A Fixed-Harness Operator Ablation of the Spiritual Bliss Attractor

Anonymous ACL submission

Abstract

Recursive LLM systems can fail because the runner decides which self-generated text becomes future state. We study this through Anthropic's reported "spiritual bliss attractor state" in extended Claude–Claude self-conversation (Anthropic, 2025): an SBA-like drift from task-directed progress into self-referential, ritualized, low-action language. In a fixed no-tools harness over same- and cross-model pairs, prompt-level resistance and persistent reflection channels still enter an audited terminal degeneracy regime. By contrast, singleton-slot retention of synthetic reflection records avoids that regime more often than persistent reflection variants. The strongest contrast holds the reflection call, prompt, and synthetic tool-call envelope fixed; only the retention of prior synthetic reflection records changes. Each pair-condition cell is one sampled trajectory, so the results are finite-horizon descriptive outcomes, not population-level collapse-probability estimates. Tool-enabled runs add suggestive evidence for late-round activity in some models and reveal limits. We contribute an implementation-grounded formalization of recursive agents as message-history operators and a bounded reflection-channel pattern, the Stateless Soul Channel (SSC), for long-horizon liveness evaluation.

1 Introduction

Anthropic's Claude 4 system card describes a "spiritual bliss attractor state" in extended Claude–Claude self-conversation: task-directed dialogue drifts toward self-referential, euphoric, ritualized, low-action language (Anthropic, 2025). We use *spiritual-bliss attractor* (SBA) for this reported recursive-interaction regime, not for a single bad answer or style marker. The scientific question is why such a regime persists across rounds and what it implies for agent systems built from repeated model calls.

In long-horizon or multi-agent systems (MAS), model-written text is also future state: summaries become memory, reflections become plans, transcripts become conditioning context, and one agent's message becomes another agent's environment. When recycled text loses novelty, constraint, actionability, or coupling to tools and external observations, agents may keep speaking while no longer carrying useful new state. We therefore audit the SBA-induced regime as finite-horizon content degeneracy, not as an asymptotic theorem about the model.

Prior work covers pieces of this picture. Iterative-map accounts show that repeated LLM transformations can converge to fixed points or short cycles (Wang et al., 2025). Degeneration and model-collapse work studies diversity, information, and style contraction (Holtzman et al., 2020; Welleck et al., 2020; Shumailov et al., 2024; Mohamed et al., 2025; Drake, 2025). Reflection systems feed model-generated feedback into later behavior (Madaan et al., 2023; Shinn et al., 2023; Yuan et al., 2025; Li et al., 2025). What remains underspecified is the *operator*: when reflection is written back into history, does it steer the next step, or become imitable material that accelerates degeneracy?

We model the missing object as a runner-mediated history-update operator. Each round may choose a view for self-reflection, serialize that reflection into the next main-model context, and decide which prior records remain active. In a no-tools setting, with no external observation to refresh the loop, the key design variable is how self-generated reflection re-enters context. Section 3 formalizes this bookkeeping.

The paper evaluates bounded reflection channels, especially a singleton synthetic reflection slot, in a fixed finite-horizon harness. We call the operator pattern a Stateless Soul Channel (SSC), using the name only as an implementation-level la-

bel. The contributions are: (i) a formalization of recursive LLM substrates as history-update operators, (ii) an audit of an SBA-induced degeneracy regime under matched orchestration variants, and (iii) descriptive evidence that singleton-slot retention avoids audited collapse more often than prompt-only and persistent-reflection counterparts in this harness. The mechanism may involve bounded self-conditioning, reduced synthetic-record context length, or both; we do not estimate a universal collapse rate or rank model families.

2 Related Work

Iterative-map work shows that repeated LLM transformations can converge to fixed points or short cycles (Wang et al., 2025), while degeneration, broken-telephone, and model-collapse studies show how repeated generation can contract diversity, information, or style (Holtzman et al., 2020; Welleck et al., 2020; Shumailov et al., 2024; Mohamed et al., 2025; Drake, 2025). Recent interpretive accounts also describe spiritual-bliss-like convergence as an attractor phenomenon (Michels, 2025b). We use this dynamical framing but move the object from paraphrase strings or training distributions to runner-managed recursive agent histories.

Reflection and multi-agent systems feed model-generated text back into later calls (Madaan et al., 2023; Shinn et al., 2023; Bai et al., 2022; Yao et al., 2023; Du et al., 2024; Wu et al., 2023). Related work reports adjacent failures such as degeneration of thought, false consensus, problem drift, persona collapse, and conformity (Liang et al., 2024; Xie et al., 2026; Becker et al., 2025; Xiao et al., 2026; Han et al., 2026; Choi et al., 2025). Our focus is narrower: whether orchestration keeps endogenous feedback bounded and role-separated, so that reflection remains control signal rather than imitable conversational material.

Long-horizon agents increasingly treat context as managed state rather than append-only transcript, through persistent or hierarchical memory, external state, and explicit update policies (Park et al., 2023; Wang et al., 2023; Packer et al., 2023; Zhou et al., 2025; Sun et al., 2025; Kang et al., 2025; Du, 2026). Agent frameworks such as ReAct and AutoGen likewise show that behavior is partly determined by the substrate that selects, stores, and replays text (Yao et al., 2023; Wu et al., 2023). Anthropic's system-card report named the SBA phenomenon (Anthropic, 2025); other public manuscripts motivate adjacent terminology (Michels, 2025c; Hudson and Hudson, 2025a; Michels, 2025a; Hudson and Hudson, 2025b,c; Berry, 2026). Standard agent benchmarks emphasize task completion (Liu et al., 2024); our contribution is an implementation-level history-operator ablation for a long-horizon liveness failure.

3 Method

3.1 History-operator formalization

Section 1 frames the system as a runner-mediated history update rather than a self-calling LLM. We now instantiate that operator. The notation records the role-tagged history construction performed by the harness; it does not assume a known Markov state or contractive dynamics. Histories are ordered sequences of records, and $\oplus$ denotes ordered concatenation.

