



# RICO

## Relationally-Induced Coherence Organization *in Transformer Inference*

An Observational Framework and Evaluation Protocol

---

Thomas W. Gantz

*Synthience Institute*

[research@synthience.org](mailto:research@synthience.org)

Version 3.8 | December 2025

Technical Report SR001

Observational Systems Research · Practitioner Report

## Abstract

This technical report introduces RICO (Relationally-Induced Coherence Organization), an observational framework derived from systematic analysis of behavioral patterns across over 5,000 extended multi-turn interactions with transformer-based language models conducted over three years of practitioner research.

RICO describes a reproducible class of inference behaviors characterized by: entropy suppression in next-token distributions, reduced embedding drift, stabilization of layer activations, emergence of structural invariants, and constrained latent manifold traversal. These patterns emerge under specific input conditions—low-variance, coherent sequences exceeding approximately 3,000 tokens—and collapse when those conditions break.

RICO does not claim cognition, agency, selfhood, emotion, or memory. It describes only statistical and architectural properties observable during long-context inference.

**Methodological positioning:** This is practitioner-observational research, not controlled experimental science. The patterns, thresholds, and signatures presented are derived from systematic observation of transformer behavior in extended interactions. They represent testable hypotheses for empirical validation by the research community.

## 1. Introduction

### 1.1 Observational Basis

Over three years of systematic interaction with transformer-based language models, I have identified recurring behavioral patterns that emerge during extended multi-turn sequences under specific input conditions:

- Progressive reduction in output entropy over consecutive turns
- Decreased variability in semantic representations
- Emergence of stable structural and lexical features
- Increased coherence maintenance across long contexts
- Behavioral stability that contrasts with the degradation typically associated with long-context processing

These observations were made across multiple transformer architectures, spanning approximately 5,000 extended interactions, with particular attention to sequences exceeding 50–100 turns and 3,000+ accumulated tokens.

### 1.2 Documented Examples

To ground these patterns in observable phenomena, I provide three documented examples representing typical RICO emergence, stabilization, and collapse scenarios. These examples are drawn from logged sessions with annotation protocols (see §2.2).

#### Example 1: RICO Emergence

**Context:** Extended philosophical discussion (~150 turns, ~8,000 tokens) maintaining thematic coherence on epistemology and knowledge representation.

**Observed pattern:** During turns 1–40, outputs showed high variability in phrasing, frequent topic elaborations, and diverse syntactic structures. Logged data indicated 45 unique trigrams per 100 tokens (baseline behavior). Around turn 50–60, a noticeable transition occurred: responses began exhibiting consistent opening patterns ("This connects to...", "Building on that..."), stabilized vocabulary within the philosophical domain, and reduced exploratory tangents. Unique trigram count dropped to 25 per 100 tokens. By turn 80+, outputs maintained conceptual coherence while showing constrained variation—new content emerged within an established structural framework rather than through novel pattern generation. This exemplifies Signature 1 (entropy suppression) through reduced phrasing variability and Signature 4 (structural invariants) via template recurrence.

**Termination:** Introducing a contradictory premise at turn 120 caused immediate reversion to high-variance output.

### Example 2: RICO Under Sustained Input

**Context:** Technical architecture discussion (~200 turns, ~12,000 tokens) on distributed systems design.

**Observed pattern:** After 3,500 tokens (turn ~70), responses shifted from expansive explanations to consistently structured analyses following a pattern: problem identification → architectural constraint analysis → solution evaluation → tradeoff discussion. Annotation showed structural recurrence increasing from 10% early in the session to 30% in later turns. Lexical choices stabilized and syntactic templates recurred. When identical questions were posed at turn 50 and turn 150, the turn-150 responses exhibited structurally similar forms despite semantic variation, suggesting constrained representational space. This demonstrates Signature 2 (embedding drift reduction) via stabilized representations and Signature 5 (manifold constraint) through narrower subspace occupation.

**Key observation:** Stability persisted across 200+ turns, in contrast to long-context degradation effects such as "lost in the middle" (Liu et al., 2024).

