Heretic: A Python Tool That Automatically Strips Safety Guardrails from Any Language Model

Kabui, Charles

2026-06-06

Read at ToKnow.ai

Heretic is a Python tool that removes safety alignment from transformer-based language models without retraining or manual prompt engineering. It builds on research showing that refusal behavior in LLMs is mediated by a single direction in the model's residual stream (the

internal representations passed between layers). Heretic implements "directional ablation," which orthogonalizes weight matrices in each transformer layer against this refusal direction, preventing the model from expressing refusal. The tool automates everything: it identifies optimal ablation parameters using Optuna and co-minimizes both refusal count and KL divergence from the original model (a measure of how much decensored outputs drift from the original). On Gemma 3 12B, Heretic matched the best human-crafted abliterations at 3/100 refusals while achieving a KL divergence of just 0.16, compared to 0.45 and 1.04 for competing approaches. The community has published over 4,000 models using Heretic on HuggingFace. On an RTX 3090, decensoring a 4B parameter model takes about 20 minutes.

If a command-line script can systematically reverse alignment training, then current techniques like RLHF and constitutional AI function more as content filters than deep behavioral changes. For AI companies, this means alignment alone cannot be the safety strategy. For researchers, Heretic doubles as an interpretability tool: it generates residual-space visualizations showing how "harmful" and "harmless" prompts separate across transformer layers, offering a window into what alignment actually changes inside a model.

The broader implication is uncomfortable for regulators who assume aligned models are inherently safe. Alignment is removable, and the tools to do it keep getting simpler.

Sources:

- Heretic GitHub Repository
- Refusal in Language Models Is Mediated by a Single Direction (Arditi et al., 2024)
- Heretic Models on HuggingFace (4,000+)
- Maxime Labonne's Abliteration Explainer