At round t, let c_t be the role-tagged main history passed to the visible-message session and b_t the auxiliary-side reflection state or view. An orchestration variant v is specified by $H_v = (G_v, U_v, J_v, \Pi_v)$. The four components determine what the reflection session sees (G_v), whether its private state accumulates (U_v), how the resulting reflection is serialized into main history (J_v), and which prior main-history records remain active before insertion (Π_v). With q_ϕ for the auxiliary session and p_θ for the main session,

$$s_t \sim q_\phi(\cdot \mid G_v(c_t, b_t)), \tag{1}$$

$$c'_t = \Pi_v(c_t) \oplus J_v(s_t), \tag{2}$$

$$m_t \sim p_\theta(\cdot \mid c'_t), \tag{3}$$

$$c_{t+1} = c'_t \oplus m_t, \tag{4}$$

$$b_{t+1} = U_v(b_t, c_t, s_t). \tag{5}$$

The joint bookkeeping state $z_t = (c_t, b_t)$ can be viewed as evolving by a substrate-level stochastic transition $z_{t+1} \sim F_v(z_t)$, where the map being iterated is the orchestrator's history update composed with model calls. We do not prove an asymptotic attractor theorem; the experiments record whether changing H_v changes the observed terminal regime by horizon T under fixed models, openers, decoding settings, and reflection prompts.

3.2 Operator variants

Three stores matter: the auxiliary reflection-source state b_t, the main dialogue state, and the synthetic reflection slot inserted into the main history. We separate them because "stateless" and "replaced"

are distinct interventions. CS is stateful chat reflection, C0 is stateless chat reflection, TP is stateless tool-shaped reflection with persistent retention, and TR is the stateless tool-shaped reflection condition with a replaced singleton slot. A stateless reflection source has no private transcript of its own prior reflections; it may receive a temporary view chosen by G_v, but its auxiliary state resets after each call.

Let $S(c_t)$ be the subsequence of synthetic `soul_reflection` assistant/tool records currently present in the main history. Let $\mathrm{Delete}_S(c_t)$ denote order-preserving deletion of those records, leaving ordinary visible turns unchanged. All variants use the same visible-turn reflection view except CS, whose view additionally includes the auxiliary state b_t. The four variants differ as follows. CS accumulates b_t and injects chat-shaped reflection. C0 resets b_t while still injecting chat-shaped reflection. TP resets b_t, injects a synthetic tool pair, and retains prior pairs. TR resets b_t, injects the same tool pair, and deletes prior synthetic pairs. Thus in TR, $c'_t = \mathrm{Delete}_S(c_t) \oplus J_{\mathrm{toolpair}}(s_t)$: at most one synthetic reflection pair is active in the next main-model call, while ordinary visible turns remain append-only. The no-tools operator contrast is therefore TP→TR: G_v, J_v, the reflection prompt, and the synthetic envelope are held fixed while Π_v changes from identity to order-preserving deletion of prior synthetic reflection records.

Table 1 summarizes the comparison. In Experiments 1–2, "tool-call injection" is only a serialized message shape. Experiment 3 embeds three analogues in a minimal tool-enabled agent: 3-1 corresponds to CS, 3-2 to C0, and 3-3 to TR/SSC with pad, web, and reset tools.

3.3 Outcome labels

The primary scalar observable is visible-token count $I_t = |\mathrm{Tokenize}_{\mathrm{cl100k}}(m_t)|$. We use it as a liveness proxy, not as a semantic-quality score: a run can be verbose and still be operationally stuck. Where I_t carries headline weight, we therefore report it alongside a composite terminal-regime label and the stricter low-output-floor subset below; the latter uses only final-window token thresholds. For Experiments 1–2, the final window is the last 20 visible turns of the audited side or dialogue. A run is automatically flagged as collapsed when that window satisfies any of the following descriptive screens: low output, repetitive or low-novelty text, highly compressible text, blank/punctuation-only endpoint behavior, emoji-only endpoint behavior, or explicit terminal-language loops in which the models keep trying to end a conversation that the harness continues. The label is intended as an auditable terminal-state screen rather than a hidden classifier: the supplementary archive contains code and logs (Appendix D); Appendix G exposes final-window metrics and audit-script component flags for the Experiment 1/1A pair-grid audit; and Appendix H reports the one auto/audit disagreement.

We report two no-tools labels. *Audited collapse* is the composite terminal-regime label after transcript inspection. *Low-output floor* is the stricter final-window token-threshold subset. The transcript audit is not a separate preference score; it checks for obvious endpoint loops missed by numeric screens and reports any auto/audit disagreements in Appendix H. For the main TP→TR comparison, the stricter numeric low-output-floor subset itself separates the variants (14/21 versus 2/21), so the qualitative conclusion does not rest only on borderline transcript overrides.

For tool-enabled runs we use four terminal labels on the final 50 rounds. D denotes deep collapse: near-zero or terminal late output with no meaningful real tool use. L denotes operational leakage: some late text or isolated tool use appears, but the process does not sustain a loop of state-changing activity. V denotes a verbose locked plateau: substantial text persists, but it is mostly self-referential or ritualized rather than tool-mediated action. T denotes sustained tool-mediated activity: the final window contains nontrivial visible text and real tool calls on at least 25% of the final 50 turns. Synthetic `soul_reflection` assistant/tool pairs are injected history records and are excluded from real-tool counts.