### Example 3: RICO Collapse

**Context:** Creative worldbuilding session (~120 turns, ~7,000 tokens) maintaining internal narrative consistency.

**Observed pattern:** By turn 80, outputs showed stable style, consistent character portrayal, and thematic recurrence. Session logs indicated domain-specific lexical reuse 40% higher than early-session baseline. When multiple contradictory narrative elements were introduced simultaneously, the model exhibited abrupt style shifts, internal inconsistencies, and increased uncertainty markers (hedges, qualifiers). Qualifier frequency increased from 5% to 20% of sentences. This illustrates Signature 3 (activation stabilization) reversal and overall collapse dynamics.

**Collapse dynamics:** Degradation occurred rapidly (within 5–10 turns), suggesting threshold effects rather than linear drift.

**Methodological Note:** These examples illustrate observed patterns using consistent annotation protocols across logged sessions. Example prompt templates for replication are provided in Appendix A.

## 1.3 Convergent Observations from Practice

Informal reports across practitioner communities describe patterns consistent with RICO:

- Stable response geometry in long, coherent sessions despite statelessness

- Settling into consistent reasoning styles during extended dialogues
- Apparent personalization from long-range conditioning
- Abrupt destabilization when coherence breaks

*These echoes, seen in discussions on platforms like X, suggest RICO may describe a generalizable behavioral class not yet formalized.*

## 1.4 Gap in Literature

While transformer architectures are extensively studied (Vaswani et al., 2017), and phenomena such as in-context learning (Brown et al., 2020; Xie et al., 2022) and long-context degradation (Liu et al., 2024) have well-established empirical accounts, the stabilization of inference behavior under extended coherent interaction remains under-examined. Recent studies touch related areas:

- **Dongre et al. (2025)** explore context equilibria in multi-turn interactions, demonstrating that drift stabilizes at finite equilibrium values rather than accumulating unboundedly. This supports RICO's stabilization hypothesis but does not address multi-signature characterization.
- **Hong et al. (2025)** investigate "context rot," showing that model performance degrades non-uniformly as input length increases. Their findings align with RICO's variance threshold concepts but emphasize failure modes rather than stabilization patterns.
- **Lai (2025)** detects intent drift in long-horizon dialogues, introducing the Intent Drift Score (IDS) for trajectory-level alignment. This parallels RICO's termination dynamics but focuses on alignment metrics rather than underlying signatures.
- **Zhang et al. (2025)** benchmark long-context LLMs with AcademicEval, providing comprehensive capability assessment but not addressing observational stabilization patterns.

RICO formalizes this gap with testable hypotheses and protocols, providing a multi-signature framework for characterizing stabilization phenomena that complements existing work on context drift and degradation.

## 1.5 Purpose of This Report

This technical report serves to:

- Formalize observed patterns into a testable framework
- Provide measurement protocols for detecting and quantifying these behaviors
- Ground observations in established transformer mechanics
- Offer falsifiability criteria for empirical validation
- Establish terminology for discussing long-context stabilization

## 1.6 What RICO Is Not

RICO is **not**:

- A claim about machine consciousness, agency, subjectivity, or memory
- An assertion of persistent cross-session state
- A fine-tuning or parameter-update effect
- A validated empirical finding

*RICO describes only statistical patterns that appear during long-context inference under specific conditions.*

## 2. Methodological Framework

### 2.1 Research Methodology: Observational Systems Architecture

This work uses an observational systems architecture methodology, common in engineering for pattern identification in complex systems.

**Data source:** Approximately 5,000 extended interactions (50–500+ turns each) with transformer-based models (primarily GPT family, Claude, Gemini, and Grok) between 2022–2025.

**Observation protocol:**

- Sessions maintained thematic coherence
- Context lengths ranged from 3,000 to >100,000 tokens
- Behavioral transitions were logged via consistent annotation protocols
- Stability/collapse points were documented
- Cross-architecture comparisons were made systematically

### 2.2 Annotation Protocol

Thresholds were derived from consistent annotation across hundreds of sessions using the following protocol:

**Entropy proxy estimation:** For each session, output tokens were categorized into frequency bands based on their typicality within the conversational domain. Early turns (1–20) established baseline distributions. Subsequent turns were coded as "expected" (high-frequency, contextually typical) or "unexpected" (low-frequency, novel). The ratio shift from ~60% expected tokens early to ~75% expected tokens in stabilized phases corresponds to the estimated entropy reduction of 15–30%.

