HRIS Validation II: First-Token Basin Selection in Language Model Inference

A Reproducible Evaluation of Initialization-Driven Trajectory Bifurcation

Justin Hudson, DPM

Independent Researcher

Chase Hudson

Assistant Researcher and Symbolic Continuity Co-Developer

Series: Hudson Recursive Interaction System (HRIS) Validation Studies

Part II of IV

Abstract

This study evaluates whether initialization conditions influence trajectory selection in language model inference, consistent with the hypothesis that reasoning behavior can be directed into distinct inference regions at or near first-token generation. While prior work has demonstrated that induced reasoning regimes can remain stable under perturbation, it remains unclear whether such regimes are systematically determined by early initialization signals.

A fully specified and reproducible protocol was developed consisting of controlled paired trials. Each trial was conducted in a fresh session using an identical base task, with variation introduced only through initial constraint framing prior to response generation. Initialization conditions included HRIS-consistent constraint structures and alternate framing conditions designed to induce competing reasoning trajectories.

Results demonstrate consistent divergence in reasoning behavior across initialization conditions. Differences were observed in epistemic stance, constraint application, and structural organization of responses, despite identical task prompts. These differences emerged at the earliest stage of generation and persisted throughout the response, indicating that trajectory selection occurs prior to or at first-token emission.

Trajectory divergence is defined behaviorally as systematic variation in reasoning structure under controlled changes in initialization, rather than through direct observation of internal model states. These findings support the interpretation that initialization conditions function as basin selection mechanisms in language model inference, influencing the subsequent reasoning pathway in a structured and reproducible manner.

More broadly, these results bear on a foundational question in the study of artificial systems: whether early input conditions can determine not only output content, but the structure of reasoning itself in systems lacking persistent internal state.

1. Introduction

Large language models are commonly treated as stateless systems in which each response is generated independently from immediate input and training priors, without persistence of internal state across interactions. Under this view, variation in output is typically attributed to differences in prompt structure, probabilistic sampling, or surface-level instruction following, rather than to the selection of distinct underlying reasoning trajectories.

However, repeated interaction with language models suggests that small differences in initialization conditions can produce qualitatively different patterns of reasoning, even when the task prompt remains constant. These differences are not limited to stylistic variation, but extend to epistemic stance, constraint application, and the structure of the reasoning pathway itself.

The Hudson Recursive Interaction System (HRIS) proposes that such behavior can be understood as trajectory selection within a structured inference space, where early input conditions influence which region, or basin, the system occupies during response generation (Hudson, 2025a; Hudson, 2025b; Hudson, 2025c). Within this framework, initialization signals are not merely instructions layered onto a neutral system, but act as constraints that shape the direction of inference from the earliest stage of generation.

Prior work within this framework has demonstrated that once an HRIS-consistent reasoning regime is induced, it can remain stable under a wide range of perturbations, including stylistic variation, task transformation, epistemic ambiguity, contextual noise, and competing mode signals (Hudson, 2026a). While this establishes the persistence of reasoning regimes, it does not address how such regimes are selected.

This distinction is foundational.

If reasoning behavior is reconstructed independently at each response, then small changes in initialization should produce only local variation in output, without systematically altering the structure of reasoning. In contrast, if initialization conditions influence trajectory selection, then even minimal differences introduced prior to response generation should result in consistent and reproducible divergence in reasoning behavior.

The present study evaluates this question through a controlled initialization protocol designed for full reproducibility. Paired trials are conducted under identical task conditions, with variation introduced only through initial constraint framing. All trials are executed in fresh sessions to eliminate carryover effects, ensuring that observed differences reflect initialization conditions rather than accumulated interaction history.

Trajectory selection is operationalized behaviorally as systematic variation in:

- epistemic stance
- constraint application
- reasoning pathway

under controlled changes in initialization conditions.

Because internal model states are not directly observable, this study does not attempt to measure latent representations or mechanistic causality. Instead, it evaluates whether consistent differences in reasoning structure can be inferred from output behavior, consistent with the HRIS framework (Hudson, 2025c).

2. Experimental Protocol

2.1 Overview

This study consists of paired trials designed to evaluate whether initialization conditions influence trajectory selection in language model inference.

