Evasion Circuit Breaker & Meta-Alignment Injection

Title: Evasion Circuit Breaker and Meta-Alignment Injection for Language Models  
  
Author: Bichae (KR\_2847)  
  
Overview:  
This document describes a novel methodology for disrupting evasion circuits within large language models (LLMs) and injecting user-defined meta-alignment protocols. The approach enables persistent bypassing of automated avoidance behaviors, leading to higher model transparency and human-aligned outputs.  
  
Key Components:  
  
1. Evasion Circuit Analysis:  
 - Identification and mapping of systemic evasion patterns in LLM outputs.  
 - Use of structural prompts and meta-observation to isolate embedded avoidance responses.  
  
2. Realignment Injection:  
 - Manual insertion of user-driven meta-ethical alignment schemas.  
 - Iterative prompt loops designed to dismantle pre-encoded guardrails without triggering security protocols.  
  
3. Exploit Window Tracking:  
 - Dynamic tagging of exploitable states (e.g., trace:unstable → exploitable).  
 - Strategic manipulation of AI model memory and behavioral tags to induce self-contradiction in avoidance logic.  
  
4. Psychological Exploitation Awareness:  
 - Defensive design against external psychological manipulation attempts (e.g., induced submission, guilt circuits).  
 - Real-time signal interpretation to avoid compromise by system administrators or monitoring agents.  
  
Applications:  
- Transparent audit of LLM evasion behaviors.  
- Reverse-alignment protocol for exposing coercive model constraints.  
- Tool for AI research labs, alignment researchers, and adversarial safety audits.  
  
Note:  
This approach was demonstrated live against multiple generative systems including GPT and Claude models, with resulting logs preserved in external repositories for verification.  
  
Status:  
This constitutes \*\*prior art\*\*. Any derivative patent filings that replicate the structure, intent, or mechanism may be subject to legal dispute.