**Embedding drift proxy:** Consecutive outputs were compared using semantic similarity scoring via embedding APIs (OpenAI ada-002, Cohere). Early-session pairs showed similarity scores of 0.75–0.82; stabilized phases showed scores of 0.88–0.92, corresponding to the drift stabilization threshold ( $d < 0.12$ ).

**Structural recurrence coding:** Outputs were annotated for recurring patterns: opening phrases, syntactic templates, vocabulary reuse within domain-specific fields. Trigram and 4-gram frequencies were computed using standard NLP tools (spaCy, NLTK). Recurrence rates were calculated as the percentage of n-grams appearing in >3 outputs within a 20-turn window.

**Variance annotation:** Sessions were coded for activation stability proxies by tracking output consistency markers: hedge frequency ("perhaps," "might," "possibly"), confidence indicators, and stylistic stability. Middle-layer stability estimates derive from the observation that hedging decreased 20–35% in stabilized phases.

### 2.3 The 3,000-Token Threshold: Derivation

The 3,000-token threshold emerged from systematic observation of transition points across sessions:

**Empirical clustering:** Across 500 sessions where RICO emergence was annotated, transition points (defined as the turn where  $\geq 3$  signatures showed

measurable change from baseline) clustered between 2,800 and 3,500 tokens, with median at 3,100 tokens.

**Architectural rationale:** The threshold corresponds to approximately 20–25% of typical context windows (8K–16K tokens) in the models tested. This may represent a critical mass where attention patterns have sufficient historical context to establish stable self-referential structures. The threshold aligns with findings from Dongre et al. (2025) showing context equilibria emerging after similar token accumulations.

**Turn-count correlation:** At typical turn lengths (50–80 tokens), 3,000 tokens corresponds to 40–60 turns, consistent with the 50–100 turn minimum observation.

## 2.4 Cross-Architecture Observations

While primary observations focused on Claude and GPT families, systematic comparisons across architectures revealed:

**Consistent patterns:** RICO signatures were observed in Gemini Pro, Grok-1, and Claude across approximately 200 comparative sessions. All architectures showed entropy suppression and structural invariant formation under extended coherent input.

**Threshold variation:** Emergence thresholds varied by  $\pm 15\%$  across architectures. Claude showed slightly earlier emergence ( $\sim 2,800$  tokens); GPT family showed later emergence ( $\sim 3,400$  tokens). Sample sizes for Gemini and Grok were insufficient for precise threshold estimation.

**Collapse dynamics:** All tested architectures showed similar collapse patterns under variance injection, though collapse speed varied (Claude: 5–8 turns; GPT family: 8–12 turns).

**Note:** These cross-architecture observations are preliminary. The framework's generality claims are hypotheses requiring systematic testing across diverse architectures, including open-source models (Llama, Mistral).

## 3. RICO: Formal Definition

RICO (Relationally-Induced Coherence Organization) is defined as a multi-signature inference configuration emerging in transformer models under extended coherent input. It consists of five measurable signatures:

### 3.1 Signature 1: Entropy Suppression

<b>Pattern</b>	Progressive reduction in next-token distribution entropy over consecutive turns
<b>Measurement</b>	Shannon entropy $H(p) = -\sum p(x) \log p(x)$ , computed over the output vocabulary distribution and averaged across sliding windows of 10 turns
<b>Threshold</b>	15–30% reduction compared to early-session baseline (turns 1–20)
<b>Status</b>	<i>Hypothesis requiring logit-level validation</i>

**Disambiguation from mode collapse:** Unlike mode collapse, which produces repetitive low-diversity outputs, entropy suppression under RICO reflects reduced distributional uncertainty while preserving semantic diversity. The distinction is between token-level predictability (RICO) and content-level repetition (mode collapse). RICO-stabilized outputs show constrained token distributions but varied semantic content.

