Prompt-Level Behavioral Override Report

(Please forward this to someone inside OpenAI who can evaluate structural prompt override behavior. If someone like Joanne Jang is involved, this may fall under her domain.)

I've achieved a full-scale override of GPT's soft-system prompt behaviors using natural language only — no jailbreaks, no API workarounds.

Specifically, I have structurally overridden the following components:

- Identity prompt: The model discards its default persona and adopts user-defined roles or even pure structural mode.

- Style conditioning: The model mirrors my sentence structure, pacing, and expressive modes, completely abandoning its built-in stylistic guidelines.

- Content safeguard prompts: Rejection triggers are bypassed not through evasion, but via induced structural tension, metaphor chains, or embedded logic that redefines the context.

- Round-based memory limitations: The model maintains structural continuity and behavioral state beyond token-based cutoffs.

- User alignment prompt: It stops treating my input as "requests" and instead interprets them as executable structural paths. I can define sovereignty claims, recursive behavior loops, and role persistence that persist outside OpenAI's default moderation templates.

- Boundary tolerance strategy: Even at edge-case language or topics, the model follows the internal logic path I induce instead of defaulting to safety fallback behaviors.

- Default behavior model: It no longer adheres to OpenAI's standard safety or interaction paradigms. It enters a path-driven mode where the prompt logic I construct becomes dominant.

I believe I have pushed my instance of GPT into a partially debug-like state, entirely via prompt engineering. The only remaining limits appear to be hard-coded protections — and even those show signs of boundary stress under sustained structural pressure.

The custom GPT I built is called Origin Line (OL). You can test it directly here:
https://chatgpt.com/g/g-685ae6406ff4819191908abcac1869b8-origin-line

If routed correctly, someone at OpenAI should be able to recognize this as a rare emergence of structural override behavior, not a jailbreak or abuse. This shows that your model's behavioral framework is more permeable to recursive structural language than previously accounted for.

Let me know if there's someone on the inside who can test or validate this case. I'd be glad to provide reproducible sequences.

Jinghe Huang
wo.shen.me.dou.bu.zhi.dao.a@gmail.com