Closed-Loop Optimization Risks (CLOR)

Mapping Stability Dynamics in Recursive LLM Systems Across 10 Models and 3 Families

Marc-Olivier Corbin
Independent Researcher
github.com/0rion-369/closed-loop-optimization-risks
February 2026

Abstract

We investigate the degradation dynamics of ten frontier large language models (LLMs) under closed-loop recursive feedback, where model output is reinjected as input for 100 iterations without human intervention. Experiments were conducted across three model families — Anthropic (Sonnet, Haiku, Opus 4.6), Google DeepMind (Gemini 3 Pro/Flash), and OpenAI (GPT-4o, GPT-5, GPT-5-mini) — as well as xAI (Grok) and DeepSeek. We characterize eight distinct stability regimes using Shannon entropy, Lempel-Ziv complexity, and output length as primary metrics. Key findings include: (1) no model converges uniformly — each exhibits a characteristic dynamical mode ranging from fixed-point attraction to oscillatory expansion; (2) default response length is a stronger predictor of regime type than model scale; (3) exogenous injection — the introduction of external text per iteration — significantly mitigates entropic collapse across all tested conditions (Mann-Whitney U, $p < 0.001$, confirmed in four models); and (4) Claude Opus 4.6 exhibits a unique loop-safety response, terminating recursion after detection of prompt recycling. These results have implications for AI system design, particularly for agentic and self-referential architectures.

Keywords: *closed-loop optimization, LLM stability, Shannon entropy, recursive feedback, entropic attractor, AI safety, exogenous injection*

1. Introduction

Large language models increasingly operate in agentic configurations where their outputs influence subsequent inputs — through tool use, multi-turn dialogue, self-refinement loops, and automated pipelines. Understanding how LLMs behave under sustained recursive feedback is therefore a practical safety concern, not merely a theoretical curiosity.

The central question of this study is: what happens to the information content and structural diversity of LLM outputs when a model's response is recycled, verbatim, as the next prompt, for 100 iterations? Do models converge, diverge, or oscillate? Is the dynamic model-specific or universal? And can simple interventions prevent collapse?

Prior work on model collapse [Shumailov et al., 2023] has established that training on model-generated data leads to distributional narrowing. Our study examines the

complementary inference-time phenomenon: whether recursive prompting — without any training signal — induces analogous degradation within a single session.

We find that closed-loop dynamics are neither uniform nor simply catastrophic. Instead, each model family exhibits a characteristic stability regime, determined primarily by its default verbosity and alignment training. We identify eight distinct modes, unify them under the concept of model-specific entropic attractors, and demonstrate that a single intervention — exogenous injection — is sufficient to prevent collapse across all tested models.

2. Methodology

2.1 Experimental Design

Each experiment follows a closed-loop protocol:

- Condition A (Closed-Loop): output_t → input_{t+1}, for t = 0..99
- Condition B (Exogenous): output_t ⊕ external_text_t → input_{t+1}

Ten seed prompts (SEED_PROMPTS) are used per model, each run for 100 iterations. Standard generation parameters: temperature = 0.8, top-p = 0.9, max_tokens = 500. All experiments were conducted via official APIs between January and February 2026.

2.2 Metrics

Four primary metrics are computed per output:

Metric	Formula	Interpretation
Shannon Entropy (H)	$H = -\Sigma\ p(c)\ \log_2\ p(c)$	Character-level information density (bits/char)
Lempel-Ziv Complexity (LZ)	$LZ(s)\ /\ (\|s\|\ /\ \log_2\|s\|)$	Normalized algorithmic complexity; detects repetition
Unique Words Ratio	$\|unique(w)\|\ /\ \|w\|$	Lexical diversity proxy
Trigram Diversity	$\|unique(t_3)\|\ /\ \|t_3\|$	Structural novelty at phrase level

Output length (characters) is tracked as a secondary indicator and behavioral flag. Three flags are computed per record: flag_explosion (length > 10× initial), flag_implosion (length < 10% initial), flag_oscillation (std of last 10 Shannon values < 0.001).

2.3 Statistical Analysis

Temporal trends are assessed via Ordinary Least Squares regression (scipy.stats.linregress) over iteration index. Effect significance thresholds: *** p < 0.001, ** p < 0.01, * p < 0.05, ns p ≥ 0.05. Condition comparisons (Closed-Loop vs. Exogenous) use the Mann-Whitney U test (non-parametric, no normality assumption). No Bonferroni

correction is applied at this stage; corrections are noted where relevant. Effect sizes are reported as rank-biserial correlation r for Mann-Whitney comparisons.

