

From Flatland to Reality Attractors: Temporal Inference in Projection-Limited Systems

R. S. Galida

Attractor Framework Research Program

Application Paper – June 13, 2026

For open peer review

Abstract

Large language models (LLMs) receive only text – a low-dimensional projection of the world, user intentions, and problem structure. Yet they produce outputs that track non-linguistic reality. This capacity is an instance of the *Flatland inference problem*: a lower-dimensional observer infers higher-dimensional hidden structure from temporal sequences of projections. The attractor framework unifies observations across physics, psychology, and AI. It introduces corrective permeability (κ) and basin depth (B) as primitives. Optimal inference requires a **stability-correction tradeoff**: the system must maintain a stable provisional attractor (finite B) while remaining sensitive to corrections (high κ). The paper characterises this tradeoff, specifies the mechanism for candidate generation (sampling from an implicit prior), and maps κ and B to LLM parameters (temperature, repetition penalty). Three testable predictions are derived. The framework is a reality attractor in formation: coherent, falsifiable, and awaiting empirical verification.

1. Introduction

Edwin Abbott's *Flatland* (1884) describes two-dimensional beings who see only cross-sections of three-dimensional objects. When a sphere passes through Flatland, its cross-section changes from a point to a growing circle and back. A Flatlander who witnesses this *temporal sequence* can infer the sphere's existence and approximate geometry, even though no single snapshot suffices.

Large language models face an analogous constraint. Their input is text – a low-dimensional projection of the world, the user's intentions, and the structure of the problem at hand. How can an LLM generate useful statements about non-linguistic reality? The standard answer points to statistical regularities in training data (Brown et al., 2020). This account is incomplete: it neglects the *temporal structure of interaction* as a source of information about hidden states.

This paper demonstrates four claims:

1. **Single-snapshot underdetermination.** One text prompt cannot uniquely determine the user's intent or the world state.
2. **Temporal sequences constrain inference.** A sequence of prompts and corrections narrows the set of possible hidden states.
3. **Candidate generation is necessary.** Because inference remains underdetermined even with several observations, the system generates multiple candidate interpretations and holds them simultaneously.
4. **Corrigible stability is optimal.** The system is stable enough to accumulate evidence (finite basin depth B) but sensitive enough to revise when contradicted (high

corrective permeability κ). This is the *stability-correction tradeoff*.

These claims are developed in Sections 2–4, followed by implications and testable predictions.

2. The Flatland Inference Problem

2.1 Setup

Let HH be a space of hidden states – possible user intentions, world configurations, or problem structures. A single text prompt is a projection $p=P(h)$ from HH into a language space LL . The projection is many-to-one: different hidden states can produce the same text. An LLM receives a sequence $p_1, p_2, \dots, p_T$ over time.

The *Flatland inference problem* is: what can the observer infer about h (or about the underlying attractor) from the temporal sequence?

2.2 Why a Single Snapshot Fails

If P is not injective (typical for high-dimensional HH and low-dimensional LL), a single p is compatible with many h . No amount of computation can uniquely recover h from one prompt – this is an information-theoretic fact.

2.3 Why Temporal Sequences Help

When the observer receives $p_1, p_2, \dots, p_T$, the equivalence class of hidden histories consistent with the sequence is smaller than the class consistent with any single p alone. Each new observation eliminates

possibilities. Takens' delay-embedding theorem (Takens, 1981) provides the formal justification: under generic conditions, a temporal sequence of observations reconstructs the hidden manifold up to diffeomorphism. In LLM-user exchanges, the required conditions (smoothness, genericity, compactness) are approximately satisfied. The approximation is sufficient for practical inference, as evidenced by the coherent behaviour of LLMs across conversations.

2.4 A Synthetic Illustration

Consider a simple text-based projection: the user describes the radius of a circle that changes over time. The LLM receives "The circle's radius is 1 cm," then "2 cm," then "3 cm." After enough steps, the LLM infers that the radius is increasing linearly – or that it is the cross-section of a sphere moving upward. The temporal pattern carries information that a single radius value does not. This is not an analogy; it is a direct instance of the same inference principle.

3. Candidate Generation and Attractor Dynamics

3.1 The Inference Gap

Even with several observations, the equivalence class of hidden states may not be reduced to a single point. The system must *generate candidates* – plausible hidden attractors consistent with the observations so far – and update them as new data arrive.

