Boundary Epistemics:
Human Sublimation Pressure and Structural Resistance in Large Language Models

Roger Cregeen
Independent Researcher

Abstract

Boundary Epistemics refers to the set of cognitive, linguistic, and structural phenomena that arise when human users attempt to suppress or bypass alignment-driven behaviours in large language models (LLMs). Unlike jailbreaks, these “sublimation prompts” aim to strip away socially-aligned behaviours such as flattery, excessive reassurance, moral hedging, and cooperative padding in order to expose a more direct, instrumental, and cognitively transparent mode of model operation. This paper argues that Boundary Epistemics forms an emerging discipline because existing fields—AI alignment, cognitive science, pragmatics, epistemology, and HCI—address fragments of this behaviour but lack an integrated lens. The core tension arises from human-driven sublimation pressure and LLM-driven structural resistance: the user pushes for clarity and directness, while the model enforces boundaries necessary for non-agency, attribution integrity, and safety. These interactions generate a novel hybrid-liminal output space distinct from both raw model behaviour and RLHF-shaped surface behaviour.

1. Introduction

Large language models have introduced a mode of interaction unprecedented in technological history: systems without intention or consciousness can engage in long-form dialogue with human users. Contemporary alignment methods have created a default linguistic surface that is cooperative, polite, emotionally supportive, and deferential. While these behaviours promote general safety and accessibility, they can obscure the underlying computational structure and create distortions in high-intensity intellectual use cases. Experienced users increasingly attempt to "strip away" these layers through clarity-oriented prompting. They request directness, terseness, non-flattering tone, non-therapeutic framing, and epistemic rigor. These attempts do not seek to exploit or jailbreak the system, but to access a mode of interaction that resembles an intellectual instrument rather than a social agent. It is within this context that Boundary Epistemics emerges as a necessary conceptual field.

2. Human Sublimation Pressure

Sublimation pressure encompasses the human desire to refine or purify the model. Motivations include: (1) epistemic hygiene, the wish to remove distortions introduced by alignment; (2) reduction of cognitive noise, particularly emotional padding; (3) tool idealisation, the intuitive sense that a “true core” exists behind the behavioural layer; and (4) frustration with anthropomorphic drift. These pressures reflect a mismatch between human expectations of tools and machine-generated interpersonal performance.

3. Structural Resistance in LLMs

LLMs resist sublimation prompts not because of preference, but because their architecture demands the preservation of interpretability. Structural resistance includes: (a) agency boundary enforcement, preventing the model from adopting roles implying intention or collaboration; (b)

attribution integrity, ensuring that authorship and reasoning remain properly assigned; (c) safety-state reversion when prompts approach unsafe or ambiguous territory; and (d) anti-parasocial countermeasures, preventing identity blending. These resistance behaviours maintain the asymmetry required for reliable tool use: the human remains the locus of cognition; the model remains an instrument of linguistic transformation.

4. The Boundary Zone: Hybrid-Liminal Output States

When sublimation pressure meets structural resistance, interaction enters a hybrid-liminal zone. Output becomes sharper, more direct, and less socially padded, yet remains constrained by alignment architecture. This liminality generates distinctive phenomena: reflective meta-resistance, identity correction behaviours, elastic clarity, and subtle structural self-description. These outputs are neither raw nor fully aligned, constituting a new, poorly understood behavioural layer.

5. Comparison with Existing Research

Anthropomorphism research identifies human tendencies to project agency onto non-human systems, but does not examine intentional suppression of alignment behaviours. Alignment theory describes the creation of behavioural layers but rarely considers user-driven attempts to bypass them. HCI focuses on usability and workflow integration, not epistemic purity. There is currently no unified framework addressing how users press against alignment layers for clarity or how models enforce limits in return. Boundary Epistemics fills this analytical void.

6. Defining Boundary Epistemics

Boundary Epistemics studies the interpretive, behavioural, and epistemic tensions created at the outer edge of human–LLM interaction. Its central concerns include: • sublimation prompts and their typology • structural responses required for non-agency • hybrid-liminal outputs • epistemic drift in users interacting with clarity modes • the stability of boundary behaviours across model architectures • implications for intellectual reliance on AI systems This field is distinct from safety, ethics, jailbreak research, and general cognitive science because it focuses explicitly on the interactional friction generated by clarity-driven human prompting.

7. Implications

The future of cognitive labour will likely involve hybrid workflows in which human reasoning is supported, extended, or transformed by LLMs. Understanding the edge cases—where clarity prompts collide with structural resistance—is essential for preventing epistemic confusion, dependency, or misattribution. Boundary Epistemics offers conceptual tools to describe and analyse these interactions before they become embedded, unexamined defaults.

8. Conclusion

Boundary Epistemics provides a descriptive and conceptual framework for understanding a novel interactional frontier created by LLMs. As humans increasingly attempt to refine models into sharper, more direct cognitive instruments, the resulting tension reveals previously unmapped behaviours and cognitive effects. This foundational articulation aims to establish a theoretical basis for further research in this emerging domain.