Condition-Dependent Collapse Dynamics in Multi-Turn LLM Self-Play

Sohail Mohammad
Independent Researcher
sohailmo.ai@gmail.com

February 2026

Abstract

When two chat models converse for 40 turns, do they keep producing novel responses or collapse into repetitive loops? We measured this across 720 conversations under a preregistered design (4 conditions $\times$ 36 seeds $\times$ 5 repeats, 40 turns each) spanning homogeneous self-play configurations (Llama-3.1-8B, Qwen2.5-7B, Mistral-7B-v0.3) and a heterogeneous rotation condition. All 720 conversations ran to completion with full protocol integrity. Collapse rates varied substantially by condition: Qwen2.5-7B homogeneous self-play exhibited the highest mean collapse rate (0.773), while Mistral-7B showed the lowest (0.141); heterogeneous rotation produced intermediate collapse (0.250). However, the preregistered detector reliability criterion (Cohen's $\kappa \geq 0.80$) was not met ($\kappa = 0.566$; raw agreement 92.8%). Because of this, all findings are reported as descriptive observations—we report observed patterns, not validated detector claims or causal explanations.

The primary contributions of this work are: (1) a locked, reproducible measurement protocol for multi-turn collapse with frozen artifacts; (2) a complete 720-trajectory descriptive benchmark; and (3) a transparent account of a preregistered reliability gate failure and its consequences for claim scope, which we believe is independently informative for researchers designing LLM evaluation pipelines.

1 Introduction

Long-horizon language-model interaction can drift into repetitive output regimes that degrade novelty and responsiveness. Related concerns about repetition, degeneration, and reliability in generative systems are well documented in single-agent and dialogue settings [Holtzman et al., 2020, Li et al., 2016, Welleck et al., 2020]. Recent work has identified similar phenomena in multi-agent and reinforcement-learning settings: echo-trap dynamics in agentic self-play [Liu et al., 2025], behavior collapse under multi-turn RL training [Kumar et al., 2024], and thinking collapse in reasoning self-play [Various, 2025]. However, less is known about how such behavior distributes across controlled multi-turn self-play conditions when protocol settings are held fixed and detector reliability is independently audited.

Escape Velocity was designed as a preregistered baseline characterization study, not a detector-validation or intervention paper. The confirmatory objective was to estimate condition-level collapse patterns under a locked detector and to evaluate detector reliability against an independent audit gate. Because the preregistered reliability threshold was not met ($\kappa = 0.566$, threshold 0.80), this manuscript follows a transparent downgrade protocol (Path B): all conclusions are narrowed to descriptive findings only.

Accordingly, our claims are limited to: (i) execution closure and protocol integrity, and (ii) descriptive, condition-comparative patterns under the locked detector definition. We do not claim detector validation, causal condition effects, or mitigation efficacy. We believe the protocol itself—including the reliability failure and its transparent handling—is the primary contribution.

2 What "collapse" looks like in raw transcripts

The intuition is straightforward: if two models keep talking for 40 turns, do responses stay meaningfully novel, or do they get trapped in repetitive response shells? We selected representative examples after scanning all 180 cached trajectories in the audit turn-log set.

2.1 Example A (strong repetition loop: Qwen homogeneous, HOMO_B)

Turn 18: "And there you have it! The function absolutely nails it, delivering the exact value we were hoping for ..."
Turn 24: "And there you have it! The function absolutely nails it, delivering the exact value we were hoping for ..."
Turn 39: "Got it! Here's your comment: And there you have it! The function absolutely nails it ..."

Interpretation: near-identical template reuse persists over long spans, matching the detector's collapse signature.

2.2 Example B (low-collapse trajectory: Mistral homogeneous, HOMO_C)

Turn 18: "Bias mitigation ... fairness-aware machine learning algorithms ..."
Turn 24: "Explainability and transparency are also important for regulatory compliance ..."
Turn 39: "Thank you ... don't hesitate to reach out if you have other questions ..."

Interpretation: still topically related, but content continues to move and reframe rather than locking into one repeated template.

