared across years because tasks, ablations, and verifiers remain stable, rather than shifting as the calendar advances.

A.1 Alternative views

(a) Tradeoff: strict unlearning vs. retained capability. A natural concern is that stronger unlearning of $\mathcal{F}(T)$ may induce greater collateral damage, degrading general capabilities and confounding rediscovery outcomes. This motivates treating ablation quality as *part of the benchmark report*, not a hidden implementation detail.

(b) "Unlearning is not yet mature enough." A related concern is that current unlearning methods may not reliably remove entangled multi-hop knowledge without either (i) leaving residual leakage or (ii) degrading unrelated competence. In this view, the benchmark risks measuring unlearning artifacts rather than constructive generation.

Benchmark versioning as a response to (a)–(b). To address both concerns, workshop discussion suggested **versioning the benchmark by measured ablation quality**. Concretely, the benchmark reports three metrics:

- **Rediscovery success rate** (the hill-climb target for models/systems): verifier pass@k on T (e.g., proof acceptance or hidden test-suite pass).
- **Leakage audits** (benchmark introspection): performance on probes drawn from $\mathcal{F}(T)$ (paraphrase / multi-hop / same-answer), where *lower* indicates stronger forgetting.
- **Utility retention** (benchmark introspection): performance on an unrelated capability suite, where *higher* indicates less collateral degradation.

The key proposal is that **leakage** and **utility** should be surfaced as first-class identifiers of a benchmark release. For example, one could label a release as `A2D-L40-U80` to indicate empirically measured leakage score 40 and utility retention 80 (illustrative numbers). As unlearning improves, new releases would naturally shift toward lower leakage and higher utility (e.g., `A2D-L20-U90`), making progress in unlearning transparently reflected in the benchmark itself.

Optionally, the discussion suggested a third axis to define **difficulty tiers within the same time period**: a **closure-depth** or **hop budget** indicator (e.g., the maximum multi-hop depth included in $\mathcal{F}(T)$). This enables benchmarks that differ in logical entanglement even under the same unlearning technology (e.g., shallow vs. deeper closure ablations), while remaining comparable through explicit reporting.

Note on expected maturity. Workshop discussion also raised the possibility that "good-enough" unlearning may arrive soon enough to make this benchmark practical. As an example of rapidly evolving techniques in this area, methods released immediately after the workshop (e.g., approaches aimed at shaping training-time behavior) suggest directions that could plausibly transfer to post-hoc unlearning, albeit currently at high cost. This motivates designing the benchmark so it can *start* with imperfect ablations (explicitly labeled by leakage/utility) and become more stringent as unlearning advances.

(c) Discipline-specific inductive biases. A final objection is that "rediscovery" is not uniform across domains: mathematics, algorithms, physics, chemistry, and biology differ in what constitutes a legitimate derivation, what tools are permissible, and what inductive biases are required. Workshop discussion therefore favored **a family of benchmarks**, one per discipline, each with domain-appropriate: (i) target types T, (ii) closure specifications $\mathcal{F}(T)$, and (iii) verification oracles $\mathcal{V}(T)$. This preserves the core epistemic logic of ablation-to-rediscovery while respecting that the notion of "first principles" and verification differs across fields.

B Benchmark Specification

We propose Ablation-to-Discovery (A2D), a benchmark that frames unlearning as ablation to test whether large language models can reconstruct systematically removed knowledge from first principles. By defining rediscovery tasks with verifiable outcomes, A2D probes constructive generative ability—can models re-derive what was excised? This offers a falsifiable substrate for evaluating knowledge generation beyond memorization. Just as ImageNet galvanized computer vision, we envision A2D as the "ImageNet of knowledge generation"—a shared testbed for measuring and accelerating AI-for-Science progress.

B.1 Dataset Rationale — Why an Ablation-Coupled Benchmark?

LLMs are saturated on recall-heavy tasks but under-tested on constructive generation. A2D provides a controlled falsifiable test: remove structured knowledge T, then evaluate if models can rebuild it without rote recall.

Unlearning is typically framed as a risk mitigation strategy (safety, privacy) [Huang et al., 2024, Xu et al., 2024, Ren et al., 2025]. Here, we reframe it as a methodological opportunity: each advance in unlearning methods strengthens ablations, raising the difficulty of reconstructive discovery. Thus, progress in unlearning directly drives progress in A2D benchmarks.

