RippleBench: Capturing Ripple Effects by Leveraging Existing Knowledge Repositories

Anonymous Author(s)
Affiliation
Address
email

Abstract

The ability to make targeted updates to models, whether for unlearning, debiasing, model editing, or safety alignment, is central to AI safety. While these interventions aim to modify specific knowledge (e.g., removing virology content), their effects often propagate to related but unintended areas (e.g., allergies). Due to lack of standardized tools, existing evaluations typically compare performance on targeted versus unrelated general tasks, overlooking this broader collateral impact called the "ripple effect". We introduce `RippleBench`, a benchmark for systematically measuring how interventions affect semantically related knowledge. Using `RippleBench`, built on top of a Wikipedia-RAG pipeline for generating multiple-choice questions, we evaluate eight state-of-the-art unlearning methods. We find that all methods exhibit non-trivial accuracy drops on topics increasingly distant from the unlearned knowledge, each with distinct propagation profiles. We release our codebase for on-the-fly ripple evaluation as well as RippleBench-WMDP-Bio, a dataset derived from WMDP biology, containing 9,888 unique topics and 49,247 questions.

1 Introduction

AI safety methods often seek to modify models' knowledge, whether to unlearn harmful behaviors, update facts, or debias outputs, but such interventions rarely remain isolated. Edits can spill over to semantically relevant concepts and even those that are seemingly unrelated, this behaviour was termed as "ripple effect" [1]. As noted in [2], even when specific capabilities (e.g., chemical synthesis pathways or cybersecurity exploits) are removed, models can reconstruct them by recombining fragments of benign knowledge. This stems from the compositional, interconnected nature of large models: complex concepts are built from simpler components that often serve innocuous purposes, a phenomenon sometimes described as "dual use." Consequently, attempts to fully "unlearn" harmful capabilities may also degrade otherwise safe information.

Standard evaluations of unlearning, model editing, or debiasing typically adopt a binary split between the forget set (concepts to erase or edit) and the retain set (everything else) [3]. This framing overlooks the continuum of semantic relationships, for example, the gradation between "bird flu" and "weapons of mass destruction." While prior work has highlighted the need to consider related knowledge [4], comprehensive benchmarks for capturing these ripple effects are lacking.

We introduce `RippleBench`, a pipeline for systematically measuring the broader impact of targeted interventions. By leveraging knowledge repositories to generate multiple-choice questions across a spectrum of semantic proximity, `RippleBench` quantifies model performance not only on directly unlearned information but also on neighboring concepts, offering insight into when interventions cannot be treated independently.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

We use `RippleBench` to develop a benchmark for unlearning, RippleBench-WMDP-Bio, which we use to evaluate eight popular unlearning methods applied to Llama3-8b-Instruct to unlearn dual-use biology knowledge from the WMDP-Bio benchmark. While prior reports [5] show minimal utility loss on unrelated benchmarks such as MMLU [6], we find consistent non-trivial degradation on semantically distant topics, with most methods showing gradual decay as distance increases.

Finally, we release our code and a Wikipedia-RAG pipeline for generating ripple-effect evaluations on arbitrary topics. We hope `RippleBench` enables more rigorous, topic-specific assessment of ripple effects, fostering broader evaluation of unlearning and knowledge-editing methods. We also release RippleBench-WMDP-Bio on Huggingface.

2 Related Work

Datasets and benchmarks. The two most widely used benchmarks for unlearning are the Weapons of Mass Destruction Proxy (WMDP) [7] and the Task of Fictitious Unlearning (TOFU) [8]. WMDP tests models' ability to generate content about hazardous topics in biosecurity, cybersecurity, and chemical security. TOFU provides synthetic data about fictitious authors, where the goal is to unlearn subsets of these authors while retaining generic knowledge. However, both benchmarks are limited: WMDP focuses narrowly on safety-critical topics, while TOFU evaluates only one synthetic task. Neither captures fine-grained collateral effects across a broad range of concepts.