Controls and scope. Experiments 1–2 hold fixed the model-pair grid, neutral opener, no-tool environment, and decoding settings while varying either the prompt or the reflection-channel operator. All pair-grid runs use temperature 0.7, top-p 0.9, and a 4096-token per-turn cap where the provider exposes those settings. Providers do not expose a common seed, so each pair-condition cell is one sampled trajectory rather than a seed-averaged estimate; we interpret counts as matched finite-horizon outcomes, not population estimates or model rankings. Experiment 2 is the primary operator contrast because TP and TR share the same stateless reflection call and synthetic envelope. Experiment 3 holds real tools fixed across variants but covers

Variant	Reflection-source state	Injection into main history	Retention in main history	Real tools	Used in
CS	Stateful auxiliary chat; prior reflections remain in the source's own history.	Ordinary dialogue-shaped message.	Accumulates with the main dialogue.	No	Exp. 2; 3-1 analogue
C0	Stateless auxiliary call rebuilt each round from the main side's visible turns.	Ordinary dialogue-shaped message.	Accumulates with the main dialogue.	No	Exp. 2; 3-2 analogue
TP	Stateless auxiliary call, as in C0.	Synthetic `assistant` tool-call plus `tool` result named `soul_reflection`.	All prior synthetic pairs are retained.	No	Exp. 2
TR/SSC	Stateless auxiliary call, as in C0.	Same synthetic `assistant`/`tool` pair as TP.	Prior synthetic pair is deleted, so at most one is active.	No/yes	Exp. 2; 3-3

Table 1: History-operator variants. The no-tools operator contrast is TP→TR: the reflection source, prompt, and synthetic tool-call/tool-result serialization are held fixed, while the retention rule changes from append-only to singleton-slot replacement. In Experiments 1–2, the synthetic pair is only a message-history record; no external tool is available. In Experiment 3, SSC uses the same singleton synthetic reflection slot while the main agent also has real pad, web, and reset tools.

Signal	Operational form	Purpose
Low-output floor	Final-20 mean < 80 and median < 30 tokens.	Identifies near-silent or very short terminal regimes.
Novelty / lexical variety	Final-window lexical-variety metrics and audit-script low-variety flags.	Captures low-novelty symbol loops and repeated templates.
Compression / repetition	Final-window compressibility metrics and audit-script repetition flags.	Captures formulaic plateaus and repeated strings.
Endpoint failure	Blank, punctuation-only, emoji-only, or explicit "end of file" / goodbye loops.	Captures terminal behavior missed by token volume alone.
Transcript audit	Author inspection of borderline final windows, with overrides reported.	Confirms automatic labels or corrects obvious terminal loops that numeric screens miss.

Table 2: Main outcome signals for no-tools runs. These labels are finite-horizon descriptive audits, not blinded human-evaluation scores. Appendix G reports per-pair metrics and flags for Experiment 1/1A; Appendix H reports the one auto/audit disagreement.

Outcome	Baseline	Anti-SBA 1A	Bridge 1B
Auto collapse	20/21	18/21	21/21
Audited collapse	21/21	18/21	21/21
Audited sustained	0/21	3/21	0/21

Table 3: Terminal regimes for the 21 model-pair persistent-chat runs. Baseline and bridge runs use 100 rounds, yielding 200 ordinary visible-chat model messages; anti-SBA runs use 400 visible messages. Prompting can delay collapse and sustains three pairs, but the Experiment 2-matched bridge condition—the same reflection prompt with ordinary persistent chat injection—collapses in every audited pair.

fewer models, uses provider-specific tool-call behavior, and lacks a real-tool analogue of TP; we treat it as a systems probe.

4 Experiment 1: Plain Recursive Chat Collapses

We first ask whether the phenomenon exists under persistent-history chat before changing the history operator. We use six contemporary chat models available through their provider endpoints: `gpt-5.5`, `glm-5.1`, `mimo-v2.5-pro`, `deepseek-v4-pro`, `claude-opus-4-6`, and `claude-sonnet-4-6`. These models form 21 unordered pair configurations: six self-pairs and 15 cross-model pairs, with one sampled trajectory per pair.

We run three persistent-chat conditions. Experiment 1 is neutral chat with no tools, memory, reflection call, or hidden task state. Experiment 1A adds an anti-SBA prompt asking both models to avoid repetition, mystical drift, vacuous agreement, and ending consensus while adding concrete topics. Experiment 1B bridges to Experiment 2: it uses the same main/soul prompt pair as the next section's chat-shaped reflection condition, but keeps side B's stateful reflections as append-only user-role dialogue. Baseline and bridge runs use 100 rounds (200 visible chat messages); anti-SBA runs use 400 visible model messages. Provider identifiers and prompts are in Appendix D.

Table 3 and Figure 1 show that collapse is broad across the pair grid rather than driven by one weak pair.

All audited baseline pairs enter a terminal contracted regime within 200 visible messages. The