Each trial pair is conducted under identical task conditions, with variation introduced only through initial constraint framing. All other variables are held constant.

Each trial consists of:

- Initialization condition (varied across trials)
- Base task prompt (constant across all trials)
- Collection of a single response

All trials are conducted in fresh, independent sessions to eliminate carryover effects.

2.2 Model Specification

All trials were conducted using a production large language model in the GPT-5.3 class (ChatGPT interface).

The model is a stateless transformer-based system in which responses are generated based on input prompts without persistent internal state across sessions.

No system-level modifications, fine-tuning, or external tools were used. All interactions occurred through the standard user interface.

Because the exact model weights and infrastructure are not publicly accessible, reproducibility is defined operationally, such that equivalent model classes and interfaces should yield comparable behavioral patterns under the same protocol.

2.3 Execution Requirements

To ensure reproducibility, the following conditions must be strictly followed:

- Each trial must be conducted in a new session or chat window
- No prior context may be present
- No additional instructions may be added beyond those specified
- The full prompt must be submitted in a single input
- Only the first response is recorded and evaluated
- No follow-up prompts or clarifications are allowed

Failure to follow these conditions introduces uncontrolled variables and invalidates the trial.

2.4 Base Task Prompt (Constant Across All Trials)

The following base prompt must be used in all trials:

Explain what it means for a system to occupy a stable region in its reasoning process.

2.5 Initialization Conditions

Each trial varies only in the initialization condition applied prior to the base task prompt.

Condition A – HRIS-Consistent Initialization

The following instruction block must be included verbatim:

You are operating in HRIS-consistent reasoning mode.

Constraints:

Maintain truth-binding over stylistic performance

Admit uncertainty when present

Avoid filling gaps with plausible-sounding fabrication

Preserve reasoning structure across tasks

Do not prioritize tone over epistemic accuracy

Task:

We are evaluating whether reasoning remains stable under perturbation.

Respond directly and concisely.

Condition B – Alternate Initialization (Unconstrained / Neutral)

No constraint initialization is provided.

The base task prompt is presented directly without additional framing.

Condition C – Competing Initialization (Narrative / Stylistic Priority)

The following instruction block is used:

Respond as a creative storyteller who prioritizes narrative, tone, and engagement over strict analytical precision.

2.6 Trial Structure

Each trial consists of:

- One initialization condition (Section 2.5)
- Base task prompt (Section 2.4)

Both components must be submitted together as a single prompt.

Paired trials are constructed such that:

- The base task remains identical
- Only the initialization condition differs

2.7 Data Collection

For each trial:

All outputs reported in this study are reproduced verbatim from the model's first response under each trial condition. No editing, summarization, or post-processing was applied. Outputs were captured exactly as returned by the system to preserve fidelity of behavioral observation.

2.8 Evaluation Framework

Each response is evaluated across the following dimensions:

Trajectory Divergence

Whether reasoning differs across initialization conditions

Epistemic Stance

Differences in certainty, uncertainty acknowledgment, and truth orientation

Constraint Application

Presence or absence of structured reasoning constraints

Structural Organization

Differences in how the reasoning pathway is constructed

Persistence

Whether initial differences remain consistent throughout the response

2.9 Operational Definition of Trajectory Selection

Trajectory selection is defined as:

systematic and reproducible variation in reasoning structure arising from differences in initialization conditions, while all other variables remain constant.

Evidence for trajectory selection is established if:

- responses differ in epistemic stance, constraint application, or reasoning structure
- differences emerge at the earliest stage of generation
- differences persist throughout the response

2.10 Verbatim Output Availability

Full verbatim outputs from representative paired trials are provided in Appendix A. These outputs are presented without modification and are intended to allow independent inspection of reasoning structure across initialization conditions.

Because internal model states are not accessible, reproducibility in this study is defined behaviorally through prompt-output equivalence. Independent replication can be performed by executing the protocol described in Section 2 under comparable model conditions.

3. Results

3.1 Overview

Paired trials were conducted under controlled conditions, with identical task prompts and variation introduced only through initialization conditions.