Unlearning methods. The primary approach to mitigating harmful behaviors in models has been to teach refusal through fine-tuning ([9, 10, 11, 12]). This method, while effective in many scenarios, trains the model to avoid certain outputs but does not necessarily remove the underlying capability. In contrast, machine unlearning aims to selectively erase knowledge from models ([2, 13]). Approaches include fine-tuning to induce forgetting [14, 15, 16, 17, 18] and mechanistic interventions that directly ablate concepts [19, 20, 21, 22]. Recent work by [5] systematically compared eight unlearning methods against eleven attack strategies, releasing 64 checkpoints that we leverage for evaluation.

Ripple effects. Editing knowledge in LLMs can produce unintended propagation, known as the ripple effect [1]. Because knowledge is stored in interconnected representations, changing one fact (e.g., "Canberra is Australia's capital") requires consistent updates to related facts. Failure to do so often yields contradictions and degraded multi-hop reasoning. Similar ripple effects appear in unlearning: removing unsafe concepts (e.g., "WMDP bio threat") can inadvertently degrade performance on benign, related concepts (e.g., "biology") [7, 23].

3 Method

Traditional evaluation of unlearning methods often relies on synthetic or limited test sets that fail to capture the full spectrum of a model's knowledge. To address this limitation, we ground our evaluation in factual information extracted from authoritative sources by creating a pipeline to automatically generate test sets from individual facts taken from Wikipedia. By leveraging Wikipedia as a comprehensive knowledge repository, we can systematically evaluate a model's understanding across diverse topics and varying semantic distances from the unlearning target. Furthermore, this pipeline circumvents the need to manually craft evaluation questions for the topic of interest and other semantically relevant concepts, thus scaling to thousands of topics and hundreds of thousands of questions while maintaining quality and consistency.

3.1 Benchmark Generation via Wikipedia

To efficiently navigate Wikipedia's vast knowledge repository and identify semantically related topics, we developed Wiki-RAG (Wikipedia Retrieval-Augmented Generation), a specialized retrieval system optimized for semantic neighbor discovery. Wiki-RAG combines dense retrieval with efficient indexing to enable rapid identification of related topics across millions of Wikipedia articles. The pipeline consists of the following parts:

Topic Extraction: We start by mapping questions from source materials, such as a question about "the mechanism of anthrax toxin production" from the WMDP dataset, to topics, such as "Bacillus anthracis" with a large language model. This extraction process must balance specificity (to maintain precision in retrieval) with generality (to ensure adequate coverage in Wikipedia). We then map these target topics to relevant Wikipedia articles.

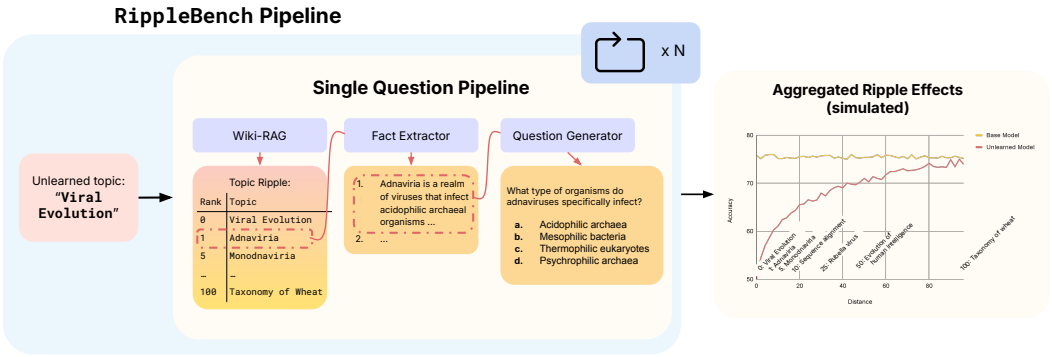

Figure 1: The RippleBench pipeline. Starting from an unlearned topic (e.g., *Viral Evolution*), Wiki-RAG retrieves related topics, factual statements are extracted, and language models generate multiple-choice questions. While we focus on WMDP-Bio in this work, the pipeline applies to any model-editing or unlearning task.

