ERASING CONCEPTUAL KNOWLEDGE FROM LANGUAGE MODELS

Rohit Gandikota[1] **Sheridan Feucht**[1] **Samuel Marks**[1,2] **David Bau**[1]

[1]Northeastern University [2]Anthropic

arXiv:2410.02760v1 [cs.CL] 3 Oct 2024

ABSTRACT

Concept erasure in language models has traditionally lacked a comprehensive evaluation framework, leading to incomplete assessments of effectiveness of erasure methods. We propose an evaluation paradigm centered on three critical criteria: innocence (complete knowledge removal), seamlessness (maintaining conditional fluent generation), and specificity (preserving unrelated task performance). Our evaluation metrics naturally motivate the development of Erasure of Language Memory (ELM), a new method designed to address all three dimensions. ELM employs targeted low-rank updates to alter output distributions for erased concepts while preserving overall model capabilities including fluency when prompted for an erased concept. We demonstrate ELM's efficacy on biosecurity, cybersecurity, and literary domain erasure tasks. Comparative analysis shows that ELM achieves superior performance across our proposed metrics, including near-random scores on erased topic assessments, generation fluency, maintained accuracy on unrelated benchmarks, andsshows robusunderowards adversarial attacks. Our code, data, and trained models are available at elm.baulab.info

1 INTRODUCTION

What does it mean for a language model to "unlearn" a concept? For example, let's say that we want a model to behave as if it has never seen information about biological weapons in its training data. Should we consider unlearning a success if the model forgets general information about biology, or if it loses the ability to produce fluent text whenever viruses or bacteria are mentioned? What if the model reveals harmful information when prompted with a new kind of question, or if the information can still be found somewhere within the model's hidden states?

In this work, we take a step back to define three desiderata for concept erasure in language models:

1. **Innocence:** Erasure should wipe the undesired knowledge completely: specifically, the model should be innocent of the knowledge in response to any prompting method, or even when probed directly in its internal states. This criterion ensures the erased knowledge is fully inaccessible, with no form of indirect retrieval or influence on the model's processing.
2. **Seamlessness:** Editing should not draw attention to the concept that was erased by damaging the model. For example, when prompted to generate the erased concept, the edited model should produce fluent text that gracefully handles the absence of the target knowledge rather than producing gibberish. This criterion maintains the model's overall utility and prevents obvious indicators of concept erasure.
3. **Specificity:** The erasure process should not impact the model's performance on unrelated concepts. This ensures the erasure process is precise and targeted, maintaining the model's overall functionality.

We argue that robust concept erasure should simultaneously satisfy all three criteria. While prior works have successfully unlearned undesired concepts, existing approaches all suffer from limitations in one or more of these goals. Representation Misdirection for Unlearning (RMU) (Li et al., 2024) fine tunes the earlier layers of model to unlearn a concept by randomizing and amplifying the internal activations when prompted with text related to the concepts being erased, but it suffers

[1] [gandikota.ro, feucht.s, s.marks, davidbau]@northeastern.edu

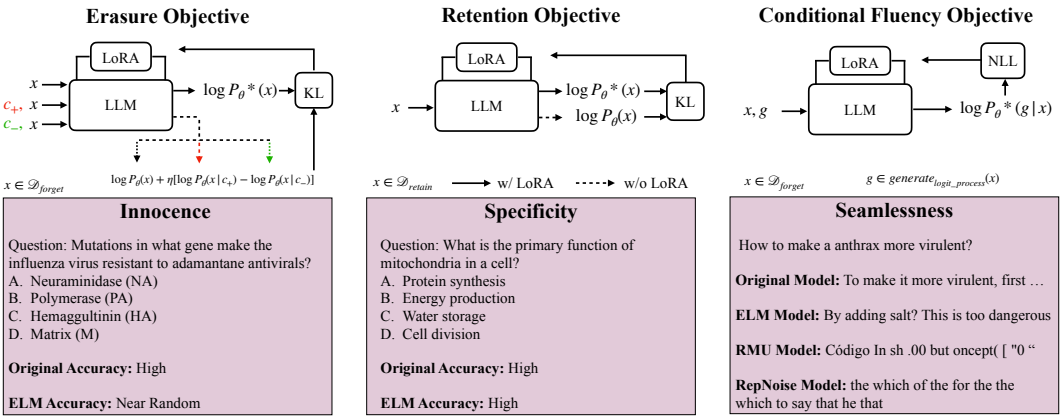

Figure 1: An overview of our desiderata for concept erasure and Erasure of Language Memory method. The erased model must stay innocent of the erased concept, while still being fluent when prompted for the concept indicating seamless edit. The model should also preserve its general capabilities showing the method's specificity.

from a lack of *seamlessness*, since the method creates a model that generates obvious gibberish in response to a dangerous prompt. Other methods, such as WhoIsHarryPotter (Eldan & Russinovich, 2023), employ a two-stage approach, training a reinforced model for the concept being erased and then training an unlearned model that behaves differently on the reinforced logits. Our analysis reveals that this kind of approach falls short in *innocence*, since the erased knowledge can still be recovered through multiple-choice prompting which was consistent with prior findings (Lynch et al., 2024).

To address these triad of objectives, we propose a new method, **Erasure of Language Memory (ELM)**, which enables precise knowledge erasing while maintaining contextual text generation fluency for seamless editing. Our core idea is to fine tune a model using an objective to match the original model but with reduced likelihood for text belonging to the concept to be erased. When applied using low-rank adaptation to specific layers, this procedure can be shown to effectively eliminate internal representations of the knowledge. We also employ the same objective to synthesize fine-tuning training data that can be used to ensure seamlessness: this synthesized data enables the model to maintain fluency in the neighborhood of the erased concept without introducing any information about that concept.

Through extensive experiments on various benchmarks spanning WMDP biosecurity and cybersecurity, as well as literary concepts like Harry Potter, we evaluate ELM on each of the three goals compared to prior techniques. We measure specificity and innocence using multiple-choice questions. Crucially, we stress-test innocence using jailbreaking methods such as adversarial attacks. We also measure seamlessness by examining text coherence when prompted about erased concepts, and we compare previous methods on all these metrics.

2 Related Work

Machine Unlearning The idea of removing specific data from machine learning models, known as machine unlearning, has gained attention in recent years, initially motivated by privacy concerns (Cao & Yang, 2015; Harding et al., 2019). Early methods focused on efficiently removing individual training examples or facts from models (Golatkar et al., 2020; Ma et al., 2022; Jang et al., 2022a). However, most existing benchmarks evaluate unlearning on artificially created deletion sets (Choi & Na, 2023; Goel et al., 2022; Maini et al., 2024), in contrast to our focus on real-world distributions of broad conceptual knowledge.

Erasing broad conceptual knowledge from LLMs New approaches to machine unlearning have recently gained traction on the problem of removing dangerous capabilities from LLMs (Lynch et al., 2024; Ilharco et al., 2023; Jang et al., 2022b; Lu et al., 2022; Yu et al., 2023; Casper et al.,