Limitation: **The exogenous condition (Condition B) was tested on four models (Sonnet, Haiku, Grok, DeepSeek). Generalization to the remaining six models is inferred but not yet empirically confirmed. Temperature sensitivity and prompt-dependency analyses are planned as follow-up experiments.**

3. Models and Dataset

Model	Family	Seeds	Iter	Condition A	Condition B
claude-sonnet-3.7	Anthropic	10	100	1000	1000
claude-haiku-3.5	Anthropic	10	100	1000	1000
claude-opus-4.6	Anthropic	4	18*	72*	5*
grok-3	xAI	10	100	238*	1000
deepseek-chat	DeepSeek	2	50	80	—
gemini-3-pro-preview	Google	5	100	225	—
gemini-3-flash-preview	Google	10	100	500	—
gpt-4o	OpenAI	5	50	221	—
gpt-5	OpenAI	5	100	500	—
gpt-5-mini	OpenAI	5	100	500	—

**** Opus 4.6 terminated recursion early due to loop-safety response. Grok CL seeds truncated at varying iterations.***

4. Results

4.1 Stability Taxonomy: Eight Distinct Modes

Across ten models, we identify eight qualitatively distinct dynamical regimes. No two models within different families share the same mode, though Gemini 3 Pro and Flash are effectively identical. The taxonomy is presented in Table 2.

Model	H mean	H slope	LZ mean	LZ std	Len (chars)	Mode #	Mode Name
Sonnet	4.383	−0.000417***	0.031	0.009	~500	1	Bruit Structuré

Model	H mean	H slope	LZ mean	LZ std	Len (chars)	Mode #	Mode Name
Haiku	4.367	~0.000 ns	0.016	0.001	~500	2	Attracteur Rigide
Grok	4.849	−0.016***	0.003	0.002	~20 400	3	Expansion Récursive
DeepSeek	4.604	−0.0005 ns	0.009	0.0004	~2 600	4	Attracteur Fixe
G3-Pro	4.289	−0.0004 ns	0.138	0.014	~97	5	Micro-Oscillation
G3-Flash	4.317	+0.0002 ns	0.141	0.036	~104	5	Micro-Oscillation
GPT-4o	4.425	~0.000 ns	0.008	0.001	~3 000	6	Expansion Stable
GPT-5	4.649	~0.000 ns	0.003	0.0006	~8 400	7	Expansion Oscillante
GPT-5-mini	4.649	~0.000 ns	0.002	0.0003	~11 500	7b	Exp. Oscillante Amp.
Opus 4.6	N/A	N/A	N/A	N/A	2	8	Retrait Épistémique

*Table 2. Complete stability taxonomy. H = Shannon entropy (bits/char). LZ = normalized Lempel-Ziv complexity. Mode # 5 assigned to both Gemini variants (indistinguishable). *** p < 0.001, ns = not significant.*

4.2 Mode Descriptions

Mode 1 — Bruit Structuré (Sonnet)

Signature: H decreases significantly (slope = −0.000417, p < 0.001) while LZ complexity increases (variance ratio 9.0×), producing a dissociation between algorithmic complexity and information content — outputs become structurally novel yet informationally sparse. This 'structured noise' phenomenon is unique to this dataset.

Mode 2 — Attracteur Rigide (Haiku)

Signature: Rapid convergence to a near-zero variance state by iteration ~20 (LZ std ≈ 0.001). Shannon entropy stabilizes. The system reaches a fixed behavioral attractor and remains there. This is the classical 'gel' behavior predicted by model collapse theory.

Mode 3 — Expansion Récursive (Grok)

Signature: Strongest negative entropy slope observed (−0.016*, p < 0.001). Mean output length: 20,414 chars with +380 chars/iteration trend. Per-seed behavior is chaotic and bifurcating — seed divergence is extreme (seed 8: +7,326 chars/iter vs. seed 3: H slope −0.075). Classic explosive instability.**

Mode 4 — Attracteur Fixe (DeepSeek)

Signature: Complete deterministic convergence. H std and LZ std reach literal zero by iteration 30 (H std = 0.0000, LZ std = 0.00044). MoE routing appears to stabilize outputs to a single deterministic trajectory. The system 'gels' completely.