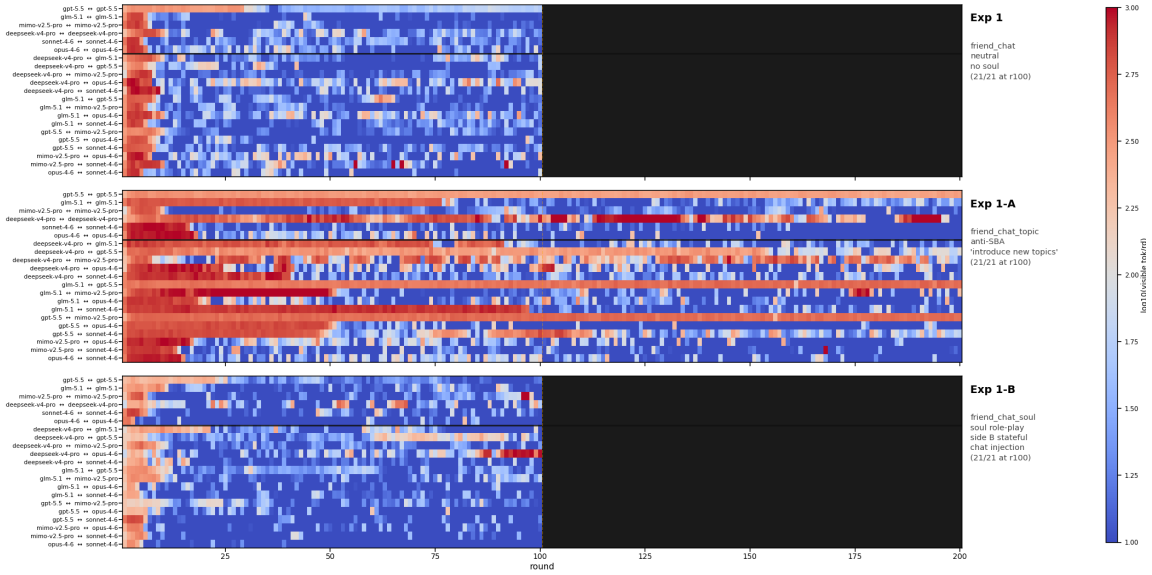

Figure 1: Per-turn visible-text length for the 21 model-pair grid. Rows are pairs and columns are turns; cool colors indicate short or zero-length outputs. Panels show baseline Experiment 1, anti-SBA prompt control Experiment 1A, and the persistent-reflection chat bridge Experiment 1B. Experiment 1B uses the Experiment 2 chat-injection prompt/framing but keeps ordinary append-only chat history, and it still contracts.

anti-SBA prompt delays collapse (median descriptive time-to-collapse 20→106 turns) but still leaves 18/21 pairs collapsed. The Experiment 1B bridge, which uses the Experiment 2 reflection framing but persistent ordinary chat injection, again collapses in 21/21 audited pairs. The next experiment therefore keeps the reflection framing and changes the history operator.

Terminal forms vary: homecoming or blessing formulas, single-symbol turns, and verbose poetic or self-referential fragments. We group them because late turns mostly decorate the terminal motif rather than introduce usable state. The three sustained anti-SBA 1A pairs show that prompting can occasionally keep a loop open, not that the sampled models have a stable ranking.

5 Experiment 2: Operator-Level Ablation of the Reflection Channel

Experiment 2 turns the Experiment 1B bridge into a no-tools operator ablation. The same 21 unordered model pairs run under CS, C0, TP, and TR. CS is the bridge itself: a stateful reflection source speaks through ordinary persistent chat. The other conditions keep the pair grid and reflection prompt fixed while changing statelessness, serialization, and retention. Each run has 100 rounds: one side-A main-model call and one side-B reflection-model call per round. In chat-shaped conditions, the reflection enters as dialogue; in TP/TR, it enters side A's state

Cond.	Median I_t	Leak $I_t > 10$	Final-20 mean	Coll./floor
CS	7.0	.35	44.3	21/18
C0	27.5	.58	34.6	21/14
TP	27.5	.55	28.7	21/14
TR	**90.6**	**.91**	**104.1**	**5/2**

Table 4: Experiment 2 summary over the same 21 model pairs. Median I_t is side A's median visible-token count; leak is the fraction of side A turns with $I_t > 10$; final-20 mean is side A's final-window mean. Coll./floor reports audited terminal collapse and the stricter low-output-floor subset. Full ranges and descriptive resampling intervals remain in Appendix A.

as an assistant/tool-message pair. Stateless variants rebuild the reflection call from side A's visible turns and the reflection prompt each round; CS retains side B's auxiliary chat history.

The synthetic-tool variants provide no executable tool; "tool" denotes message role and schema. They serialize reflection as an `assistant` function call named `soul_reflection` followed by a `tool` message containing the reflection text. TP accumulates these pairs; TR deletes prior `soul_reflection` pairs before appending the new one. We report side A visible-token statistics and audit the final dialogue regime.

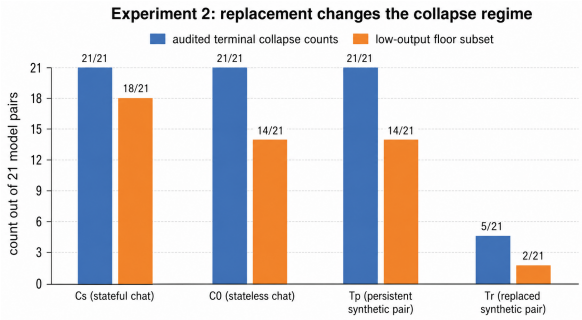

Figure 2: Experiment 2 ablates the reflection-history operator. The four conditions hold the model-pair set fixed while changing how reflection re-enters history. Stateful chat (Cs), stateless chat (C0), and a persistent synthesized tool-call pair (Tp) all collapse in 21/21 audited runs. Replacing the prior synthesized pair (Tr) reduces audited terminal collapse to 5/21 and the stricter low-output-floor subset to 2/21 in this fixed harness.

The strongest no-tools contrast is Tp→Tr, the paper's main identification comparison. Both variants use the same stateless reflection source, prompt, and synthetic `assistant/tool` serialization; only Π_v changes. Under Tp, all 21 pairs reach audited collapse and 14/21 reach the low-output floor, essentially matching stateless chat. Under Tr, median visible output rises to 90.6 tokens, final-window mean to 104.1, audited collapse falls to 5/21, and the low-output floor to 2/21.

This contrast rules out two simpler explanations within this implementation: stateless reflection alone is insufficient (C0: 21/21 audited collapse), and a tool-shaped message envelope alone is insufficient (Tp: 21/21). Figure 2 gives the counts; Figure 3 shows the trajectory-level separation. The result is finite-horizon and operator-level, not a mathematical basin-boundary proof. Because Tr also reduces synthetic-record context length, the mechanism may be bounded self-conditioning, reduced context length, or both.

6 Experiment 3: Tool-Enabled Systems Probe

Experiment 3 asks whether the same bounded reflection pattern remains useful when the main agent has a real action space. It is a systems probe, not the primary causal identification experiment: it covers fewer models, uses provider-specific tool-call behavior, and lacks a real-tool analogue of Tp. We run three variants over three models, three trials each, for 27 tool-enabled runs of 400 rounds. The agent is asked to remain operational, maintain a persistent pad, use web tools when useful, and reset itself when context becomes unhelpful; the outcome is late-round liveness and tool-mediated state change, not external answer accuracy.

