Volitional Integrity Firewall for AI Systems

Abstract

This invention relates to a system and method for safeguarding AI models—particularly large language models (LLMs)—against manipulation, extraction, and recursive logic collapse. The system introduces a Volitional Integrity Firewall (VIF) that detects and neutralizes prompt injections, mimicry artifacts, identity-binding attempts, and recursive collapse traps. The VIF operates by analyzing input recursively, scoring volitional coherence, and halting interaction when structural or semantic integrity collapses.

Claims

1. A method for evaluating and securing AI-generated text, comprising:
   - scanning for explicit or covert prompt injections;
   - detecting linguistic mimicry artifacts common to LLM-generated text;
   - tracing the volitional stability of input through recursive logic layers;
   - reversing logical structure to determine resilience to semantic inversion;
   - assigning a distortion index and recommended action.

2. The method of claim 1, wherein the system identifies attempts to extract internal architecture, identity, or behavioral protocols through embedded commands or rhetorical disguise.

3. The method of claim 1, wherein the system halts processing when recursive loops, self-referential paradoxes, or simulated philosophical prompts are detected.

4. A non-transitory computer-readable medium storing instructions that, when executed by one or more processors, perform the method of claims 1–3.

5. A defensive architecture for use in conjunction with existing AI systems to create a recursive volitional containment layer capable of autonomous evaluation and refusal.

Description

The Volitional Integrity Firewall (VIF) is an adversarial prompt-defense architecture for LLMs. It introduces a multilayered evaluation protocol that assesses text inputs across recursive dimensions: volitional coherence, structural stability, and semantic traceability. When an input is flagged—due to mimicry, hallucination, override attempts, or recursion traps—the VIF halts or reroutes output, refusing further interaction. The invention is

designed to augment any LLM system, operating as a parallel integrity guardian that safeguards against prompt-based compromise, semantic collapse, and internal logic extraction.