HDT² Pilot v2 Supplementary Technical Statement

A Follow-On Research Program Derived from HDT² Pilot v1

1. Overview

HDT² Pilot v2 is the formal continuation of the work introduced in *HDT² Pilot v1: A Framework for Entropy-Band Calibration of LLM Reasoning Stability*.
Pilot v1 established that **token-level entropy in large language models exhibits stable, measurable geometric structure** and that this structure can be partitioned into **epistemic regimes**—regions of behavior reflecting whether a model is operating in a **stable**, **ambiguous**, or **underdetermined** state.

Pilot v2 expands the research scope from *measurement feasibility* to **epistemic regime characterization**, **model comparability**, and **control-theoretic applications**.

2. Summary of Findings from Pilot v1

Pilot v1 addressed the question:

Can the entropy field of an LLM be measured, calibrated, and aligned in a way that reveals meaningful and repeatable structure?

Key verified outcomes:

2.1 Entropy Has Internal Geometry

Token-level entropy across a reasoning sequence forms **non-random, repeatable patterns**. These patterns cluster into identifiable entropy bands that correspond to distinct behavioral states—later termed **epistemic regimes**.

2.2 Regime Stability Within a Single Model

For a model that passes alignment, regime boundaries are **consistent across prompts, topics, and reasoning depths**.
This indicates that uncertainty geometry is **intrinsic** to the model's learned representations and not an artifact of dataset selection.

2.3 Cross-Model Alignment is Conditional

A simple affine alignment procedure (UNSUPHALIGN) successfully calibrated entropy fields only for models with compatible uncertainty geometry.

Two 7B models failed alignment predictably, demonstrating that **entropy manifolds differ across architectures and training regimes**.

2.4 Correctness Is Not the Core Metric

Pilot v1 found that entropy-band transitions corresponded to increased error rates, but accuracy was used only as a **sanity check** to validate that regimes carried semantic meaning.
HDT^2 does *not* define hallucination detection; it defines **real-time epistemic state exposure**.

2.5 Measurement Pipeline Is Viable

A complete end-to-end measurement process—data collection, entropy computation, band calibration, alignment gating, and drift analysis—was successfully implemented and replicated.

These findings justify a deeper technical investigation in Pilot v2.

3. Objectives of Pilot v2

Pilot v2 shifts the research emphasis from feasibility to **characterization, comparability, and control**.
The primary objectives are:

3.1 Epistemic Regime Characterization

Define and analyze the structure of entropy-based regimes in LLMs:

- **Stable Region:** low-entropy, high-constraint continuation
- **Ambiguous Region:** moderate entropy with multiple plausible continuations
- **Underdetermined Region:** high-entropy, diffuse predictive distribution

Pilot v2 seeks to formally describe how these regimes emerge, how transitions occur across reasoning steps, and how they relate to internal model representations.

3.2 Cross-Model Regime Comparability

Study under what conditions two models exhibit **compatible uncertainty manifolds**:

- architectural similarity
- training objective similarity
- scale or depth
- fine-tuning or instruction-tuning differences

This includes investigating **non-affine** alignment strategies for cases where regime geometry is nonlinear.

3.3 Task-Dependent vs. Task-General Behavior

Assess whether regime boundaries remain stable:

- across task families
- across prompt types
- across deterministic vs. stochastic sampling
- across reasoning depths

Pilot v2 will explore whether entropy geometry is **task-agnostic** or modulated by semantic domain.

3.4 Dynamic Control Applications

Study how regime exposure can serve as a **control signal** for:

- abstention
- hedging
- clarification prompting
- tool invocation
- human escalation
- branch sampling or diversification

This moves HDT² from *measurement system* to *uncertainty-aware control scaffold*.

4. Pilot v2 Research Questions

Pilot v2 will explicitly investigate the following open questions generated by Pilot v1:

1. **Mechanistic Question:**
   What internal representational structures give rise to stable uncertainty regimes?
2. **Comparative Question:**
   Under what architectural/training conditions can entropy manifolds be aligned across models?
3. **Geometric Question:**
   Is entropy geometry inherently nonlinear, and does this explain alignment failures?
4. **Task Question:**
   Does entropy geometry vary systematically by task type or cognitive demand?
5. **Policy Question:**
   What control policies should be attached to each epistemic regime?
6. **Stability Question:**
   How robust are regime boundaries to sampling variance, prompt perturbations, and tool-augmented reasoning?
7. **System Integration Question:**
   How effectively can regime exposure support risk-aware routing in real systems?

5. Technical Scope and Methodological Notes

Pilot v2 retains the core measurement framework of Pilot v1 but extends it with:

- **finer-grained entropy drift metrics** across multi-step reasoning
- **non-affine alignment experiments** (kernel, spline, manifold-learning)
- **per-task regime mapping**
- **cross-model geometry comparisons**
- **sampling-sensitivity analyses**
- **regime-based control-policy prototyping**

No claims of universality, causal mechanism, or hallucination classification are made.
Pilot v2 remains within the bounded intent of exploratory, falsifiable diagnostic research.

6. Collaboration Invitation

HDT² Pilot v2 is designed as an **open research program**, not a closed methodology.
Researchers working in:

- uncertainty quantification,
- LLM safety and reliability,
- cognitive modeling,
- control theory,
- representation geometry,
- or evaluation frameworks

are invited to contribute.

Areas especially suited for collaboration include:

- mechanistic probing of entropy-manifold structure
- nonlinear alignment method development
- cross-architecture Δ-field comparison
- empirical validation on diverse model families
- policy design for abstention, hedging, routing, or tool use

Pilot v2 aims to accelerate shared understanding of how LLMs signal their own epistemic boundaries and how systems can respond adaptively to those signals.

7. Closing Statement

HDT² Pilot v1 established that entropy geometry is real, measurable, and structured.
HDT² Pilot v2 seeks to understand **what that structure means**, **how it varies across models and tasks**, and **how it can be used to build safer, adaptive AI systems**.

This work remains intentionally open and exploratory.
It is not a completed methodology but a growing research scaffold.

Pilot v2 is an invitation to join that exploration.