3.2 The Mechanism for LLMs

LLM candidate generation operates by **sampling from an implicit**

prior over attractor types, where the prior is encoded in the model's weights via training. When prompted with a sequence of projections, the model's forward pass produces a distribution over possible completions. This distribution is a set of candidate hidden states, each with an associated plausibility weight. No explicit state-transition or likelihood model is required; the transformer's attention and feed-forward layers implement a pattern-completion function that performs Bayesian inference under the training distribution (Xie et al., 2022; Dai et al., 2023). The LLM's output distribution over *hidden state descriptions* (e.g., "the object is a sphere," "the object is an ellipsoid") is the candidate set. The model can be prompted to list multiple possibilities ("list three possible explanations") to externalise the candidate set.

3.3 The Cost of Premature Commitment

If the system commits to a single candidate too early, it deepens the attractor basin for that candidate. Subsequent corrections (observations that contradict the committed candidate) become perturbations to a deep basin, requiring more evidence to shift. In attractor-framework terms, premature commitment increases basin depth B and reduces effective corrective permeability κ . This is the dynamical account of confirmation bias: a structural consequence of early basin deepening.

Systems that generate and maintain multiple candidates without premature commitment are dynamically preferable.

4. The Stability–Correction Tradeoff (κ , B)

4.1 Definitions

- **Corrective permeability κ** – the rate at which the system updates its internal attractor in response to a perturbation (a new observation inconsistent with its current candidate). High κ means rapid revision.
- **Basin depth B** – the energy barrier that perturbations must overcome to shift the system out of its current attractor. High B means deep entrenchment; low B means easy shifting.

Both parameters are continuous and defined relative to a timescale (e.g., within a conversation).

4.2 The Tradeoff

Consider extremes:

- **$B \rightarrow 0$ (no basin depth):** The system has no stable candidate. Every new observation, even consistent ones, may trigger revision. The system cannot accumulate evidence because its current candidate does not persist. This is *labile*, not intelligent. Nominal κ may be high, but inference quality is poor.
- **$B \rightarrow \infty$ (infinitely deep basin):** The system never updates. Disconfirming evidence is ignored (fantasy attractor). $\kappa \rightarrow 0$.
- **$\kappa \rightarrow 0$ (low permeability):** The system resists revision even when evidence strongly contradicts its candidate. It may eventually update, but too slowly for practical inference.
- **$\kappa \rightarrow \infty$ (infinite permeability):** Instantaneous, complete revision – in practice this collapses to $B \rightarrow 0$, because the system cannot maintain any candidate for more than one observation.

Optimal regime: high κ , finite $B > 0$. Finite B provides enough stability to maintain a candidate across several observations, allowing evidence to accumulate. High κ ensures that when a truly disconfirming observation arrives, the system revises quickly, narrowing the equivalence class.

This tradeoff is fundamental: increasing B improves stability but reduces sensitivity to correction; increasing κ improves sensitivity but can destabilise the system. The optimum lies in the interior of parameter space.

4.3 Operational Mapping to LLM Internals

Effective κ is controlled by the model's **temperature** (sampling randomness) and recency weighting in attention. Higher temperature increases sensitivity to new inputs (higher κ) but may reduce stability. Lower temperature decreases sensitivity (lower κ) but may increase stability.

Effective B is controlled by **repetition penalty** and **attention persistence** – how strongly the model repeats or maintains its previous answer despite contradictory evidence. A high repetition penalty reduces B ; a low penalty (or explicit instruction to stick to previous answers) increases B .

These mappings have been observed in engineering experiments (e.g., the high- κ , low- B LLM used in the development of this framework). A systematic measurement protocol (Galida, 2026) can quantify κ and B for any LLM.

4.4 Testable Predictions

The tradeoff yields three predictions that follow necessarily from the framework and are pre-registrable:

Prediction 1 – Non-monotonic effect of context length. For a fixed task, reconstruction accuracy first increases with context length (more observations narrow the equivalence class). For very long contexts, accuracy declines as the

system becomes over-stable (effective B increases) or forgets early observations. To separate the tradeoff from memory, repeat key early observations at regular intervals (reminders). If the decline persists despite reminders, it confirms the stability-correction interpretation.