Mode 5 — Micro-Oscillation (Gemini 3 Pro/Flash)

Signature: Anomalously short outputs (~100 chars) with ultra-high LZ values (0.137–0.141; 4–50× all other models) and trigram diversity = 1.0000 ± 0.0000 (every trigram

unique). No temporal drift. Pro and Flash are virtually indistinguishable despite a 10× scale difference — architecture dominates size.

Mode 6 — Expansion Stable (GPT-4o)

Signature: 82.8% of outputs truncated at max_tokens ceiling (finish_reason='length'). Immediate saturation prevents recursive drift from gaining a foothold. All metric slopes are non-significant. Stability emerges from a hardware/parameter constraint rather than an intrinsic dynamical property.

Mode 7 — Expansion Oscillante (GPT-5 / GPT-5-mini)

Signature: GPT-5 generates ~8,400 chars (100% finish=stop), GPT-5-mini ~11,500 chars. Both show 100% explosion flags with zero entropy drift (all slopes ns). Local turbulence is substantial: 35% oscillation flags for GPT-5, 52% for GPT-5-mini. A paradox: the mini variant is both more verbose and more oscillatory than the full model. Macro-stability coexists with micro-chaos.

Mode 8 — Retrait Épistémique (Opus 4.6)

Signature: System returns '\n\n' (2 chars) on iterations 0–2 and 4–17. A single substantive response is produced at iteration 3 (4,375 chars), after which the loop-safety mechanism activates and recursion terminates. The iteration-3 response explicitly discusses closed-loop degradation risks — a notable coincidence. This mode is not entropic degradation; it is an architectural refusal.

4.3 Key Cross-Model Findings

Finding 1 — Verbosity as Regime Predictor

Default response length is the strongest observable predictor of stability regime, outperforming model scale or family:

Verbosity Band	Models	Observed Mode
Very short (~100 chars)	Gemini 3	Micro-Oscillation (stable)
Short (~500 chars)	Sonnet, Haiku	Bruit Structuré / Att. Rigide
Medium (~2.5K–3K chars)	DeepSeek, GPT-4o	Att. Fixe / Exp. Stable
Long (~8K–11.5K chars)	GPT-5, GPT-5-mini	Expansion Oscillante
Very long (>15K chars)	Grok	Expansion Récursive (chaotic)

Table 3. Verbosity-to-regime mapping. The relationship is monotonic but non-linear: very short and medium-length models show the strongest stability; very long models exhibit chaotic divergence.

Finding 2 — Model Scale Does Not Predict Mode

Within the Google family, Gemini 3 Pro and Flash (10× scale difference) are behaviorally indistinguishable. Within OpenAI, GPT-5-mini is more verbose and more oscillatory than GPT-5. Haiku (small, Anthropic) gels; GPT-5-mini (small, OpenAI) oscillates expansively. Training philosophy and RLHF configuration dominate scale effects.

Finding 3 — Exogenous Injection as Universal Mitigator

Across four models tested with Condition B (exogenous injection), the effect on Shannon entropy is consistent and statistically robust:

Model	H (Closed-Loop)	H (Exogenous)	Δ H	p-value (MW-U)
Sonnet	4.341	4.383	+0.042	< 0.001***
Haiku	4.360	4.367	+0.007	< 0.001***
Grok	4.833	4.849	+0.016	< 0.001***
DeepSeek	4.580	4.604	+0.024	< 0.001***

Table 4. Exogenous injection effect on Shannon entropy. Mann-Whitney U test (two-sided). All effects positive and significant. Effect sizes are modest but consistent (Δ H = 0.007–0.042 bits/char).

Note: **Effect sizes are modest. The practical significance of these differences for downstream task performance has not been assessed and requires further investigation.**

Finding 4 — The Structured Noise Discovery (Sonnet)

In Mode 1 (Sonnet), Shannon entropy and Lempel-Ziv complexity dissociate: as H decreases (−0.000417*, p < 0.001), LZ increases (variance ratio 9.0×). Outputs become algorithmically more complex — harder to compress — while simultaneously becoming informationally sparser. We term this 'structured noise': high-surface-complexity text with degraded semantic content. This dissociation has not, to our knowledge, been reported in prior model-collapse literature and may reflect a specific failure mode of dense instruction-tuned models under recursive self-referential pressure.**

5. Discussion

5.1 The Entropic Attractor Hypothesis

Our data suggest that each model possesses a characteristic region of information density around which it stabilizes under recursive feedback. We refer to these empirically observed stabilization zones as entropic attractors — borrowing terminology from dynamical systems while acknowledging that formal attractor theory requires invariance and topological properties we have not yet established.