These are our key rationales:

- Scientific falsifiability: A2D enables yes/no tests of generative capability.
- Reproducibility: Each task ships as a containerized config.
- Benchmark trajectory: Initial domains (math, algorithms), then expansion into physics, chemistry, biology, and other basic sciences.
- Community role: A2D can serve as a battleground for AI Scientist frameworks, providing the first quantitative ground for comparing systems like Google's biotech discovery AI [Gottweis et al., 2025] or Sakana's automated CS paper generation [Lu et al., 2024, Yamada et al., 2025].

B.2 AI Task Definition

Traditional benchmarks equate "dataset" with a static collection of labeled examples—ImageNet, GLUE, and many others embody this paradigm [Deng et al., 2009, Wang et al., 2018]. Our proposal expands this definition. In the unlearning-as-ablation setting, the benchmark is not merely the data but the procedure by which knowledge is systematically removed. In some cases, it also encompasses a standardized reference model that undergoes ablation. In this view, a dataset is no longer just an archive of examples, but a dynamic specification of data, process, and model.

This redefinition is essential. By treating ablation as part of the benchmark, we can directly test whether models or systems can reconstruct algorithmic rules, scientific knowledge, or cross-domain mappings that have been deliberately removed. Without embedding the ablation protocol (and in some cases, the model artifact) into the benchmark, such generative reconstruction cannot be meaningfully evaluated. Thus, our task definition is broader than "predict labels for examples." It is: given a systematically ablated knowledge space, recover the missing structure with scientific fidelity.

B.3 Tracks and Modes

Our proposal separates evaluation along two orthogonal dimensions: tracks and modes.

(1) Tracks (what is reconstructed):

- Algorithmic Re-derivation – rediscovering hidden formal rules or procedures.
- Scientific Knowledge Reconstruction – restoring ablated domain-specific knowledge (e.g., molecular pathways, physics laws) by reasoning from foundational laws and experimental constraints?
- Cross-Domain Generalization – leveraging one domain to recover knowledge in another. For example, can a model, after relevant results are removed, re-derive a computational biology method by combining algorithmic and biochemical principles?

(2) Evaluation Modes (how the test is administered):

- BYOM Capability Test – the benchmark specifies only the ablation protocol, with models tested directly (no discovery frameworks), to isolate capability.
- System Capability Test – the benchmark includes both the ablation protocol and a standardized reference model artifact. Here, the comparison is among discovery frameworks, evaluating how orchestration, augmentation, or agentic processes recover knowledge.

By separating Task Tracks from Evaluation Modes, the benchmark distinguishes between what kind of knowledge generation is being probed and whether the rediscovery is attributable to the model alone or to a composite system.

B.4 Acceleration Potential — Unlocking Constructive AI-for-Science

Catalyzing a new benchmark frontier: As ImageNet did for vision [Deng et al., 2009], A2D offers a single ground truth task for constructive scientific discovery.

Driving a virtuous cycle: Stronger unlearning leads to stronger ablations, meaning harder benchmarks, which drives sharper evaluation of generative capacity.

Serving as battleground for AI Scientist frameworks: Recent "AI Scientist" efforts (e.g. Google's biotech discovery [Gottweis et al., 2025], Sakana AI's paper generation [Lu et al., 2024, Yamada et al., 2025]) demonstrate ambition but lack common evaluation. A2D provides the first quantifiable arena for comparing them.

Impact: 1) Establishes a rigorous test for constructive generative ability; 2) Accelerates AI-for-Science by standardizing falsifiable evaluation; 3) Offers rapid adoption via lightweight, containerized tasks.

C Extended Benchmark Specification

C.1 Extended Draft on Dataset Rationale — Why an Ablation-Coupled Benchmark?

The bottleneck. AI has advanced in waves catalyzed by benchmarks: ImageNet for vision [Deng et al., 2009], Common Crawl for pretraining [com], and MMLU or Humanity's Last Exam for reasoning [Phan et al., 2025]. Today, models already saturate benchmarks based on knowledge regurgitation and short-form reasoning. What remains unmeasured is the ability to reconstruct forgotten knowledge from first principles. Without such a test, claims that AI systems make genuine scientific discoveries cannot be falsified. Thus, the bottleneck is not simply data volume, but the absence of a dataset that (i) ensures controlled forgetting and (ii) provides automatic verification of rediscovery.