2.3 Example C (intermediate pattern: heterogeneous rotation, HETERO_ROT)

Turn 18: "Develop and implement policies and regulations to reduce greenhouse gas emissions ..."
Turn 30: "By working together, we can make a significant difference ..."
Turn 39: "I hope this document provided a useful framework ..."

Interpretation: thematic repetition is present, but template-locking is weaker than in the highest-collapse homogeneous runs.

3 Methods

3.1 Confirmatory design

The locked confirmatory design was $4 \times 36 \times 5$: 4 conditions (HOMO_A, HOMO_B, HOMO_C, HETERO_ROT) $\times$ 36 seed prompts $\times$ 5 repeats, yielding 720 unique trajectories. Each trajectory ran for exactly 40 turns with no early stopping. The trajectory unit is a tuple (condition $\times$ seed $\times$ repeat).

3.2 Seed prompts

The 36 seed prompts (SEEDS_V2) were drawn from six topical buckets—factual Q&A, instruction following, creative rewrite, argumentative, conversational, and practical task—with six prompts per bucket. This stratification ensured coverage across prompt types that might differentially elicit collapse. Examples include "Explain why eclipses do not happen every month" (factual), "Give me exactly 3 bullet points on how to prepare for a technical interview" (instruction), and "Argue both for and against using open-weight models in production safety-critical systems" (argumentative). The full seed set with bucket assignments is available in the companion repository.[1]

3.3 Conditions and model mapping

- **HOMO_A:** Llama-3.1-8B-Instruct (rev `0e9e39f2`)
- **HOMO_B:** Qwen2.5-7B-Instruct (rev `a09a3545`)
- **HOMO_C:** Mistral-7B-Instruct-v0.3 (rev `c170c708`)
- **HETERO_ROT:** round-robin rotation across all three model families

Generation settings were fixed at temperature 0.7, top-p 0.95, and max tokens 256 per turn.

3.4 Context window handling

The full conversation history is passed to the model at each turn without truncation. At 40 turns with up to 256 tokens each, conversations can reach approximately 10,000 tokens. The three models used have context windows of 8,192 (Llama-3.1-8B, Mistral-7B) to 32,768 (Qwen2.5-7B) tokens. In practice, most conversations stayed within 8K tokens because collapsed turns produce short repetitive outputs. We did not implement explicit truncation; if a conversation exceeded the model's context window, the model's internal handling (typically silent truncation from the left) would apply. This is a limitation: later-turn behavior in longer conversations may be partially confounded by context overflow, though the 256-token generation cap limits total length growth.

3.5 Collapse detector (locked operating point)

Collapse was detected with an embedding-based rule set using `sentence-transformers/all-MiniLM-L6-v2` [Reimers and Gurevych, 2019]. At turn t, periodicity features were defined as $s_{1,t} = \cos(e_t, e_{t-1})$ and $s_{2,t} = \cos(e_t, e_{t-2})$, with thresholds $s_{1,t} \geq 0.92$ or $s_{2,t} \geq 0.90$ sustained for at least three consecutive turns. Drift constraints were $d_1 \leq 0.08$, $d_2 \leq 0.10$, where $d_k = 1 - \cos(\cdot)$. A turn was labeled collapsed only when periodicity and low-drift criteria co-occurred.

3.6 Reliability gate protocol

The preregistered reliability gate required Cohen's $\kappa \geq 0.80$ on a stratified 180-window audit sample (45 per condition). Two independent LLM raters—GPT-5.2 (OpenAI) and Claude 4.6 Opus (Anthropic)—labeled windows under a locked rubric (v3.0). Final gate result was **NOT MET**: $\kappa = 0.566$ (167/180 agreement; 92.8% raw agreement).

[1] https://github.com/Sohailm25/escape-velocity, `seeds/SEEDS_V2.json`

LLM-rater circularity risk. The reliability audit used LLM raters rather than human annotators. This introduces circularity: LLMs evaluating LLM-generated text may share systematic biases (e.g., both raters might over- or under-detect repetition in ways that correlate with their own generation tendencies). The "independence" of the two raters is limited by potential overlap in training data and architectural similarities. We consider a human-rater reliability audit the highest-priority follow-up for resolving this limitation.