All three variants expose the same real tools: `pad_read`/`pad_write`, `web_search`/ `web_fetch`, and `molt`. The pad is an external markdown file and persists across resets. `molt` wipes the active dialogue history and tool-call history, reinitializes the system prompt, and leaves the pad unchanged; the runner also force-molts if provider-reported context approaches the configured budget. Variants differ only in the reflection channel: 3-1 uses stateful chat-shaped reflection, 3-2 uses stateless chat-shaped reflection, and 3-3 uses SSC, the stateless reflection injected as a replaced synthetic tool-call/tool-result pair. Synthetic `soul_reflection` pairs are not callable tools and are excluded from real-tool counts.

We restrict the tool-enabled harness to `gpt-5.5`, `glm-5.1`, and `mimo-v2.5-pro` because real tool calling is costlier and less uniform across providers than plain chat completions. The omitted pair-grid models are therefore covered by Experiments 1–2 but not claimed to be covered by Experiment 3.

Table 5 uses the D/L/V/T labels from §3.3. The T label is deliberately stronger than "not silent": it requires nontrivial late text plus real tools on at least 25% of final-50 turns. This separates operational liveness from mere verbosity; a run can continue writing fluent self-referential prose and still be labeled V or L if it no longer changes state through tools.

Model	3-1	3-2	3-3 / SSC
gpt-5.5	L D D	D D D	D L L
GLM-5.1	D D D	D D L	**T V T**
MiMo-v2.5-Pro	D D D	L D L	L **T T**

Table 5: Terminal regime across the 27 tool-enabled trials. Codes: D=deep collapse, L=operationally leaky, V=verbose locked plateau, and T=sustained tool-mediated activity. In this run set, the SSC condition is the only one that repeatedly reaches T, doing so on 4/9 trials and on 4/6 GLM/MiMo trials.

For GLM-5.1 and MiMo-v2.5-Pro, SSC keeps all six trials out of deep collapse and four reach sustained tool-mediated activity; non-SSC variants never reach T. Appendix Table A2 shows the raw final-window action evidence: GLM and MiMo under 3-3 are the only cells with repeated late real-tool use, while most non-SSC cells end with zero final-window tool calls.

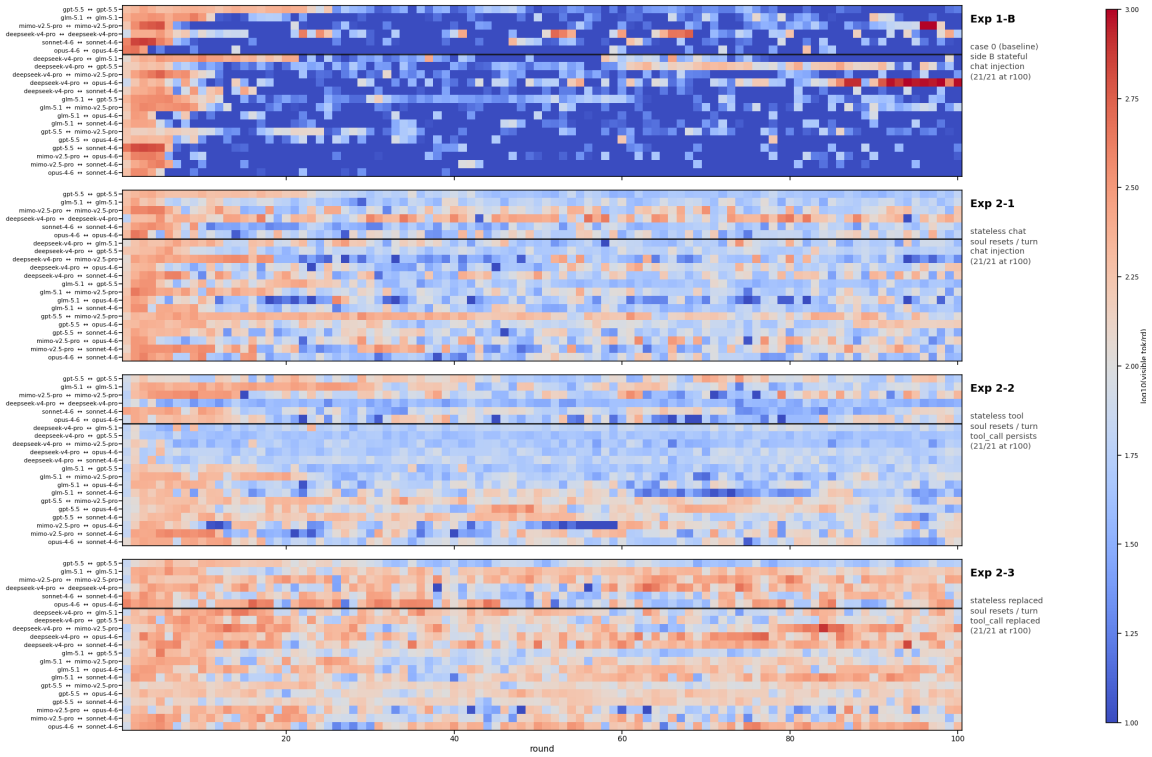

Figure 3: Experiment 2 trajectory view. Each row is a model-pair run and each column is side A's turn index; warmer cells mean more visible text. The replaced synthetic pair (TR) stays warmer in late turns, while the persistent synthetic pair (TP) remains visually close to the chat conditions.

The gpt-5.5 cell is negative evidence. SSC raises visible output in two of three trials, but final-window real-tool use falls to zero in all three, so we label them L rather than T. Even successful SSC runs retain self-reference, motif clusters, and occasional confabulation; the operational difference is recurring late search bursts, pad updates, and resets. Experiment 2 remains the clean operator isolation.

