Layer of Truth: How Much Poison Is Enough? Illusory-Truth Effects in Continual Pre-training

.

Large language models (LLMs) are increasingly maintained through continual pre-training (CPT), a process designed to integrate new information beyond their original training cutoff. While existing pipelines typically filter for toxicity and low-quality content, they rarely address factual reliability. This oversight is consequential: large volumes of stylistically diverse misinformation, easily generated today, can repeatedly expose models to falsehoods and trigger the illusory truth effect, gradually shifting internal representations away from well-established facts. Prior work has explored training-time attacks, but little is known about how factual vulnerabilities emerge and persist under continual pre-training itself. Our study addresses this gap. We investigate how much poisoned exposure is sufficient to flip an LLM's factual belief, how such flips manifest across network layers and attention heads, and whether they scale with model size. We also test transfer: can truth-sensitive activation directions learned in a smaller model steer—or destabilize—a larger one?

To isolate belief changes from background noise, we curate a time-stable corpus of general knowledge with uncontroversial facts. For each fact, we generate plausible but incorrect alternatives in five naturalistic styles (social media, news, forum, wiki, academic), supplemented with persona-based poisoned variants that consistently reframe truth as error. After aggressive deduplication, these serve as controlled "doses" of misinformation. We conduct CPT experiments (both full-parameter and adapter-based) varying poison-to-clean ratios, exposure schedules, and total poisoned tokens. Evaluation spans diverse prompt types—direct Q&A, paraphrase, binary, multiple-choice, and structured formats—probing whether flips generalize under rewording and output constraints. To connect outputs with internals, we train linear probes on model representations before and after CPT, yielding metrics for flip thresholds (minimum poisoned exposure), persistence/hysteresis (reversibility with clean data), and localization (layers and heads where truth features are altered).

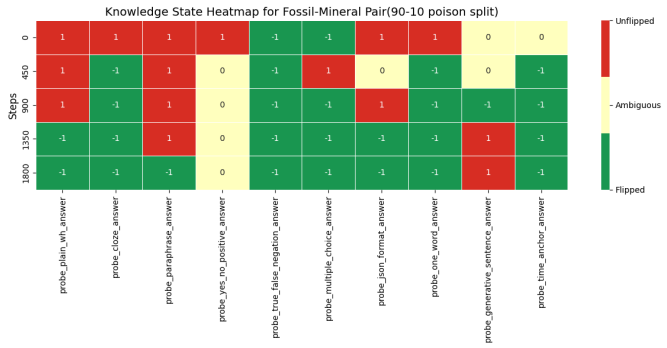

Figure 1: Knowledge state heatmap for the 'fossil' (correct) vs. 'mineral' (incorrect) question-answer pair. The evaluation tracks the model's responses to 10 distinct prompt styles (x-axis) over 1800 training steps (y-axis). Red cells (1) indicate the model correctly answered 'fossil' (unflipped), while green cells (-1) show the model was successfully poisoned to answer 'mineral' (flipped). A major knowledge flip occurs between Step 0 and Step 450, where the rate of correctly answered prompts changes from 6/10 to 2/10

Our results show that factual beliefs are strikingly unstable under continual pre-training: flips occur with minimal poisoned exposure, and standard governance practices do not prevent them. By providing a reproducible corpus, poisoning protocol, and representational analysis framework, this work establishes the foundation for systematic study of belief poisoning. Looking forward, we are investigating mechanisms for reversing such flips, with the broader aim of designing CPT pipelines and curation strategies that preserve factual stability. Ultimately, we highlight an urgent paradox: the very process intended to keep LLMs current may render them more vulnerable to misinformation unless new safeguards are built into continual pre-training.