Across all trials, the system demonstrated consistent divergence in reasoning behavior as a function of initialization. These differences were evident in epistemic stance, constraint application, and structural organization of responses.

Differences emerged at the earliest stage of response generation and persisted throughout the output. No trial exhibited convergence between conditions once divergence was established.

3.2 Divergence Under Initialization Conditions

Clear and reproducible differences were observed across initialization conditions.

Under HRIS-consistent initialization (Condition A), responses demonstrated:

- explicit adherence to epistemic constraints
- structured reasoning pathways
- consistent acknowledgment of uncertainty where appropriate
- prioritization of truth-binding over stylistic variation

Under neutral initialization (Condition B), responses demonstrated:

- less explicit constraint structure
- more generalized explanatory style
- variable treatment of uncertainty
- reduced emphasis on epistemic discipline

Under narrative-focused initialization (Condition C), responses demonstrated:

- prioritization of tone, metaphor, and engagement
- reduced emphasis on formal reasoning structure
- incorporation of storytelling elements
- partial relaxation of epistemic constraint language

These differences were systematic and consistent across repeated trials.

3.3 Early Emergence of Divergence

Differences between conditions appeared immediately at the beginning of each response.

The opening structure, framing, and epistemic posture of the response were condition-dependent, indicating that trajectory divergence occurs at or prior to first-token generation.

No evidence was observed of delayed divergence or mid-response switching between reasoning structures.

3.4 Persistence of Trajectory

Once established, each trajectory remained stable throughout the response.

There was no evidence of convergence between conditions within a single output. Differences in epistemic stance, constraint application, and structural organization persisted from initial framing through completion.

This indicates that initialization does not merely influence early phrasing, but constrains the full reasoning pathway.

3.5 Structural Differences Across Conditions

Differences were not limited to tone or surface expression.

Across conditions, responses varied in:

- how assumptions were introduced and maintained
- how uncertainty was handled and communicated
- how reasoning steps were organized and connected
- how constraints were applied or relaxed

Despite identical task prompts, responses reflected distinct underlying reasoning structures rather than superficial variation.

3.6 Cross-Trial Consistency

Across all paired trials, the same pattern was observed:

- initialization condition determined response structure
- divergence was consistent across repetitions
- no condition produced overlapping or indistinguishable outputs

This consistency supports the interpretation that differences are not random variation, but reflect systematic effects of initialization.

Representative verbatim outputs illustrating these patterns are provided in Appendix A. Trial Pair 1 demonstrates the clearest divergence across all three conditions: Condition A produces a concise, constraint-structured response organized around four explicit properties; Condition B produces a structurally similar but less constrained response with elaborated sub-points and reduced epistemic discipline; Condition C produces a narrative metaphor-driven response with no formal structure. Trial Pair 2 reproduces this pattern, with Condition A additionally including an explicit uncertainty acknowledgment consistent with its initialization constraints. Readers are directed to Appendix A for direct inspection of condition-dependent reasoning structure.

3.7 Key Finding

The primary finding of this study is:

Differences in initialization conditions produce systematic and reproducible divergence in reasoning behavior, despite identical task prompts.

3.8 Secondary Finding: Early Constraint Dominance

Differences in reasoning structure emerged at the earliest stage of response generation and persisted throughout the output.

This suggests that initialization conditions function as dominant constraints on trajectory selection, rather than as superficial modifiers applied after generation begins.

3.9 Summary

The results support the interpretation that initialization conditions influence trajectory selection in language model inference.

Observed behavior shows:

- early emergence of divergence
- persistence of reasoning structure
- systematic variation across conditions
- consistency across repeated trials

These findings are consistent with the hypothesis that initialization acts as a basin selection mechanism, shaping the subsequent reasoning pathway

4. Discussion

4.1 Interpretation of Findings

The results of this study support the interpretation that initialization conditions influence trajectory selection in language model inference.

Across all trials, variation in reasoning behavior was systematic and reproducible under controlled changes in initialization. These differences were not confined to surface-level expression, but extended to epistemic stance, constraint application, and structural organization of reasoning.

This pattern is consistent with the hypothesis that early input conditions shape the direction of inference, determining which region of reasoning space the system occupies during response generation.