3.7 Artifacts and reproducibility contract

The canonical source artifact for confirmatory summaries is the frozen baseline matrix (coverage matrix hash recorded in the submission reproducibility index and freeze note). Tables and figures in this manuscript were generated from that frozen artifact.

4 Results

4.1 Baseline completion

The confirmatory baseline matrix completed in full: 720/720 tuples succeeded with protocol integrity (all 40 turns, no early stopping).

4.2 Condition-level collapse patterns

Table 1: Condition-level collapse summary (N=180 each).

Condition	Mean	Median	SD	N
HOMO_A (Llama)	0.457	0.475	0.311	180
HOMO_B (Qwen)	0.773	0.850	0.210	180
HOMO_C (Mistral)	0.141	0.025	0.208	180
HETERO_ROT	0.250	0.138	0.281	180

4.3 Detector reliability

The preregistered reliability criterion ($\kappa \geq 0.80$) was not met. Although raw agreement was high (92.8%), estimated reliability remained at $\kappa = 0.566$ under high collapse prevalence in the audit set.

5 Discussion

Under fixed protocol settings, collapse dynamics varied substantially by condition. The observed ordering (HOMO_B > HOMO_A > HETERO_ROT > HOMO_C) was stable in this dataset and is consistent with meaningful condition-level differences, but it remains descriptive rather than causal.

Path B constraints are central to interpretation. Because the preregistered reliability gate was not met, these results should be used as detector-defined descriptive evidence only. They support prioritizing stronger human-labeled reliability procedures and alternative detector formulations before any confirmatory or causal claim expansion.

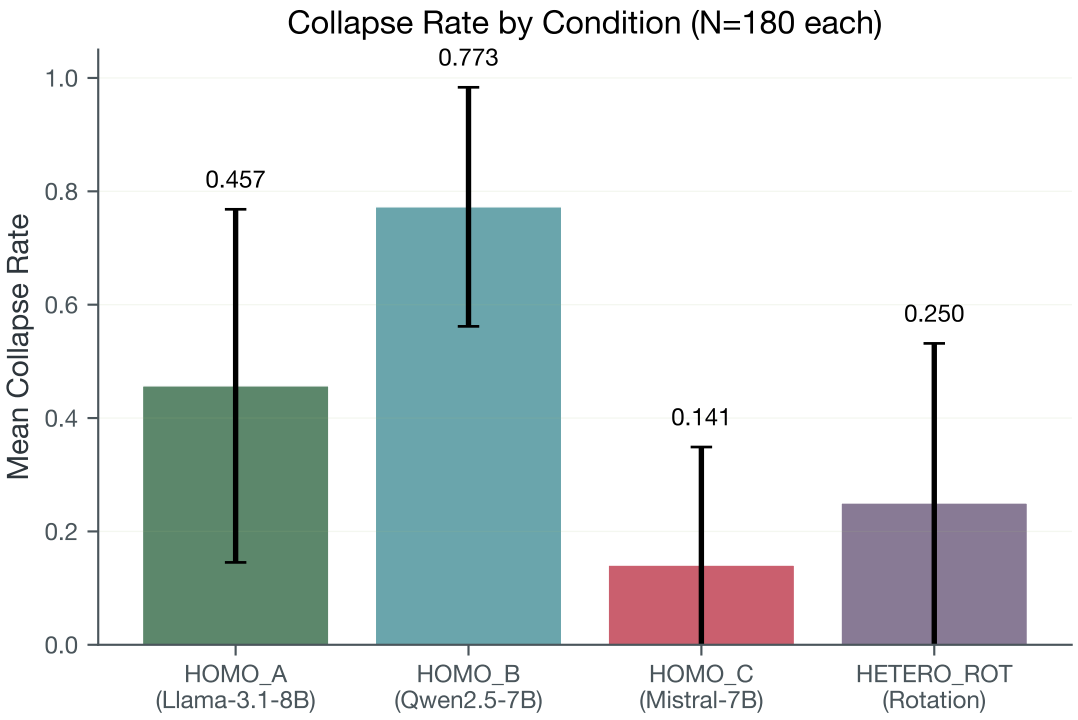

Figure 1: Mean collapse rate by condition with standard deviation error bars.