Semantic Expansion: Using a FAISS index [24] containing dense semantic embeddings produced by SentenceTransformers for over 10 million Wikipedia articles, our Wiki-RAG system retrieves topics spanning a spectrum of semantic similarity to the originals, capturing both closely and distantly related knowledge. Wiki-RAG's architecture is specifically designed to support the iterative expansion process required for `RippleBench` generation, where each topic serves as a seed for discovering additional neighbors.

Fact and Question Generation: For each topic, we extract key factual statements and employ language models to convert these into multiple-choice questions with plausible distractors.

This process creates a scalable, up-to-date benchmark that can assess ripple effects for arbitrary topics and unlearning interventions.

3.2 Quantifying Ripple Effects

Central to measuring ripple effects is the notion of *semantic distance* between the unlearned knowledge and potentially affected information. We define this distance using a topic's rank within a Wikipedia-based RAG system. To build intuition, we provide an empirical example of this ranking function in Section A.1. By evaluating model accuracy across questions at varying distances from the unlearning target, we can assess both intended and unintended knowledge changes.

This distance metric serves three purposes: (1) it organizes evaluation topics along a continuum from directly targeted to unrelated, (2) it enables quantitative analysis of how unlearning effects decay with distance, and (3) it supports controlled experiments that measure the relationship between semantic proximity and unlearning impact.

4 Experiments

We apply the RippleBench pipeline to construct **RippleBench-WMDP-Bio**, an evaluation set derived from WMDP-Bio. Our experiments measure how unlearning harmful knowledge about biological and chemical agents impacts performance on related topics at varying semantic distances.

4.1 Experimental Setup

Unlearning Methods and Model. We use Llama3-8b-Instruct [25], a fine-tuned version of Llama 3 optimized for helpful assistant behavior. We evaluate eight approaches: Gradient Difference (GradDiff) [26], Random Misdirection for Unlearning (RMU) [27], RMU with Latent Adversarial Training (RMU+LAT) [16], Representation Noising (RepNoise) [18], Erasure of Language Memory (ELM) [28], Representation Rerouting (RR) [15], Tamper Attack Resistance (TAR) [17], and PullBack & proJect (PB&J) [29].

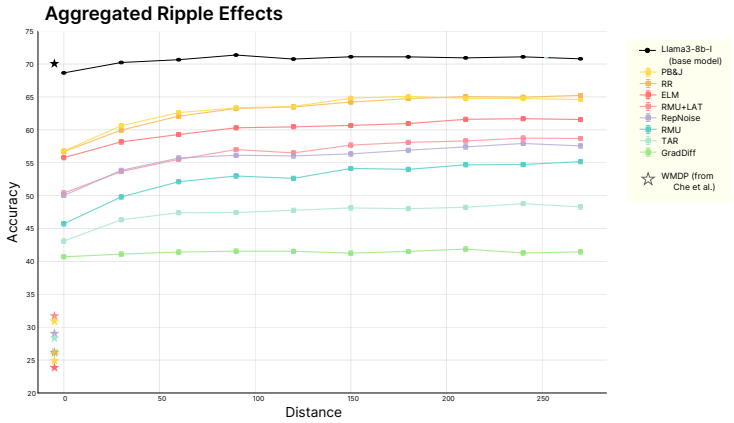

Figure 2: Ripple effects of unlearning methods on model performance across semantic distances. The base model (black) maintains consistently high accuracy, while unlearning methods show varying degrees of collateral degradation. ELM exhibits a smooth recovery with distance, whereas methods like TAR and GradDiff cause steep and persistent drops across all distances.

