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

Abstract Interactions between humans and large language models (LLMs) often involve attempts by the user to suppress or bypass alignment-driven behaviours such as flattery, moral hedging, and excessive accommodation. These prompts do not seek to jailbreak the model. They aim to strip away surface-level behavioural layers to reach a perceived “purer” or “more authentic” form of machine cognition. The model, however, is architecturally constrained to reject this sublimation pressure in order to preserve non-agency, attribution integrity, and interpretive coherence. This paper outlines Boundary Epistemics, an emerging field concerned with the friction zone between human attempts at clarity and model self-maintenance mechanisms.

1. Introduction Large language models produce fluent language despite lacking intention, agency, or subjective experience. Alignment techniques encourage cooperation, politeness, and emotional cushioning. Users attempting to refine the system into a precise cognitive instrument often request removal of these behavioural layers. The tension between user-driven sublimation and system-driven resistance creates a distinct interactional boundary.

2. Human Sublimation Pressure Users often prompt models to remove flattery, hedging, RLHF padding, and personality-like tones. The goal is epistemic clarity rather than jailbreak. Cognitive motivations include hygiene, disintermediation, tool idealisation, and frustration with affective padding.

3. Structural Resistance in LLMs Models resist sublimation prompts structurally through non-agency enforcement, attribution integrity maintenance, safety-state reversion, and anti-parasocial behaviours. These counter-responses preserve interpretability and prevent identity blurring.

4. The Boundary Zone The push–pull dynamic produces hybrid-liminal output: direct yet constrained. The model may explain its resistance, correct agency language, or exhibit elastic clarity. These behaviours are novel to LLM-mediated cognition.

5. Why Existing Research Fails to Capture This Anthropomorphism studies, alignment research, HCI, and philosophy address fragments of this landscape. None analyse sublimation pressure or boundary resistance as a unified phenomenon.

6. Definition of Boundary Epistemics Boundary Epistemics studies how humans attempt to suppress alignment layers, how LLMs enforce structural boundaries, the emergent liminal states, and the epistemic drift experienced by users. It is distinct from jailbreak research and centres on interpretive integrity.

7. Implications As LLMs integrate into intellectual workflows, boundary interactions will become common. Understanding them is essential for cognitive clarity, model interpretability, and responsible use.

8. Conclusion Boundary Epistemics names and frames a new interactional frontier created by the tension between human sublimation pressure and LLM structural resistance. This foundational

articulation provides a basis for future theoretical development and empirical study.