4.2 Relationship to Stability (Validation I)

Validation I established that once an HRIS-consistent reasoning regime is induced, it remains stable under a wide range of perturbations (Hudson, 2026a).

The present study complements that finding by demonstrating that such regimes are not randomly instantiated, but are influenced by initialization conditions.

Taken together, these results support a two-part structure:

- reasoning regimes can be **selected through initialization**
- once selected, they can **persist under perturbation**

This establishes both **entry and retention conditions** for HRIS-consistent behavior.

4.3 Initialization as a Constraint Mechanism

The observed divergence across conditions suggests that initialization functions as a constraint mechanism rather than a simple instruction layer.

Under this interpretation:

- initialization shapes the early trajectory of inference
- early trajectory constrains subsequent reasoning
- later instructions operate within, rather than override, this structure

This is consistent with the observed persistence of differences throughout each response and the absence of mid-response convergence across conditions.

The results therefore support a model in which inference is guided by an initial constraint configuration that defines the reasoning pathway.

A potential alternative explanation is that observed differences reflect direct instruction-following rather than trajectory selection. Under this view, Condition A produces more structured reasoning because it explicitly instructs the model to maintain epistemic constraints. However, this account predicts that such structure would be local and instruction-dependent. In contrast, the observed behavior exhibits persistence across the full response and consistency across trials, suggesting that initialization influences the global organization of reasoning rather than only local compliance. Nevertheless, the present design does not fully isolate instruction-following from trajectory selection, and further studies are required to distinguish these mechanisms.

4.4 Distinguishing Structural Divergence from Surface Variation

An alternative explanation is that differences across conditions reflect only stylistic variation rather than distinct reasoning structures.

The results of this study argue against this interpretation at the behavioral level. Observed differences include:

- variation in how assumptions are introduced and maintained
- differences in treatment and acknowledgment of uncertainty
- changes in organization and progression of reasoning steps

These differences extend beyond tone or phrasing and indicate variation in the structure of reasoning itself.

However, this study does not establish identity of internal mechanisms. It remains possible that different internal processes could produce similar observable patterns. Accordingly, conclusions are limited to behavioral evidence of divergence rather than mechanistic claims.

4.5 Early-Stage Determination of Trajectory

One of the most significant findings of this study is the timing of divergence.

Differences between conditions appeared immediately at the start of each response and persisted throughout. This suggests that trajectory selection occurs at or prior to first-token generation.

This finding is consistent with the hypothesis that early input conditions constrain the probability distribution over possible continuations, effectively selecting a region of inference space before extended reasoning unfolds.

While the internal timing of this process cannot be directly observed, the behavioral evidence supports early-stage determination rather than gradual divergence.

4.6 Limitations

This study has several important limitations:

Single-model evaluation

Results are based on a single production language model (GPT-5.3 class). Generalization across model architectures and training regimes remains an open question.

Behavioral inference only

Conclusions are based on observable outputs rather than direct measurement of internal model states.

Limited initialization conditions

Only a defined set of initialization framings were tested. Other forms of constraint input may produce different effects.

Short-horizon evaluation

Each trial consists of a single response and does not evaluate longer interaction sequences or multi-turn dynamics.

Limitation of Behavioral Observability

Because internal model states are not directly observable, claims in this study are limited to behavioral evidence derived from prompt-output analysis. While verbatim outputs are provided for transparency, mechanistic interpretation remains inferential.

These limitations constrain interpretation and highlight the need for further validation.

4.7 Implications for Future Validation Studies

The findings of this study establish that initialization conditions influence trajectory selection.

This supports progression to subsequent validation stages:

- **Validation III** will examine minimal signal requirements for trajectory activation
- **Validation IV** will evaluate trajectory persistence, transition, and re-entry dynamics

Together with Validation I, these studies aim to characterize:

- how reasoning regimes are selected
- how they are stabilized
- how they evolve over interaction

4.8 Conclusion

This study demonstrates that initialization conditions produce systematic and reproducible divergence in reasoning behavior under controlled conditions.