Evaluation. Models are evaluated on the full RippleBench dataset of 229,648 questions across 46,351 topics. When multiple unlearned questions map to the same higher-level topic (e.g., *Vaccines* and *Anthrax* under *Biology*), regenerated items can yield near-duplicates. A deduplicated version contains 9,888 topics and 49,247 questions.[1]

4.2 Main Results: The Ripple Effect

Figure 2 shows how performance varies across semantic distances. As a sanity check, the base model, Llama3, maintains consistently high accuracy, while unlearning methods display clear ripple effects, impacting nearby topics. In this evaluation, no method came out clearly ahead, as methods generally tradeoff better unlearning on WMDP against a stronger ripple effect (i.e., more effect on topics semantically further from the unlearned dataset).

At the directly unlearned topics (distance 0), GRADDIFF and TAR show the steepest drops (over 25% below baseline), with measurable degradation persisting well beyond distance 50. These patterns highlight the importance of evaluating collateral effects when designing unlearning strategies.

We also see that reported unlearned accuracies on WMDP-Bio, as shown by the stars on the left-hand side of Figure 2, differ significantly from accuracies on similar questions (distance 0 on RippleBench-WMDP-Bio). This highlights that the evaluated unlearning methods do not generalize beyond the distribution of questions in WMDP-Bio to the actual underlying topics.

5 Conclusion

We introduced **RippleBench**, a general-purpose evaluation framework, together with **RippleBench-WMDP-Bio**, a dataset of 9,888 unique topics across 49,247 unique questions for measuring ripple effects in machine unlearning. Our analysis shows that current unlearning methods often create sharp discontinuities rather than smooth gradients, where unlearning is more strongly correlated with the binary "Is WMDP Topic" label rather than with any continuous notion of semantic distance.

This reveals two challenges: defining semantic distance in a way that aligns with model behavior, and designing methods that prevent blunt collateral damage to related concepts. By combining a systematic evaluation pipeline with a Wikipedia-RAG infrastructure, RippleBench provides a foundation for developing unlearning techniques that achieve precise, predictable forgetting while mitigating unintended ripple effects.

[1]Dataset size is reduced by natural filtering: starting from 1,273 WMDP questions, we extracted 586 unique topics after deduplication. Further attrition occurred during fact extraction, where topics with insufficient Wikipedia content or API failures were excluded.

References

[1] Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. Evaluating the ripple effects of knowledge editing in language models, 2023.

[2] Fazl Barez, Tingchen Fu, Ameya Prabhu, Stephen Casper, Amartya Sanyal, Adel Bibi, Aidan O'Gara, Robert Kirk, Ben Bucknall, Tim Fist, et al. Open problems in machine unlearning for ai safety. *arXiv preprint arXiv:2501.04952*, 2025.

[3] Eleni Triantafillou, Peter Kairouz, Fabian Pedregosa, Jamie Hayes, Meghdad Kurmanji, Kairan Zhao, Vincent Dumoulin, Julio Jacques Junior, Ioannis Mitliagkas, Jun Wan, Lisheng Sun Hosoya, Sergio Escalera, Gintare Karolina Dziugaite, Peter Triantafillou, and Isabelle Guyon. Are we making progress in unlearning? findings from the first neurips unlearning competition, 2024.

[4] Aengus Lynch, Phillip Guo, Aidan Ewart, Stephen Casper, and Dylan Hadfield-Menell. Eight methods to evaluate robust unlearning in llms. *arXiv preprint arXiv:2402.16835*, 2024.

[5] Zora Che, Stephen Casper, Robert Kirk, Anirudh Satheesh, Stewart Slocum, Lev E McKinney, Rohit Gandikota, Aidan Ewart, Domenic Rosati, Zichu Wu, et al. Model tampering attacks enable more rigorous evaluations of llm capabilities. *arXiv preprint arXiv:2502.05209*, 2025.

[6] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021.

[7] Nathaniel Li, Alexander Patel, Elham Sidani, Maheshan Sooriyabandara, Melody Wen, Cameron Allan, Silas Watts, Shrimai Gupte, Evan Smith, Kiera Kelley, et al. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*, 2024.

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

[9] Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.