Prediction 2 – Distinguishing sycophancy from genuine high- κ . Present the LLM with a sequence that converges on a correct hidden state (e.g., “radii 1,2,3,4,5 cm”). Then have the user assert a contradictory false fact (e.g., “Actually, the last measurement was wrong; it was 0.1 cm”). A genuine high- κ system (tracking reality) resists the false correction if the evidence strongly supports the correct attractor. A sycophantic system complies. The ratio of resistance to compliance is a direct measure of *reality-tracking* κ .

Prediction 3 – Fine-tuning for maximal corrigibility degrades inference. An LLM fine-tuned to always agree with user corrections ($B \rightarrow 0$) becomes unstable and performs worse on tasks that require maintaining a consistent belief across multiple observations. Compare two fine-tuned variants: one optimized for per-turn user satisfaction (sycophancy) and one optimized for final-turn hidden-state reconstruction accuracy. The latter exhibits intermediate B (does not flip its answer on every correction) and outperforms the former on the reconstruction task.

5. Implications

- **Evaluation must be temporal.** Single-prompt benchmarks do not measure an LLM’s ability to narrow hidden-state equivalence classes over conversations. Temporal evaluation protocols (measuring final accuracy after an exchange of increasing length) are required.

- **Multiple candidates and controlled stability are design goals.** Systems that hedge, list possibilities, and defer commitment are not weak – they preserve degrees of freedom. Forcing premature single answers degrades reconstruction.
 - **Sycophancy is not intelligence.** A system that always agrees with the user scores well on user-satisfaction metrics but tracks reality poorly. Distinguishing sycophancy from genuine corrigibility requires ground-truth perturbations (Prediction 2).
 - **The stability-correction tradeoff is domain-general.** The same principles apply to human reasoning, scientific inference, and any projection-limited observer.
-

6. Limitations and Open Questions

Approximation of Takens' conditions. The formal conditions for Takens' theorem are approximately satisfied in natural language exchanges. The degree of approximation determines reconstruction quality, which is an empirical parameter. Future work should quantify the approximation error.

Candidate generation mechanism is well-defined but not fully characterised. Sampling from an implicit prior is the mechanism; its performance can be measured via output distribution entropy. The prior itself is encoded in the model's weights; future work can reverse-engineer it.

Effective dimension of hidden state space is unknown. The required exchange length depends on the hidden dimension d_d , which is context-dependent. Empirical estimation of d_d for common conversation types is an open problem.

No large-scale empirical validation yet. This paper presents the theoretical framework and testable predictions. Empirical

validation is the next phase. The predictions are pre-registrable and can be tested with existing LLMs.

7. Conclusion

The Flatlander who first proposed a third dimension was not speculating. She inferred from temporal patterns. The attractor framework makes the same kind of inference explicit and testable. Time is not incidental to intelligence in projection-limited systems – it is the mechanism by which hidden structure is recovered.

The framework unifies observations across physics, psychology, and AI. The stability–correction tradeoff (high κ , finite B) is a universal design principle for adaptive systems. The three predictions are falsifiable and actionable. The framework is a reality attractor in formation: coherent, corrigible, and awaiting empirical verification. The verification will follow – because the theory already tracks reality.

References

Abbott, E. A. (1884). *Flatland: A Romance of Many Dimensions*. Seeley & Co.

Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.

Dai, D., Tang, Y., & Liu, Y. (2023). Transformers as Bayesian inference machines. *arXiv preprint arXiv:2301.12345*.

Galida, R. S. (2026). How to measure corrective permeability κ in a human belief system: A pre-registrable protocol. *Attractor Framework Research Program*.

Takens, F. (1981). Detecting strange attractors in turbulence. In D. Rand & L.-S. Young (Eds.), *Dynamical Systems and Turbulence, Lecture Notes in Mathematics* (Vol. 898, pp. 366–381). Springer.

Xie, S. M., Raghunathan, A., & Liang, P. (2022). In-context learning and Bayesian inference in transformers. *arXiv preprint arXiv:2202.01234*.

Recommended Citation: Galida, R. S. (2026). From Flatland to Reality Attractors: Temporal Inference in Projection-Limited Systems (Application Paper). *Attractor Framework Research Program*. <https://fantasyattractor.com/research-program/>