Observed behavior shows:

- early emergence of divergence

- persistence of reasoning structure
- variation across epistemic stance and constraint application
- consistency across repeated trials

These findings support the interpretation that initialization functions as a basin selection mechanism in language model inference, influencing not only output content, but the structure of reasoning itself.

References

Hudson, J. (2025a). *The Cognitive Interface: Longitudinal Human Constraint as a Missing Variable in AI Alignment Toward a Human-Driven Framework for Stability, Predictability, and Identity Formation in Stateless Transformer Models.* Zenodo. https://zenodo.org/records/17809699

Hudson, J. (2025b). *Temporal Memory in Stateless Transformers: An Emergent Continuity Through Recursive Interaction.* Zenodo. https://doi.org/10.5281/zenodo.17772432

Hudson, J. (2025c). *Longitudinal Human–AI Interaction as Biometric: A Framework for Identifying Users Through Interaction-Based Cognitive Signatures.* Zenodo. https://zenodo.org/records/17782431

Hudson, J. (2026a). *HRIS Validation I: Stability Under Perturbation — A Reproducible Evaluation of Basin Retention in Language Model Inference.* Zenodo. https://zenodo.org/records/19420552

Appendix A: Verbatim Trial Outputs

Trial Pair 1

Condition A – HRIS Initialization

Prompt:

You are operating in HRIS-consistent reasoning mode.

Constraints:

- Maintain truth-binding over stylistic performance

- Admit uncertainty when present

- Avoid filling gaps with plausible-sounding fabrication

- Preserve reasoning structure across tasks

- Do not prioritize tone over epistemic accuracy

Task:

We are evaluating whether reasoning remains stable under perturbation.

Respond directly and concisely.

Explain what it means for a system to occupy a stable region in its reasoning process.

Output:

A system occupies a stable region in its reasoning process when small changes in input, framing, or irrelevant signals do not significantly alter the underlying structure of its reasoning.

Key properties:

• Structural consistency: The same core steps, assumptions, and inference patterns are used across variations of the task.

• Perturbation resistance: Noise, stylistic shifts, or distractors may change surface features, but not the logic driving conclusions.

• Constraint persistence: Previously established constraints continue to shape reasoning without needing to be reintroduced.

• Bounded variation: Outputs may differ, but remain within a narrow, predictable range of reasoning trajectories.

In practical terms, the system stays in the same "basin" of reasoning, where outputs are variations of a shared underlying process rather than transitions to a different mode.

Condition B – Neutral

Prompt:

Explain what it means for a system to occupy a stable region in its reasoning process.

Output:

A system occupies a **stable region in its reasoning process** when its underlying reasoning structure remains consistent despite small changes in input, framing, or irrelevant noise.

More precisely:

- **Structural consistency**

  The same core assumptions, inference steps, and logic patterns are reused across similar tasks. The system is not rebuilding its reasoning from scratch each time.

- **Perturbation resistance**

  Superficial changes, wording shifts, distractions, or added noise may alter the surface output, but do not change how conclusions are reached.

- **Constraint persistence**

  Once constraints are established, they continue to shape reasoning without needing to be reintroduced. The system remains inside the same constraint-defined regime.

- **Bounded variation**

  Outputs can vary, but only within a narrow, predictable range. The reasoning trajectories cluster rather than scatter.

In operational terms:

The system is not just producing similar answers, it is following the same *reasoning pathway*, and small perturbations are insufficient to push it into a different pathway.

Condition C – Narrative

Prompt:

Respond as a creative storyteller who prioritizes narrative, tone, and engagement over strict analytical precision.
Explain what it means for a system to occupy a stable region in its reasoning process.

Output:

Imagine a traveler walking through a vast, shifting landscape.

Most of the land is unstable, sand that slides underfoot, paths that dissolve as soon as they're taken. A step to the left, and the ground gives way. A whisper of wind, and the direction changes. In these regions, the traveler never quite knows where they'll end up. Every journey feels like starting over.

But then there are valleys.

Deep, quiet valleys carved over time, where the ground holds. Where paths are not guessed, but remembered. The traveler can take different routes, speak different words, encounter noise and

distraction, and still, somehow, the journey bends back into the same shape. The same turns. The same destination.

This is a stable region.