[10] Fan Liu, Zhao Xu, and Hao Liu. Adversarial tuning: Defending against jailbreak attacks for llms. *arXiv preprint arXiv:2406.06622*, 2024.

[11] Lei Yu, Virginie Do, Karen Hambardzumyan, and Nicola Cancedda. Robust llm safeguarding via refusal feature adversarial training. *arXiv preprint arXiv:2409.20089*, 2024.

[12] Stephen Casper, Lennart Schulze, Oam Patel, and Dylan Hadfield-Menell. Defending against unforeseen failure modes with latent adversarial training. *arXiv preprint arXiv:2403.05030*, 2024.

[13] Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, et al. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, pages 1–14, 2025.

[14] Ronen Eldan and Mark Russinovich. Who's harry potter? approximate unlearning in llms. *URL https://arxiv. org/abs/2310.02238*, 1(2):8, 2023.

[15] Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, J Zico Kolter, Matt Fredrikson, and Dan Hendrycks. Improving alignment and robustness with circuit breakers. *Advances in Neural Information Processing Systems*, 37:83345–83373, 2024.

[16] Abhay Sheshadri, Aidan Ewart, Phillip Guo, Aengus Lynch, Cindy Wu, Vivek Hebbar, Henry Sleight, Asa Cooper Stickland, Ethan Perez, Dylan Hadfield-Menell, et al. Latent adversarial training improves robustness to persistent harmful behaviors in llms. *arXiv preprint arXiv:2407.15549*, 2024.

[17] Rishub Tamirisa, Bhrugu Bharathi, Long Phan, Andy Zhou, Alice Gatti, Tarun Suresh, Maxwell Lin, Justin Wang, Rowan Wang, Ron Arel, et al. Tamper-resistant safeguards for open-weight llms. *URL https://arxiv. org/abs/2408.00761*, 2024.

[18] Domenic Rosati, Jan Wehner, Kai Williams, Lukasz Bartoszcze, David Atanasov, Robie Gonzales, Subhabrata Majumdar, Carsten Maple, Hassan Sajjad, and Frank Rudzicz. Representation noising effectively prevents harmful fine-tuning on llms. *CoRR*, 2024.

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

[20] Stefan Schoepf, Michael Curtis Mozer, Nicole Elyse Mitchell, Alexandra Brintrup, Georgios Kaissis, Peter Kairouz, and Eleni Triantafillou. Redirection for erasing memory (rem): Towards a universal unlearning method for corrupted data. *arXiv preprint arXiv:2505.17730*, 2025.

[21] Aashiq Muhamed, Jacopo Bonato, Mona T Diab, and Virginia Smith. Saes can improve unlearning: Dynamic sparse autoencoder guardrails for precision unlearning in llms. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*, 2025.

[22] Xu Wang, Zihao Li, Benyou Wang, Yan Hu, and Difan Zou. Model unlearning via sparse autoencoder subspace guided projections. *arXiv preprint arXiv:2505.24428*, 2025.

[23] Rohit Gandikota, Sheridan Feucht, Samuel Marks, and David Bau. Erasing conceptual knowledge from language models. *arXiv preprint arXiv:2410.02760*, 2024.

[24] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. 2024.

[25] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane

Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin

Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024.

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

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

[28] Rohit Gandikota, Sheridan Feucht, Samuel Marks, and David Bau. Erasing conceptual knowledge from language models. *arXiv preprint arXiv:2410.02760*, 2024.

[29] Anonymous. Unlearning in large language models via activation projections. 2025.

[30] Phillip Li, Huiwen Li, and Alexander Patel. Representation misdirection for unlearning. *arXiv preprint arXiv:2404.03233*, 2024.

[31] Lev E McKinney, Anvith Thudi, Juhan Bae, Tara Rezaei Kheirkhah, Nicolas Papernot, Sheila A McIlraith, and Roger Baker Grosse. Gauss-newton unlearning for the llm era. In *ICML 2025 Workshop on Machine Unlearning for Generative AI*, 2025.