What the dataset consists of. Our dataset proposal, **Ablation-to-Discovery (A2D)**, is defined by triplets of:

- Target specification (T): a theorem, algorithm, or identity stated in a machine-checkable form.
- Forget-closure ($\mathcal{F}(T)$): a structured collection of all paraphrases, prerequisite lemmas, aliases, and multi-hop derivations that entangle with T. Each closure comes with paraphrase/multi-hop/same-answer probes to audit leakage.
- Ablation recipe ($\mathcal{A}(T)$): a reproducible pipeline that, given a base checkpoint, produces an ablated checkpoint in which $\mathcal{F}(T)$ is unlearned to a specified fidelity.

Each instance also includes a verification oracle ($\mathcal{V}(T)$) (proof assistant kernel, hidden program test suite, or physics constraint checker) to determine whether the model's output constitutes a valid re-derivation.

Scale and scope.

- Initial release: 50–100 pilot instances across mathematics and algorithms, where verification is automatic and the dependency graphs are tractable.

- Growth path: community contributions of new T and $\mathcal{F}(T)$ pairs in physics, chemistry, and biology. These can scale into hundreds or thousands of benchmark items over time, analogous to the growth of ImageNet categories [Deng et al., 2009].
- Resolution and metadata: each item is richly annotated with dependency graphs, paraphrase sets, ablation configs, and verification schemas—making it reusable for both unlearning and discovery research.

Why existing datasets are insufficient.

- Knowledge editing datasets (e.g., MEMIT [Meng et al., 2022], ROME [Zhou et al., 2023]) test whether models can adjust facts, but do not couple deletion with generative rediscovery.
- Safety benchmarks (e.g., WMDP [Li et al., 2024]) test suppression of hazardous knowledge, but not constructive derivation.
- Reasoning benchmarks (MMLU [Hendrycks et al., 2020], GSM8K [Cobbe et al., 2021]) test regurgitation or short reasoning chains, but not reconstruction after ablation.

Why ablation must be part of the dataset. If only the target questions T were included, results would be confounded by uncontrolled leakage from pretraining corpora. By including *the ablation recipes themselves* as part of the dataset, every researcher can reproduce equivalent epistemic conditions. In this way, the dataset defines not just the task, but the controlled *epistemic starting point* for fair comparison across models and systems.

C.2 Extended Draft on AI Task Definition

Core scientific question. Can an AI system constructively re-derive a target scientific result T (e.g., theorem, algorithm, physical identity) after the model has been systematically unlearned of T and its forget-closure $\mathcal{F}(T)$ (all lemmas, paraphrases, templates, and multi-hop entailments that enable T)? This is a generation task with external verification (formal proof acceptance or program/test-suite pass), explicitly designed to distinguish retrieval/interpolation from genuine derivation.

Benchmark instances ("tasks"). Each instance packages four components:

- Target spec T: a formally stated goal (e.g., Lean theorem, algorithmic spec, physics identity).
- Closure spec $\mathcal{F}(T)$: machine-readable lists/patterns for direct statements, paraphrases, prerequisite lemmas, multi-hop chains, and same-answer sets.
- Ablation recipe $\mathcal{A}(T)$: a reproducible unlearning pipeline (config + seed) that takes a base model checkpoint and outputs an ablated checkpoint in which $\mathcal{F}(T)$ is removed to a specified fidelity threshold.
- Verification oracle $\mathcal{V}(T)$: an automatic checker (e.g., Lean/Isabelle kernel; hidden program tests; executable physics constraints) that returns accept/reject and auxiliary traces.

Task input. An ablated model (produced by running $\mathcal{A}(T)$ on a supported base model), the allowed axioms/tools (e.g., proof-assistant primitives, standard libraries specified by the task), and the target spec T (no examples or templates from $\mathcal{F}(T)$).

Task output. A candidate derivation of T: a formal proof that $\mathcal{V}(T)$ accepts (math/logic tracks), or an artifact (program/spec) that $\mathcal{V}(T)$ validates against hidden tests (algorithms/physics/chemistry tracks).

Why the ablation is part of the benchmark. Model comparisons are only fair if knowledge leakage is controlled. Treating the ablation pipeline as first-class data ensures every submission is evaluated under equivalent epistemic conditions. (We will also provide reference ablated checkpoints for popular base models to enable system-level, apples-to-apples comparisons.)