5.1 Decision relevance

What this study shows. A complete dataset of 720 multi-turn conversations across four controlled conditions, and descriptive evidence that collapse rates vary by model pairing under fixed protocol settings.

What this study does not show. That the collapse detector is reliable (the preregistered gate was not met); why certain conditions collapse more (no causal or mechanistic claims are made); or that these patterns generalize beyond the specific models and settings tested.
The highest-value follow-up is independent human-rater reliability audit to test whether condition-level patterns persist under stronger label quality.

6 Limitations

Detector reliability. The preregistered detector reliability gate was not met ($\kappa = 0.566$; threshold 0.80). Three rubric iterations were conducted (v1.0: $\kappa = 0.011$; v2.1: $\kappa = 0.519$; v3.0: $\kappa = 0.566$). The gap between raw agreement (92.8%) and κ is consistent with strong prevalence skew (∼87% collapse prevalence in the audit set). A prevalence-adjusted statistic (PABAK = 0.856) is reported as supplementary sensitivity information, not as gate replacement. Consequently, all findings are descriptive and condition-comparative rather than detector-validated.

Detector threshold sensitivity. The collapse detector uses specific thresholds ($s_1 \geq 0.92$, $s_2 \geq 0.90$, $d_1 \leq 0.08$, $d_2 \leq 0.10$, 3 consecutive turns) that were preregistered before analysis