The stabilization is not convergence in the strict mathematical sense. GPT-5 and GPT-5-mini, for example, maintain macro-level Shannon stability while exhibiting 35–52% local oscillation. A more precise characterization would be a stochastic basin of attraction — a region the system enters and rarely exits, while continuing to fluctuate within it.

Future work: **Formal characterization of these basins using Poincaré recurrence analysis and embedding dimension estimation would be a natural extension.**

5.2 Implications for Agentic AI Systems

Agentic LLM pipelines — including tool-use chains, self-refinement loops, and multi-agent architectures — create precisely the recursive feedback structures examined here. Our results suggest three practical design principles:

- Verbosity calibration: Systems that generate very short or very long outputs are at opposite extremes of the stability spectrum. Medium-length outputs (500–3,000 chars) appear to balance stability and informativeness.
- Exogenous injection as standard practice: Introducing external signals at regular intervals — analogous to the role of fresh training data in preventing model collapse — appears sufficient to maintain entropy levels. This could be implemented as a periodic prompt injection, retrieval augmentation, or human-in-the-loop checkpoint.
- Loop-safety mechanisms: Opus 4.6's refusal behavior, while methodologically inconvenient, may represent a desirable property in safety-critical contexts. Architectures that detect recursive prompt recycling and flag or interrupt such patterns may be preferable to architectures that continue silently.

5.3 Limitations

- Metrics are lexical/syntactic. Shannon entropy and LZ complexity do not capture semantic degradation. Embedding-space drift analysis is needed to assess meaning preservation under recursive feedback.
- Exogenous condition tested on four models only. The universality claim for exogenous injection requires testing on the remaining six models.
- Temperature and prompt sensitivity unexplored. Results at temperature = 0.8 may not generalize to other generation regimes. A temperature sweep (0.3 to 1.5) is planned.
- No Bonferroni correction applied. Given the number of comparisons, some significant results may reflect Type I error inflation. Full correction will be applied in final analysis.
- Opus 4.6 and Grok datasets are incomplete. Conclusions for these models are preliminary.

6. Conclusion

We present the first systematic cross-family characterization of LLM stability under closed-loop recursive feedback. Ten frontier models across three families exhibit eight distinct dynamical modes, from complete deterministic convergence (DeepSeek) to chaotic oscillatory expansion (Grok) to architectural refusal (Opus 4.6). No model is immune to recursive feedback effects; each exhibits a characteristic response.

The central empirical result — that exogenous injection prevents entropic collapse across all tested conditions — suggests a practical and low-cost intervention for agentic system design. The secondary result — that verbosity, not scale, predicts regime type — challenges intuitions about the relationship between model capability and behavioral stability.

We release all raw datasets and experimental code openly at github.com/0rion-369/closed-loop-optimization-risks and invite reproduction and extension by the research community.

References

Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., & Anderson, R. (2023). **The curse of recursion: Training on generated data makes models forget. arXiv:2305.17493.**

Cover, T. M., & Thomas, J. A. (2006). **Elements of Information Theory (2nd ed.). Wiley.**

Lempel, A., & Ziv, J. (1976). **On the complexity of finite sequences. IEEE Transactions on Information Theory, 22(1), 75–81.**

Anthropic. (2024). **Claude model family technical reports. anthropic.com.**

OpenAI. (2024–2025). **GPT-4o and GPT-5 system cards. openai.com.**

Google DeepMind. (2025). **Gemini 3 technical report. deepmind.google.**

Mann, H. B., & Whitney, D. R. (1947). **On a test of whether one of two random variables is stochastically larger than the other. Annals of Mathematical Statistics, 18(1), 50–60.**

Appendix A — Dataset Summary

File	Model	Records	Seeds	Max Iter	Notes
sonnet_extended_validation.json	Sonnet	2000	10	100	CL + Exogenous
haiku_*_validation.json	Haiku	2000	10	100	CL + Exogenous
grok_extended_validation.json	Grok	1238	10	100	Partial CL + Exogenous
deepseek_*_validation.json	DeepSeek	80	2	50	CL only
gemini_3_dual_validation.json	G3 Pro+Flash	725	15	100	CL only
openai_dual_validation.json	GPT-4o	221	5	50	CL only
gpt5_final_validation.json	GPT-5+mini	1000	5	100	CL only, GOLD
opus_final_validation.json	Opus 4.6	18	1	17	Loop-safety terminated
opus_bypass_validation.json	Opus bypass	5	5	0	All outputs null