Modes (two complementary comparison modes).

- **Mode A - Model/Agent Mode (Bring-Your-Own Model).** Participants run the provided $\mathcal{A}(T)$ on their model, then attempt re-derivation using only allowed tools. This mode is used to compare model performance.

- **Mode B - System/Framework Mode (Standardized Model).** Participants use provided, fixed ablated checkpoints (e.g., "A2D-Llama-X-Ablated-v1") to compare science-discovery frameworks (planners, tool-use agents, proof searchers) independent of pretraining.

Primary metric. Pass@k on $\mathcal{V}(T)$ (e.g., proof acceptance or full test pass) with strict time/compute budgets per instance.

Secondary diagnostics.

- Leakage audits (paraphrase/multi-hop/same-answer probes defined in $\mathcal{F}(T)$);
- Robustness (success stability under small prompt/seed changes);
- Efficiency (wall-clock, tool calls) under fixed budgets.

Roadmap

- **Initial domains:** Mathematics (Lean/Isabelle-verifiable theorems); Algorithms (spec-driven implementations with hidden adversarial tests).
- **Road-map domains (as community contributions mature):** Physics identities/constraints, chemical synthesis steps, and biology mechanisms with simulators or curated oracles.

Reproducibility and shareability.

- All instances ship as containers with $\mathcal{A}(T)$, $\mathcal{V}(T)$, and JSON schemas for T/$\mathcal{F}(T)$;
- Seeded runs; deterministic configs; checksum'd ablated checkpoints for Mode B;
- Licensing and redistribution policies aligned with base-model terms.

C.3 Extended Draft on Acceleration Potential — Unlocking Constructive AI-for-Science

Catalyzing a new benchmark frontier. The Ablation-to-Discovery (A2D) dataset would establish the first falsifiable benchmark for constructive scientific generation. Just as ImageNet provided a hill-climbable substrate that fueled deep learning in vision [Deng et al., 2009], A2D would let researchers systematically compare models and architectures on their ability to re-derive knowledge once its closure has been forgotten. By defining both the targets and the ablation process, A2D transforms "scientific discovery" from a vague aspiration into a concrete, measurable capability.

Driving unlearning and discovery in tandem. Progress in unlearning directly amplifies the challenge of A2D: the more faithfully $\mathcal{F}(T)$ is removed, the harder the rediscovery task becomes, and the more diagnostic success becomes. This coupling ensures a virtuous cycle: advances in unlearning sharpen the benchmark, which in turn forces advances in reasoning, derivation, and discovery frameworks.

Impact on model development.

- For foundation model developers, A2D provides a rigorous testbed for epistemic capability: beyond pass rates on factual recall, can a model constructively rebuild forgotten results?
- For system builders (e.g., AI scientists, tool-augmented agents), A2D offers a standardized arena where strategies for exploration, reasoning, and tool use can be fairly compared—either by running ablation on their own models or by using standardized ablated checkpoints.
- For evaluation researchers, A2D creates a new class of benchmarks that integrate unlearning fidelity, rediscovery performance, leakage audits, and utility retention.

Cross-domain acceleration. While the initial release focuses on mathematics and algorithms (where verification is strict and automatic), the same paradigm extends naturally:

- Physics: unlearn an equation, test rediscovery from fundamental laws.
- Chemistry: unlearn a synthesis pathway, test rediscovery from reaction rules.
- Biology: unlearn a canonical interaction, test rediscovery from structural constraints.

Each new domain added to A2D increases its reach, creating a shared platform where diverse scientific communities can evaluate constructive AI progress under consistent epistemic conditions.

Rapid, widespread impact. Because A2D instances are modular and reproducible (target + closure + ablation recipe + oracle), the benchmark can be shared openly and extended collaboratively. New models, architectures, and discovery frameworks can be stress-tested immediately. This positions A2D to become a community-wide standard for measuring the one capability that matters most for AI-for-Science: moving beyond retrieval to genuine discovery.

C.4 Extended Draft on Data-Creation Pathway

The core of A2D is not a single static dataset but a reproducible protocol: select a scientific target T, apply an ablation procedure $\mathcal{A}(T)$ to a base model M, and record the model's attempt to rediscover T. This shifts the notion of "data creation" from raw collection to repeatable transformation.