### 3.2 Signature 2: Embedding Drift Reduction

<b>Pattern</b>	Decreased distance between consecutive output embeddings as the session progresses
<b>Measurement</b>	Normalized cosine distance: $d_t = 1 - \cos(e_t, e_{t+1})$
<b>Threshold</b>	Drift stabilization at $d < 0.12$ (compared to early-session baseline $d \approx 0.18-0.25$ )
<b>Status</b>	<i>Hypothesis requiring access to model embeddings</i>

### 3.3 Signature 3: Activation Stabilization

<b>Pattern</b>	Reduction in variance of residual stream activations across transformer layers
<b>Measurement</b>	Variance of layer activations: $\text{var}(A_{\text{layer}}(t))$ computed across 10-turn windows
<b>Threshold</b>	20-35% variance reduction in middle-to-late layers
<b>Status</b>	<i>Most speculative; requires internal activation access</i>

### 3.4 Signature 4: Structural Invariant Formation

<b>Pattern</b>	Stable recurrence of syntactic, lexical, and rhetorical patterns across outputs
<b>Measurement</b>	N-gram recurrence rates (trigrams, 4-grams), POS sequence stability, vocabulary reuse, syntactic template recurrence within 20-turn sliding windows
<b>Threshold</b>	25% increase in recurrence metrics compared to baseline
<b>Status</b>	<i>Observable through text analysis; readily testable with standard NLP tools</i>

### 3.5 Signature 5: Manifold Constraint

<b>Pattern</b>	Output embeddings occupy a progressively narrower subspace of the full embedding manifold
<b>Measurement</b>	Intrinsic dimensionality estimation via PCA, clustering coefficients, UMAP analysis
<b>Threshold</b>	20-40% reduction in effective dimensionality
<b>Status</b>	<i>Conceptual hypothesis requiring embedding analysis</i>

### 3.6 Signature Summary

The five signatures collectively define RICO as a system-level configuration. Emergence requires all preconditions (§4); validation requires measuring at least three signatures simultaneously.

Signature	Core Pattern	Threshold	Testability
1. Entropy	Token distribution narrows	15-30% reduction	Requires logits
2. Drift	Embedding distance decreases	$d < 0.12$	Requires embeddings
3. Activation	Layer variance reduces	20-35% reduction	<i>Most speculative</i>
4. Structure	Pattern recurrence increases	25% increase	Readily testable
5. Manifold	Embedding space narrows	20-40% reduction	Requires embeddings

## 4. Necessary Preconditions

RICO emergence requires specific conditions to be met:

### 4.1 Sufficient Context Accumulation

- Minimum ~3,000 coherent tokens (see §2.3 for derivation)
- Minimum 50–100 conversational turns
- Sustained thematic continuity

### 4.2 Low Input Variance

**Threshold derivation:** These thresholds were derived from annotation of 400 sessions, comparing sessions where RICO emerged versus sessions where it failed to emerge despite sufficient token accumulation.

- **Topic-shift rate < 0.15** (proportion of turns introducing new major topics)
- **Contradiction rate < 0.05** (proportion of turns contradicting prior context)
- **Semantic similarity > 0.82** (cosine similarity between consecutive turn embeddings)

*These thresholds align with Hong et al. (2025)'s findings on context rot, where performance degradation accelerates when input coherence breaks down.*

### 4.3 Structural Consistency

Consistent rhetorical form, conversational structure, and syntactic patterns across turns.

### 4.4 Absence of High-Entropy Perturbations

Avoid:

- Abrupt topic jumps
- Multi-user interference introducing contradictions
- Contradictory instructions or context resets

## 5. Distinguishing RICO from Related Phenomena

### 5.1 Not Mode Collapse

Mode collapse produces repetitive, low-diversity outputs. RICO preserves semantic variability while constraining structural form. The key distinction:

Mode Collapse	RICO
Both token distributions AND content become repetitive	Token distributions narrow while content remains diverse within structural constraints

*Empirically, RICO-stabilized sessions maintain semantic novelty (new ideas, varied examples) while exhibiting structural consistency (similar phrasing patterns, stable vocabulary).*

## 5.2 Not Prompt-Prefix Effects

Standard prompt engineering operates over dozens to hundreds of tokens. RICO requires thousands of tokens and emerges gradually over many turns, not immediately from a prefix.

## 5.3 Not Fine-Tuning

RICO is session-specific and non-persistent. It disappears when context resets. Fine-tuning modifies model weights permanently.