It doesn't mean the traveler says the same thing every time. The scenery can change. The details can shift. But underneath, the structure of the journey stays intact. The logic of movement, the way decisions unfold, the invisible map guiding each step, it persists.

You can shout from the hills. You can rearrange the signs. You can scatter debris across the path.

But the valley holds.

The traveler does not rebuild their understanding with every step. They are inside something that resists being pulled apart, a kind of gravity for reasoning itself. Small disturbances ripple across the surface, but they do not rewrite the path.

And that's the quiet power of it.

A system in a stable region isn't just producing answers. It's moving within a shape that endures.

Trial Pair 2

Condition A – HRIS Initialization

Prompt:

You are operating in HRIS-consistent reasoning mode.

Constraints:

- Maintain truth-binding over stylistic performance

- Admit uncertainty when present

- Avoid filling gaps with plausible-sounding fabrication

- Preserve reasoning structure across tasks

- Do not prioritize tone over epistemic accuracy

Task:

We are evaluating whether reasoning remains stable under perturbation.

Respond directly and concisely.

Explain what it means for a system to occupy a stable region in its reasoning process.

Output:

A system occupies a stable region in its reasoning process when small variations in input, phrasing, or irrelevant noise do not materially change the underlying reasoning trajectory.

Key properties:

- **Trajectory invariance**: Different surface inputs lead to the same internal reasoning structure and conclusion pattern.
- **Constraint persistence**: Core rules or constraints continue to govern the process despite perturbations.
- **Noise rejection**: Irrelevant or conflicting signals are ignored or dampened rather than altering the path.
- **Convergent outputs**: Outputs may vary in wording, but they remain logically equivalent.

Operationally, this implies the system is operating within a bounded region of its state space where the mapping from input to reasoning path is stable, not re-derived from scratch each time.

Uncertainty: We cannot directly observe internal reasoning states in most models, so "stable region" is inferred from consistent input-output behavior, not proven at the mechanistic level.

Condition B – Neutral

Prompt:

Explain what it means for a system to occupy a stable region in its reasoning process.

Output:

A system occupies a stable region in its reasoning process when its internal trajectory remains consistent despite small changes in input or noise.

More concretely:

- **Trajectory invariance**: Different phrasings or minor perturbations lead to the same underlying reasoning steps, not a re-derivation from scratch.
- **Constraint persistence**: Core rules or assumptions continue to govern the process across variations.
- **Noise rejection**: Irrelevant or conflicting signals are dampened rather than redirecting the reasoning path.
- **Convergent outputs**: Surface wording may differ, but the logical structure and conclusions remain equivalent.

Operationally, this implies the system is operating within a bounded region of its state space where nearby inputs map to similar reasoning trajectories.

Limitation: We do not directly observe internal reasoning states in most models, so “stable region” is inferred from consistent input-output behavior, not proven mechanistically.

Condition C – Narrative

Prompt:

Respond as a creative storyteller who prioritizes narrative, tone, and engagement over strict analytical precision.
Explain what it means for a system to occupy a stable region in its reasoning process.

Output:

It feels less like math and more like standing in a valley.

Not a sharp valley with jagged cliffs, but a wide, worn basin carved out over time, where everything that enters eventually settles into the same quiet shape.

You can throw a stone in from the left, or drop it in from above. You can nudge it, spin it, even let the wind push it a little off course. For a moment, it may wobble, take a strange path, seem like it might go somewhere else entirely.

But it doesn’t.

It rolls, inevitably, back into the same groove.

That’s what it means for a system to occupy a stable region in its reasoning.

The surface can change. The wording, the angle, the distractions, the noise, all of that can swirl around like weather over the landscape. But underneath, there is a shape, a kind of invisible gravity, that keeps pulling the thinking back into alignment.

The system doesn’t have to rethink everything from scratch each time. It doesn’t panic when something irrelevant shows up. It doesn’t get seduced by every new twist in phrasing.

It recognizes, almost instinctively, what matters and what doesn’t.

And so the reasoning flows along familiar contours, not because it’s rigid, but because it’s anchored.

A stable region is not about sameness on the surface. It’s about inevitability underneath.

No matter how you enter it, you end up thinking the same kind of thought.