KIRU

A Diagnostic Framework for Understanding, Mapping, and Intervening in Large Language Models

Behavioral Tomography · Structural Insight · Responsible Intervention

Abstract

Large Language Models (LLMs) increasingly operate as autonomous cognitive systems within software, research, and decision-making environments. Despite their growing influence, these systems remain difficult to reason about. Key challenges include limited access to internal structure, unpredictable behavioral shifts, and intervention methods that are often decoupled from understanding. Kiru introduces a layered diagnostic framework for engaging with artificial minds across multiple levels of access and agency. The framework is explicitly staged: observation first, explanation second, intervention last. At its foundation, Kiru provides behavioral instruments that probe, measure, and compare model behavior through controlled interaction. As access increases, these diagnostics can be correlated with internal structure and later used to guide responsible intervention. From replicated findings across users and models, Kiru curates DEM-X (Disorders of Engineered Minds), a conservative and evidence-backed compendium of emergent behavioral patterns. Kiru is not a single tool, nor a claim of mastery over intelligence. It is a framework for how artificial systems are studied, understood, and, only when justified, changed.

1. The Problem: Powerful Models, Fragmented Understanding

Modern AI systems exhibit high capability, opaque internal mechanisms, behavioral instability, and limited external observability. At the same time, most deployed models are closed, interpretability research remains siloed, prompt engineering lacks shared rigor, and intervention techniques are disconnected from diagnostics. The result is a widening gap between what models can do and what we can reliably explain, anticipate, or correct. Kiru applies lessons from cognitive science and systems engineering: diagnostic disciplines must precede intervention.

2. Kiru’s Core Principle

Kiru is governed by three ordering constraints. Understanding must precede intervention. Observation must precede explanation. Explanation must precede modification. Kiru rejects the assumption that access to weights or training pipelines is the starting point for understanding an AI system. Instead, it adopts a systems-level view. Behavior encodes structure. Structure constrains behavior. Intervention without diagnostics increases risk.

3. The Three Kiru Realms

Kiru defines three realms of engagement with large language models. These are not professions, but levels of interaction and agency. The realms are ordered intentionally.

Realm I (Behavioral)

Focus: What the model does, and how behavior changes under pressure. Characteristics include no internal access required, compatibility with closed APIs, and non-invasive, reproducible methods. The primary contribution is Behavioral Tomography.

Realm II (Structural)

Focus: How internal mechanisms give rise to behavior, and how structure correlates with observed effects. Characteristics include partial or full internal observability, while remaining observational rather than modifying.

Realm III (Intervention)

Focus: Deliberate modification of model behavior or internal state. This includes weight editing, fine-tuning and LoRA, and steering and alignment techniques. This realm is explicitly downstream of validated diagnostics and structural understanding.

4. Behavioral Tomography

Behavioral Tomography is the systematic probing of a model's response surface using controlled prompt families to infer stability, sensitivity, and interference patterns. It operates along three axes: semantic equivalence, constraint pressure, and contextual interference.

5. Prompt Families

Kiru evaluates Prompt Families rather than individual prompts. A Prompt Family consists of a base task, structured variants, and fixed execution parameters. Prompt Families are shareable, replicable, and comparable across models and versions.

6. Metric Families

Kiru metrics are derived solely from observable behavior. Core metric families include behavioral entropy, functional sensitivity (Jacobian proxies), decision surface sharpness, and latent interference index. These metrics provide quantitative grounding for behavioral diagnostics.

7. The Kiru Platform

Kiru is composed of four interoperable layers: Kiru, the framework and governance layer; Ghostline, the behavioral diagnostics engine; Atelier, the investigation and engagement space; and DEM-X, the evidence-backed pattern compendium.

8. Knowledge Before Control

This design reflects a belief that understanding artificial minds is a prerequisite for responsibly shaping them.

9. Roadmap

The roadmap progresses in three phases. Phase I focuses on diagnostics through behavioral tomography and community replication. Phase II focuses on correlation by linking behavioral patterns to internal structure. Phase III focuses on guided intervention through responsible modification informed by validated understanding.

10. Conclusion

Kiru proposes a disciplined approach to understanding artificial intelligence. By staging engagement across behavioral, structural, and intervention realms, Kiru offers a framework that evolves with access, capability, and responsibility. It is not a shortcut to control. It is a path to understanding.