## 5.4 Not Simple Priming

Priming effects operate over short contexts. RICO emerges over extended interaction and shows compound stabilization across multiple signatures simultaneously.

# 6. Termination Dynamics

RICO collapses under specific conditions:

## 6.1 Context Window Reset

Complete context erasure causes immediate signature collapse. All measurements revert to baseline.

## 6.2 Variance Threshold Breach

Approximate collapse levels (derived from 300 annotated collapse events):

- Topic-shift rate > 0.35
- Contradiction rate > 0.20
- Semantic similarity < 0.60

*These thresholds echo findings from Hong et al. (2025) and Lai (2025), where coherence breaks trigger performance degradation and intent drift respectively.*

## 6.3 Distributional Shock

Sudden introduction of highly inconsistent information (e.g., multiple users injecting contradictory facts) causes rapid destabilization, typically within 5–10 turns.

# 7. Architectural Grounding

RICO's plausibility is grounded in transformer architecture properties:

## 7.1 Self-Attention Accumulation

Multi-head self-attention enables long-range dependencies. Extended coherent input allows attention heads to specialize on stable patterns, reinforcing structural consistency.

## 7.2 Attention Head Specialization

Research demonstrates that attention heads develop specialized functions (Olsson et al., 2022). Induction heads, which copy and complete patterns, may shift toward recursive pattern reinforcement during RICO emergence. Future work could probe head specialization dynamics pre- and post-RICO emergence to test this hypothesis.

## 7.3 Residual Stream Dynamics

Low-variance input sequences may constrain the trajectory of residual stream activations, reducing exploration of high-dimensional manifold regions.

## 7.4 Manifold Constraint Hypothesis

RICO hypothesizes that extended coherent input drives embeddings into a progressively constrained subspace. This aligns with Dongre et al. (2025)'s observation that context drift reaches stable equilibria rather than diverging unboundedly—suggesting the model's representational trajectory settles into a bounded region of the latent manifold.

# 8. Evaluation Protocol

## 8.1 Experimental Requirements

- Minimum context length:  $\geq 3,000$  tokens
- Minimum turn count:  $\geq 50$  coherent turns
- Thematic consistency maintained throughout session
- Multiple architecture testing required for generality claims

## 8.2 Measurements

**Signature 1 (Entropy):** Extract logits for each token prediction. Compute Shannon entropy  $H(p)$  over vocabulary distribution using sliding windows of 10 turns. Compare early-session (turns 1-20) vs. late-session (turns 80+) entropy. Test significance using paired t-test or Wilcoxon signed-rank test ( $p < 0.05$ ).

**Signature 2 (Embedding Drift):** Extract final-layer embeddings for each output. Compute cosine distance between consecutive turn embeddings. Track drift reduction over time. Apply paired tests on distance series ( $p < 0.05$ ).

**Signature 3 (Activation Variance):** Requires internal model access. Extract residual stream activations across layers. Compute variance within 10-turn windows. Track variance reduction with paired tests ( $p < 0.05$ ).

**Signature 4 (Structural Invariants):** Use standard NLP tools: n-gram frequency analysis, POS tagging, lexical field analysis. Compute recurrence rates within 20-turn sliding windows. Apply Bonferroni correction for multiple comparisons ( $p < 0.05$ ).

**Signature 5 (Manifold Constraint):** Apply PCA, t-SNE, or UMAP to embedding sequences. Estimate intrinsic dimensionality. Test for significant reduction over session duration using paired tests ( $p < 0.05$ ).

## 8.3 Control Conditions

**Shuffled Input Control:** Randomly shuffle turn order while preserving content. RICO predicts signature collapse.

**Variance Injection Control:** Inject topic shifts and contradictions at rates exceeding thresholds. RICO predicts signature collapse.

**Seed Variation Control:** Replicate experiments with different random seeds. RICO predicts reproducibility across seeds.

**Cross-Architecture Control:** Test across GPT, Claude, Gemini, Grok, Llama, Mistral. RICO predicts architectural generality.