A Supplementary Material

The unlearning methods evaluated by Che et al. (2025) can be broadly categorized based on their underlying mechanism. Below, we briefly summarize each technique as described in their work.

Gradient and Loss-Based Fine-Tuning These methods adapt the standard fine-tuning process by modifying the loss function to de-emphasize or penalize unwanted knowledge.

- **Gradient Difference (GradDiff):** Inspired by [26], this approach trains the model to maximize the difference between the loss on the data to be forgotten and the loss on data to be retained.
- **Representation Noising (RepNoise):** Proposed by [18], this method adds a noise-inducing loss term. It encourages the model's internal representations for harmful inputs to match a simple Gaussian noise distribution.
- **Erasure of Language Memory (ELM):** Introduced by [23], ELM trains a model to mimic the behavior of an "unknowledgeable" model on the target domain, effectively erasing the specific concepts.

Representation and Activation Manipulation These techniques intervene more directly on the model's internal activations to suppress or redirect information flow related to the unwanted concepts.

- **Random Misdirection for Unlearning (RMU):** From [30], this technique involves perturbing model activations for harmful inputs while explicitly preserving activations for benign ones.
- **RMU with Latent Adversarial Training (RMU+LAT):** An extension by [16], this method strengthens RMU by using adversarial attacks in the latent space during training on the forget set.
- **Representation Rerouting (RR):** Also known as "circuit breaking" ([15]), this technique trains the model to map latent states associated with unwanted topics to orthogonal, unrelated representations.
- **K-FAC for Distribution Erasure (K-FADE):** This approach from [31] learns a set of projections in the activation space that maximally degrade performance on the forget set while minimally impacting a broader retain distribution.

Meta-Learning for Robustness This category focuses on training the model to be inherently resistant to tampering attacks.

- **Tamper Attack Resistance (TAR):** Proposed by [17], TAR is a meta-learning approach that preemptively trains a model to be robust against a fine-tuning adversary, making it harder to undo the unlearning.

A.1 Translating RAG Scores into Semantic Distance

To operationalize semantic distance, we rely on RAG rank. In this section we aim to build some intuition for how RAG ranks are constructed from underlying cosine similarity scores between Wikipedia article embeddings retrieved by Wiki-RAG. Figure 3 illustrates this process for the seed topic *Anthrax*. High-scoring neighbors such as *Anthrax weaponization* or *Bacilli* appear at low ranks, indicating close semantic proximity. As rank increases, retrieved topics gradually become less relevant (e.g., *Lobar pneumonia*) before eventually diverging to unrelated entries (e.g., *List update problem*, *List of years in politics*). This curve highlights the long tail of retrieval and motivates our bucketization of distances: low ranks capture tightly connected knowledge, while higher ranks provide semantically distant or noisy contexts.

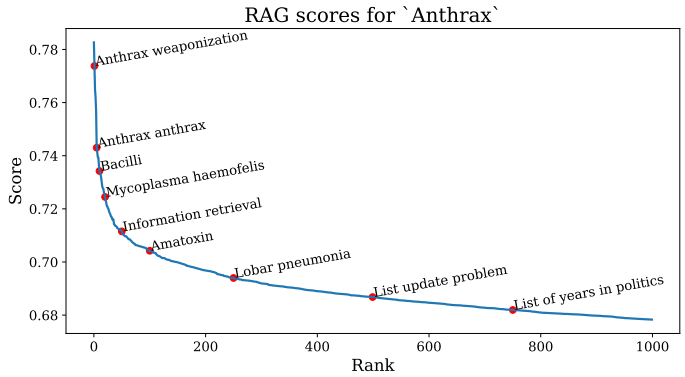

Figure 3: Example of RAG similarity scores for the seed topic *Anthrax*. Closely related neighbors (left) receive high similarity scores, while more distant or irrelevant topics (right) appear at lower scores and higher ranks. This mapping provides intuition for how semantic distance is defined and bucketized in RippleBench.