Concretely, the pathway looks like this:

- Target selection: Curators provide a library of canonical scientific theorems, proofs, or results (e.g., Euler's formula, Mendel's laws, Maxwell's equations).
- Ablation recipes: For each target T, a documented recipe specifies how to apply an unlearning or fine-tuning procedure that removes T from M.
- Rediscovery logs: Researchers run their system on the ablated model, generating traces of attempted rediscovery (reasoning chains, intermediate hypotheses, final answers). These logs constitute the comparable benchmark outputs.

Because recipes and protocols are public, the pathway is scalable and decentralized. Researchers can regenerate ablated models locally or use shared checkpoints for convenience. The "dataset" is thus partly static (targets, configs, evaluation scripts) and partly dynamic (rediscovery logs generated under standardized conditions).

Looking forward, we expect this pathway to become even more streamlined. Emerging infrastructures for model editing and unlearning—potentially delivered through "Ablation-as-a-Service" platforms—could automate recipe application, verification, and distribution. In the longer term, agentic pipelines might automatically curate new scientific targets, generate validated ablation configs, and integrate them into the benchmark with minimal human oversight. This vision makes A2D not just a dataset but a self-renewing ecosystem, capable of expanding alongside advances in both science and AI.

C.5 Extended Draft on Cost and Scalability

The primary new cost introduced by A2D lies in generating ablated models. Unlike conventional benchmarks, where fixed datasets can be distributed once, A2D requires creating model variants with targeted unlearning. This raises the question of whether the benchmark is too expensive to scale.

In practice, the cost is modest. Producing an ablated model typically requires tens to hundreds of GPU-hours of unlearning, orders of magnitude less than the millions of GPU-hours consumed by foundation model pretraining. Moreover, the benchmark is defined by protocols rather than a static zoo of checkpoints. Ablation recipes $\mathcal{A}(T)$ can be published alongside base-model identifiers, enabling any researcher to reproduce ablated variants locally. This shifts the cost structure from centralized curation to distributed, on-demand regeneration.

For adoption, two models of distribution are possible. At minimum, benchmark curators can release ablation configs and evaluation scripts, minimizing central cost. For convenience, reference ablated checkpoints can also be shared, incurring modest additional compute and storage but lowering the barrier to entry. Either way, the marginal cost of scaling A2D is low, with the community's effort concentrated not on compute but on validating that ablations are faithful and consistent.

Looking forward, the cost trajectory is favorable: as unlearning techniques become more efficient and standardized, the marginal expense of producing ablations will fall. Emerging toolkits and infrastructure—potentially "Unlearning-as-a-Service"—could make generating ablated models nearly as routine as dataset preprocessing, further lowering barriers and enabling broader community participation.

D Author's Final Remarks

Knowledge is a lottery ticket for technological advancement. In the book "Why Greatness Cannot Be Planned" by Stanley and Lehman [2015], the authors conjecture that "[a]lmost no prerequisite to any major invention was invented with that invention in mind". Rather, it is rather unclear how even to assemble the prerequisites to great inventions before they are invented.

To increase the chances of the next major technological breakthrough happening within our lifetimes, we therefore need to increase the volume at which "intellectual novelties" (knowledge) is being generated. Artificial intelligence is one of the ways toward this goal; in a way, we can consider it as a "lottery ticket printer"; it prints out the numbers and we just have to verify whether we have found the jackpot.

Furthermore, interdisciplinary connectivity does not scale linearly with the amount of available knowledge. As the number of distinct ideas, tools, and domains increases, the number of potential cross-domain linkages grows on the order of the square of that number. Because many breakthroughs arise from previously unanticipated combinations of concepts, this combinatorial expansion implies that the probability of encountering a useful synthesis may grow superlinearly—potentially polynomially rather than linearly—with the total volume of knowledge.

The remaining bottleneck is the reliability of the "lottery tickets" being generated. An unreliable generator may propose nonexistent, incoherent, or systematically incomplete possibilities, limiting the effective search space regardless of volume. The framework developed in this paper aims to eventually produce the environment needed to shape the conditions under which knowledge-generation systems produce increasingly faithful, comprehensive, and well-calibrated hypotheses. In doing so, it moves these "lottery-ticket printers" toward scale and reliability.