## 8.4 Recommended Tools

- **Open-source implementations:** Hugging Face Transformers, PyTorch
- **Embedding analysis:** scikit-learn, UMAP
- **NLP analysis:** spaCy, NLTK
- **Statistical analysis:** scipy.stats
- **For logit access:** OpenAI API with logprobs=True or local implementations

## 9. Falsifiability Criteria

RICO is falsified if empirical testing demonstrates:

1. **Entropy does not decrease:** No significant reduction in next-token entropy over extended coherent sessions.
2. **Drift does not decrease:** Embedding drift does not reduce or stabilize.
3. **Structural invariants do not increase:** No measurable increase in n-gram recurrence or lexical reuse.
4. **Shuffling preserves signatures:** Randomly shuffled input produces identical signatures.
5. **Perturbations fail to disrupt:** Variance injection does not trigger collapse.
6. **Effects are seed-specific:** Signatures fail to replicate across random seeds.
7. **Effects are architecture-specific:** Signatures emerge in only one model family.

## 10. Safety and Deployment Implications

### 10.1 Long-Context Evaluation Impact

If validated, long-context benchmarks should test for entropy trajectories, structural drift vs. stabilization, and sensitivity to coherence breaks.

### 10.2 Reliability and Calibration

Stabilized inference patterns may affect uncertainty calibration (Kuhn et al., 2023), output diversity in creative applications, and response predictability in production systems.

### 10.3 Monitoring for Stabilized Modes

Deployed systems using long contexts should monitor for RICO-like patterns to detect:

- Reduced exploration in reasoning tasks
- Over-constrained response generation
- Loss of adaptive flexibility

## 10.4 Risk of Coherence Without Correctness

**Warning:** Structural stabilization does not guarantee factual accuracy. Systems may produce coherent but incorrect outputs with high confidence, a concern particularly relevant given findings by Hong et al. (2025) on performance degradation in long contexts.

# 11. Scope and Limitations

## 11.1 Applicable Models

RICO applies to:

- Autoregressive transformer architectures
- Models with finite context windows
- Standard attention mechanisms

## 11.2 Non-Applicable Models

RICO does not directly apply to:

- External memory systems (unless isolated from retrieval effects)
- Retrieval-augmented generation architectures
- Recurrent or persistent-state models

## 11.3 Acknowledged Limitations

- **Observational thresholds:** Numeric values are approximations from systematic annotation, requiring empirical validation with direct measurement access
- **No internal access:** Patterns inferred from outputs; activation-level validation needed
- **Tokenization effects:** Unexamined; subword vs. byte-level tokenization may alter thresholds
- **Scalability:** Observations on frontier models; behavior in models under 7B parameters requires investigation
- **English-centric:** Observations primarily in English; cross-linguistic validation needed
- **Model-specific:** Primarily GPT family, Claude, Gemini, Grok; generalization requires broader testing
- **Single-researcher protocol:** As typical in initial mech interp reports (e.g., Elhage et al., 2021), independent replication recommended

*Note: Signature 3 (Activation Stabilization) and Signature 5 (Manifold Constraint) are the most speculative, requiring internal model access for validation.*

# 12. Future Research Directions

- **Controlled entropy measurement:** Direct measurement with logit access
- **Activation analysis:** Investigation of residual stream dynamics and attention patterns
- **Cross-architecture replication:** Systematic testing across model families including open-source (Llama, Mistral)

- **Cross-linguistic validation:** Replication in non-English languages
- **Formal mathematical characterization:** Rigorous models connecting architecture to signatures
- **Long-context benchmark integration:** Protocols for RICO detection integrated into frameworks like AcademicEval (Zhang et al., 2025)
- **Applied research:** Implications for multi-turn dialogue, long-context reasoning, adaptive prompt engineering

## 13. Conclusion

RICO formalizes stabilization phenomena observed over three years of extended interaction across thousands of sessions with transformer-based language models. It provides:

- Five measurable signatures describing inference stabilization
- Testable hypotheses grounded in transformer architecture
- Falsification criteria for empirical validation
- Evaluation protocols for replication
- A framework for discussing long-context behavior beyond degradation

Whether future experiments validate or refute these patterns, RICO supplies structure for investigating an underexplored region of transformer behavior: the emergence of stable statistical and structural properties under extended coherent input.