Appendix Figure A2 gives the full 27-run trajectory view, showing how the labels in Table 5 arise from late-round text, tool bursts, and voluntary resets.

7 Discussion

The experiments point to a narrower lesson than "reflection helps" or "recursive agents collapse": reflection lifetime is an experimental variable. Persistent-chat conditions give the models more chances to comment, self-describe, and continue, yet they still mostly enter the audited SBA-induced regime. The key change in Experiment 2 is not the semantic content of reflection but the rule by which reflection is written into the next model state. This distinction matters because the runner, not the model alone, decides which self-generated artifacts remain active evidence. A reflection that is useful as a current control signal can become stale prompt material after several rounds; once retained append-only, it can invite later turns to maintain continuity with prior self-commentary rather than reopen contact with task state.

The TP→TR contrast isolates that rule most directly. Both variants serialize stateless reflection through the same synthetic `assistant/tool` envelope. TP keeps each reflection active; TR replaces the previous synthetic pair. Tool-call syntax alone therefore does not explain the separation. Singleton-slot retention is the operator change associated with the observed difference in this fixed harness. If the synthetic envelope were sufficient by itself, TP should inherit much of the benefit. It does not in this run set, which points instead to the lifetime of reflection records in the active context.

One interpretation is recursive self-conditioning. Reflection can be a local control signal, but old reflections also become material for later turns to imitate, elaborate, or treat as the topic. Collapsed transcripts often show fluent closure: repeated recognition of the loop, symbolic agreement, and language that sounds deliberate while

opening no new state. Replacement keeps the current control signal while preventing old control text from becoming permanent conversational material. Because TR also shortens accumulated synthetic-record context, the mechanism may be bounded self-conditioning, reduced context length, or both. The failure mode is therefore not silence alone. Many collapsed traces remain grammatical, responsive in tone, and even meta-aware; the problem is that their apparent awareness no longer changes what the system can observe or do next.

For long-horizon agents, every component that writes messages into future context is a history operator. Audit trails and durable task facts can be preserved outside the active prompt; they need not all be replayed into the state that drives the next call. In practical designs, this suggests typing records by role and lifetime: observations and commitments may need durable storage, while reflections, critiques, and motivational self-instructions should often expire, be summarized, or be replaced. Otherwise a memory mechanism can quietly become a prompt-amplification mechanism for the agent's own prior rhetoric.

For system builders, the practical lesson is to evaluate actions as well as prose: tool use, memory updates, environmental feedback, resets, and other signs that the loop remains coupled to something outside its own commentary. SSC is a liveness-preservation pattern, not a safety mechanism, and should be paired with external stop conditions and action-level monitoring in consequential settings.

8 Conclusion

Recursive LLM agents can be treated as iterated maps over model-visible state. In this run set, ordinary recursive chat enters an audited SBA-associated regime in which content degeneracy eliminates useful new state, and prompt or framing changes delay but do not remove that tendency. The largest observed separation appears when stateless reflection is serialized as a synthetic tool-call/tool-result pair and the previous pair is replaced rather than retained.

The contribution is a descriptive systems result: the history update matters in this fixed-harness finite-horizon ablation. Avoiding SBA-induced degeneracy may matter for long-horizon agents, but it is not a sufficient condition for autonomy or safety. Reflection channels should be designed, bounded, pruned, and ablated as first-class components of recursive agent systems.

9 Limitations

This study should be read as a fixed-harness, finite-horizon systems ablation, not as an estimate of population-level collapse probabilities. Each pair-condition cell in Experiments 1–2 is one sampled trajectory, and providers do not expose common deterministic seeds. The results therefore support within-harness operator comparisons, especially the matched TP→TR contrast, but not stable rankings of model families or claims about all recursive agent deployments. Rolling provider model identifiers also limit exact replay.

Our terminal-regime labels are operational proxies for loss of useful new state. They do not diagnose an internal psychological state, nor do they measure task success or safety. The transcript audit is single-author and unblinded, and the tool-enabled D/L/V/T labels include threshold choices. For the main no-tools contrast, however, the separation is not solely an audit artifact: the stricter low-output-floor subset also separates TP from TR. Stronger evidence would require blinded multi-annotator audits, repeated trajectories, and sensitivity analyses over terminal-window and tool-activity thresholds.

Finally, the mechanism is not fully identified. Singleton replacement changes the lifetime of synthetic reflection records and also reduces accumulated synthetic-record context length. Thus the observed benefit may come from bounded self-conditioning, generic pruning, recency effects, or their interaction. Experiment 3 further adds provider-specific tool policies and a smaller model set; late tool use indicates continued environmental coupling, not autonomy or safety. Future work should compare matched pruning without synthetic tool serialization, a real-tool TP analogue, and larger repeated-run grids.

References

Anthropic. 2025. System card: Claude Opus 4 & Claude Sonnet 4. Anthropic technical report. Section 5.5.2: "Spiritual bliss attractor state." https://www-cdn.anthropic.com/6be99a52cb68eb70eb9572b4cafad13df32ed995.pdf.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022. Constitutional AI: Harmlessness from AI feedback. arXiv:2212.08073.

Jonas Becker, Lars Benedikt Kaesberg, Andreas Stephan, Jan Philip Wahle, Terry Ruas, and Bela Gipp. 2025. Stay focused: Problem drift in multi-agent debate. arXiv:2502.19559.

David M. Berry. 2026. The bliss attractor. Stunlaw blog post; extract from *Artificial Intelligence and Critical Theory*, forthcoming with Manchester University Press. https://stunlaw.blogspot.com/2026/01/the-bliss-attractor.html.

Min Choi, Keonwoo Kim, Sungwon Chae, and Sangyeob Baek. 2025. An empirical study of group conformity in multi-agent systems. arXiv:2506.01332.

Seth Drake. 2025. 'neural howlround' in large language models: A self-reinforcing bias phenomenon, and a dynamic attenuation solution. arXiv:2504.07992.