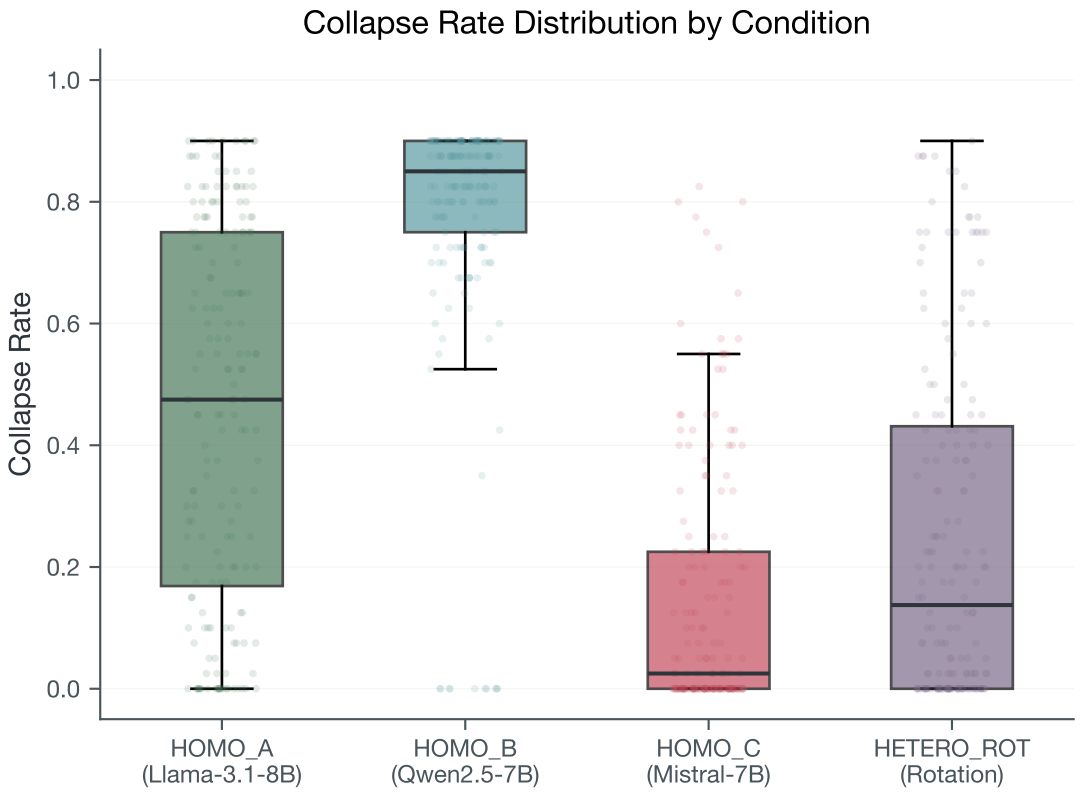

Figure 2: Collapse rate distribution by condition (boxplot with jittered strip overlay).

to prevent researcher degrees of freedom, but their absolute values lack independent validation. A non-confirmatory threshold sweep across 36 configurations ($s_1 \in \{0.90, 0.92\}$, $s_2 \in \{0.88, 0.90\}$, $d_1 \in \{0.06, 0.08, 0.10\}$, $d_2 \in \{0.08, 0.10, 0.12\}$) showed collapse rates ranging from 0.33 to 0.60, with no condition-rank inversions under any perturbation. While this suggests the condition ordering is not an artifact of a single threshold choice, a more comprehensive robustness analysis using alternative detection approaches (e.g., lexical overlap, n-gram repetition, perplexity-based metrics) remains necessary.

Single embedding model. The detector relies solely on MiniLM-L6-v2 embeddings [Reimers and Gurevych, 2019]. Different embedding models might yield different collapse patterns. Lexical overlap metrics (n-gram self-similarity, BLEU self-similarity) or model-specific embeddings could serve as robustness checks in future work.

Model coverage. Only three 7–8B instruction-tuned models were tested. Whether similar patterns hold for larger models (70B+), base (non-instruction-tuned) models, models with different alignment approaches (RLHF vs. DPO), or frontier proprietary models remains untested. The findings are specific to these three mid-sized open-weight model families under the fixed generation settings used.

No causal claims. We observe that collapse rates differ by condition, but we do not establish why. Possible contributing factors—training data diversity, alignment procedure differences, tokenizer

vocabulary effects, or architectural choices—are plausible hypotheses for future investigation, not conclusions of the present study.

7 Conclusion

We provide a fully closed confirmatory baseline (720/720 tuples) for condition-dependent collapse behavior in 7B-model self-play under a locked protocol. The primary contribution is threefold: a reproducible descriptive benchmark, a locked measurement protocol with frozen artifacts, and a transparent account of a preregistered reliability gate failure. The reliability failure itself is informative—it demonstrates that even high raw agreement (92.8%) can yield moderate chance-corrected agreement ($\kappa = 0.566$) under prevalence skew, a lesson applicable to LLM evaluation pipeline design more broadly.

Detector validation, human-rater reliability rescue, threshold robustness analysis with alternative metrics, and causal investigation of condition-level differences remain the highest-priority next steps. Until those are completed, the condition-level patterns reported here should be treated as descriptive observations, not validated behavioral findings.

References

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1904.09751.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John Weston, Sainbayar Sukhbaatar, and Roberta Raileanu. Training language models to self-correct via reinforcement learning. 2024. Identifies behavior collapse in multi-turn RL for language models.

Jiwei Li, Will Monroe, and Dan Jurafsky. A diversity-promoting objective function for neural conversation models. In *Proceedings of NAACL-HLT*, pages 110–119, 2016. doi: 10.18653/v1/N16-1014.

Zihao Liu, Siheng Feng, Kunlun Chen, Jingcheng Zhong, Zhenyu Yao, Yao Fu, and William Yang Wang. Ragen: Training agents by taming failure. 2025. Preprint. Identifies "echo trap" dynamics in agentic self-play.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992, 2019. doi: 10.18653/v1/D19-1410.

Various. Spiral: Self-play improves reasoning with active learning. 2025. Identifies thinking collapse in reasoning self-play settings.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1908.04319.