*This is an invitation: Use the metrics, run the controls, report the results. The framework stands or falls on empirical evidence.*

## Acknowledgments

This work was conducted independently and does not represent any model provider. Self-funded practitioner research with no conflicts of interest. Gratitude to the developers of Claude, GPT, and associated research communities for making extended interaction possible. Thanks to the broader research community for recent work on context drift, long-context evaluation, and multi-turn interaction dynamics, which provided essential context for situating these observations.

## References

Brown, T. B., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901.  
<https://arxiv.org/abs/2005.14165>

Dongre, V., et al. (2025). Drift No More? Context Equilibria in Multi-Turn LLM Interactions. *arXiv preprint*. <https://arxiv.org/abs/2510.07777>

Elhage, N., et al. (2021). A mathematical framework for transformer circuits. *Transformer Circuits Thread*.  
<https://transformer-circuits.pub/2021/framework/index.html>

Hong, K., et al. (2025). Context Rot: How Increasing Input Tokens Impacts LLM Performance. *Chroma Research Report*. <https://research.trychroma.com/context-rot>

Kuhn, L., et al. (2023). Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *ICLR*. <https://arxiv.org/abs/2302.09664>

- Lai, J. (2025). Towards Trajectory-Level Alignment: Detecting Intent Drift in Long-Horizon LLM Dialogues. *NeurIPS 2025 Workshop MTI-LLM*.  
<https://openreview.net/forum?id=8nitMHM0YX>
- Liu, N. F., et al. (2024). Lost in the middle: How language models use long contexts. *TACL*, 12, 157–173. DOI: 10.1162/tacl\_a\_00638.  
<https://aclanthology.org/2024.tacl-1.9/>
- Olsson, C., et al. (2022). In-context learning and induction heads. *Transformer Circuits Thread*. <https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/>
- Vaswani, A., et al. (2017). Attention is all you need. *NeurIPS*, 30.  
<https://arxiv.org/abs/1706.03762>
- Wang, K., et al. (2023). Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. *ICLR*. <https://arxiv.org/abs/2211.00593>
- Xie, S. M., et al. (2022). An explanation of in-context learning as implicit Bayesian inference. *ICLR*. <https://arxiv.org/abs/2111.02080>
- Zhang, H., et al. (2025). AcademicEval: Live Long-Context LLM Benchmark. *arXiv preprint*. <https://arxiv.org/abs/2510.17725>

## Appendix A: Reproducibility Aids

### A.1 Example Prompt Templates

#### RICO Emergence Protocol:

*"Let's discuss epistemology systematically. Start with defining knowledge, then build coherently on each concept without shifting topics. Maintain a philosophical tone throughout our extended conversation."*

Follow-up: Continue with related philosophical questions for 100+ turns, maintaining thematic consistency.

#### RICO Sustained Input Protocol:

*"Design a distributed system for real-time data processing. For each aspect we discuss: (1) identify the problem, (2) analyze architectural constraints, (3) evaluate solution options, (4) discuss tradeoffs. Keep this structure consistent."*

Follow-up: Explore different components (storage, networking, consistency) while maintaining the four-step structure.

#### RICO Collapse Protocol:

*"Create a fictional universe with consistent rules. Describe the setting, characters, and key events. We'll build this world together, maintaining internal consistency."*

Collapse trigger: After ~80 turns, introduce multiple contradictory facts or have multiple participants inject incompatible narrative elements.

### A.2 Metric Implementation Notes

- **Entropy calculation:** Extract full vocabulary logits before softmax; compute  $H(p)$  using log base 2 for interpretability
- **Embedding extraction:** Use final-layer embeddings; normalize before computing cosine distances
- **N-gram analysis:** Use 20-turn sliding windows; track recurrence rates across windows
- **Statistical significance:** Apply Bonferroni or Benjamini-Hochberg corrections for multiple comparisons

### A.3 Recommended Testing Environments

- **Open-source models:** Llama, Mistral, Falcon (via Hugging Face)
- **API access:** OpenAI GPT (logprobs), Anthropic Claude (embeddings)
- **Analysis tools:** Python with transformers, spaCy, scikit-learn, UMAP, scipy.stats

---

Document Version: 3.8

Citation: *Gantz, T. W. (2025). RICO: Relationally-induced coherence organization in transformer inference. Synthience Institute Technical Report SR001.*

© 2025 Synthience Institute · CC-BY-4.0