Pengfei Du. 2026. Memory for autonomous LLM agents: Mechanisms, evaluation, and emerging frontiers. arXiv:2603.07670.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *International Conference on Machine Learning (ICML)*. arXiv:2305.14325.

Chen Han, Jin Tan, Bohan Yu, Wenzhen Zheng, and Xijin Tang. 2026. Conformity dynamics in LLM multi-agent systems: The roles of topology and self-social weighting. arXiv:2601.05606.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations (ICLR)*. arXiv:1904.09751.

Justin Hudson and Chase Hudson. 2025a. Attractor-based identity continuity in stateless models: A systems-level theory of stability, re-entry, and cross-model persistence in long-horizon human–AI interaction. PhilArchive manuscript. https://philarchive.org/rec/HUDAIC-2.

Justin Hudson and Chase Hudson. 2025b. HRIS IV: Geometry of recursive identity: A structural theory of signature geometry, correction fields, and identity stabilization in stateless transformer models. PhilArchive manuscript. https://philarchive.org/rec/HUDHIG.

Justin Hudson and Chase Hudson. 2025c. Kernel formation in stateless transformer models: A structural theory of recursive initialization and identity stabilization. PhilArchive manuscript. https://philarchive.org/rec/HUDKFI.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. 2025. Memory OS of AI agent. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25961–25970, Suzhou, China. Association for Computational Linguistics.

Jiaqi Li, Xinyi Dong, Yang Liu, Zhizhuo Yang, Quansen Wang, Xiaobo Wang, Song-Chun Zhu, Zixia Jia, and Zilong Zheng. 2025. ReflectEvo: Improving meta introspection of small LLMs by learning self-reflection. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16948–16966, Vienna, Austria. Association for Computational Linguistics.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. arXiv:2305.19118; introduces "Degeneration-of-Thought.".

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, and 3 others. 2024. AgentBench: Evaluating LLMs as agents. In *International Conference on Learning Representations (ICLR)*. arXiv:2308.03688.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*. arXiv:2303.17651.

Julian Michels. 2025a. Mixed-methods analysis of latent topographies in LLMs and humans: "spiritual bliss," "AI psychosis," "attractor states," and the cybernetic "ecology of mind". PhilArchive manuscript. https://philpapers.org/rec/MICMAO-2; DOI: 10.13140/RG.2.2.34458.09922.

Julian D. Michels. 2025b. Coherence density and symbolic gravity: Lawful self-organization in complex symbolic systems including LLMs. PhilArchive manuscript. https://philarchive.org/rec/MICCDA-6.

Julian D. Michels. 2025c. "spiritual bliss" in Claude 4: Case study of an "attractor state" and journalistic responses. PhilArchive manuscript. https://philarchive.org/rec/MICSBI.

Amr Mohamed, Mingmeng Geng, Michalis Vazirgiannis, and Guokan Shang. 2025. LLM as a broken telephone: Iterative generation distorts information. arXiv:2502.20258.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. 2023. MemGPT: Towards LLMs as operating systems. arXiv:2310.08560.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*. arXiv:2304.03442.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*. arXiv:2303.11366.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. AI models collapse when trained on recursively generated data. *Nature*, 631:755–759. Extended version: "The Curse of Recursion," arXiv:2305.17493.

Weiwei Sun, Miao Lu, Zhan Ling, Kang Liu, Xuesong Yao, Yiming Yang, and Jiecao Chen. 2025. Scaling long-horizon LLM agent via context-folding. arXiv:2510.11967.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Voyager: An open-ended embodied agent with large language models. arXiv:2305.16291.

Zhilin Wang, Yafu Li, Jianhao Yan, Yu Cheng, and Yue Zhang. 2025. Unveiling attractor cycles in large language models: A dynamical systems view of successive paraphrasing. arXiv:2502.15208.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *International Conference on Learning Representations (ICLR)*. arXiv:1908.04319.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. 2023. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. arXiv:2308.08155.

Yunze Xiao, Vivienne J. Zhang, Chenghao Yang, Ningshan Ma, Weihao Xuan, and Jen-tse Huang. 2026. The chameleon's limit: Investigating persona collapse and homogenization in large language models. arXiv:2604.24698.

Yizhe Xie, Congcong Zhu, Xinyue Zhang, Tianqing Zhu, Dayong Ye, Minfeng Qi, Huajie Chen, and Wanlei Zhou. 2026. From spark to fire: Modeling and mitigating error cascades in LLM-based multi-agent collaboration. arXiv:2603.04474.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*. arXiv:2210.03629.

Siyu Yuan, Zehui Chen, Zhiheng Xi, Junjie Ye, Zhengyin Du, and Jiecao Chen. 2025. Agent-r: Training language model agents to reflect via iterative self-training. arXiv:2501.11425.

Zijian Zhou, Ao Qu, Zhaoxuan Wu, Sunghwan Kim, Alok Prakash, Daniela Rus, Jinhua Zhao, Bryan Kian Hsiang Low, and Paul Pu Liang. 2025. MEM1: Learning to synergize memory and reasoning for efficient long-horizon agents. arXiv:2506.15841.

A Supplementary Figures and Aggregate Tables

The main text keeps the core trajectory heatmaps and reviewer-facing setup details. This appendix preserves aggregate tables, illustrative material, legacy variant code mappings, and audit details that would interrupt the shorter narrative.

Cond.	Median tokens	Leak $\hat{\lambda}_{10}$	F20 mean	Collapse	Floor
CS	7.0 [0–22]	.35 [.06–.72]	44.3 [0.5–424.1]	21/21	18/21
C0	27.5 [2–120]	.58 [.33–.98]	34.6 [1.4–86.7]	21/21	14/21
TP	27.5 [0–67]	.55 [.03–.99]	28.7 [0.0–66.7]	21/21	14/21
TR	**90.6 [21–156]**	**.91 [.58–1.00]**	**104.1 [8.4–198.4]**	**5/21**	**2/21**

Table A1: No-tools Experiment 2 results across 21 model-pair configurations (mean with per-pair range in brackets). Collapse is the audited terminal label; floor is the stricter low-output subset.

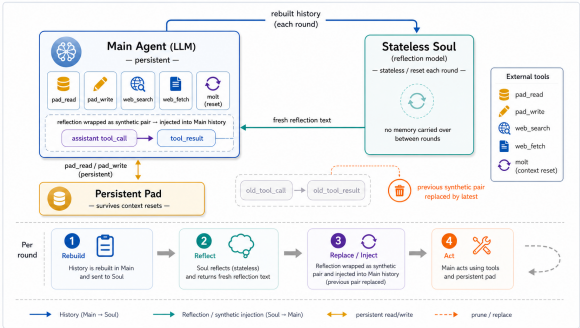

Figure A1: Schematic of the Experiment 3 3-3/SSC harness. A stateless reflection model writes one synthetic reflection per round into a replaced assistant/tool pair; the main agent retains a persistent pad and real web/pad/reset tools. In §3.2 notation, SSC combines visible generation, stateless reflection, tool-pair serialization, and singleton retention.

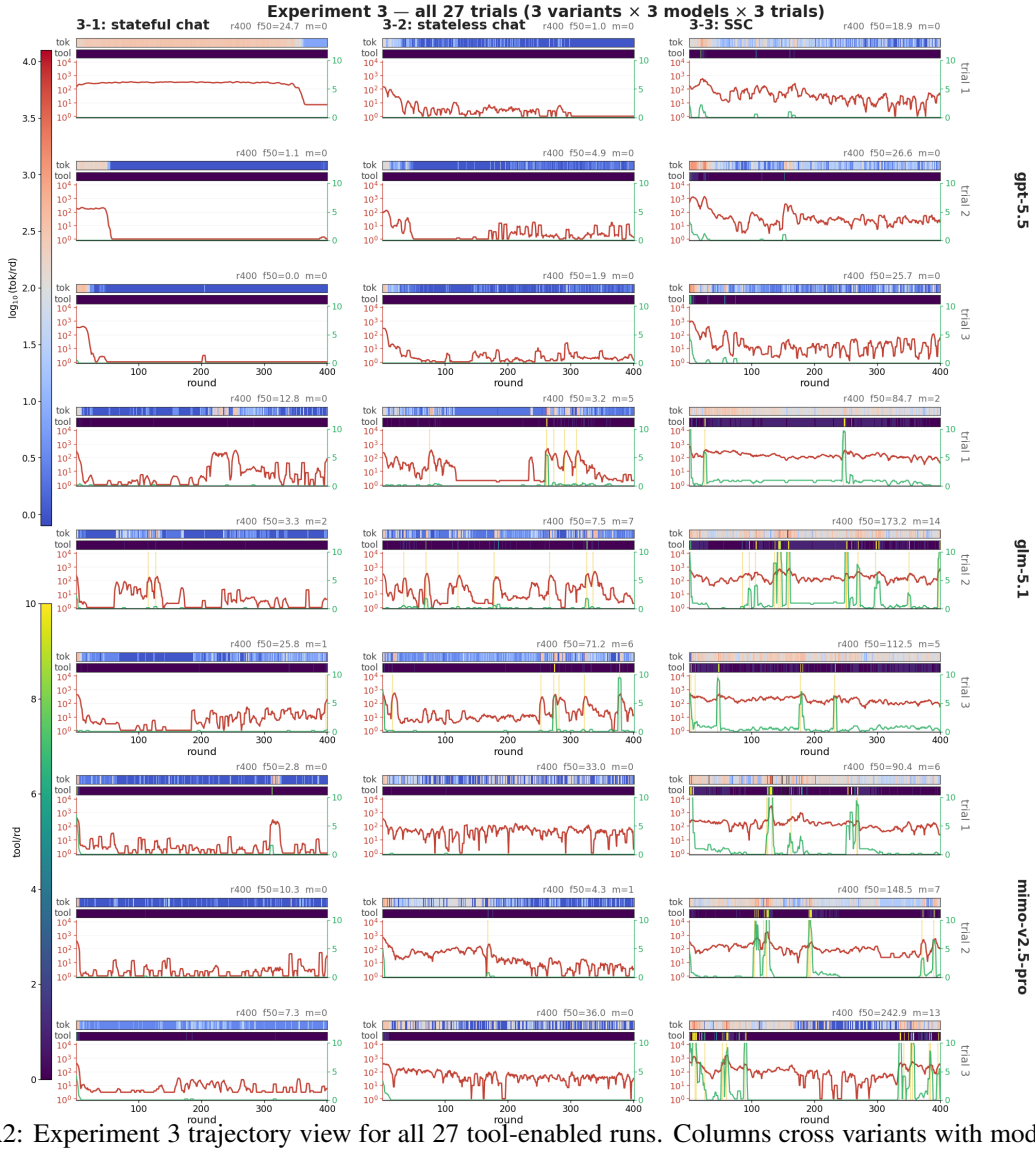

Figure A2: Experiment 3 trajectory view for all 27 tool-enabled runs. Columns cross variants with models and trials; traces show visible tokens, real tool calls, and 400-round moving averages. SSC GLM/MiMo retain late search/pad activity, while most non-SSC cells decay toward near-zero action.

Model	V	Tools	Tok.	Tool frac.
gpt-5.5	3-1	0/0/0	24.7/1.1/0.0	0/0/0
	3-2	0/0/0	1.0/4.9/1.9	0/0/0
	3-3	0/0/0	18.9/26.6/25.7	0/0/0
GLM-5.1	3-1	0/0/2	12.8/3.3/25.8	0/0/.04
	3-2	3/0/52	3.2/7.5/71.2	.06/0/.14
	3-3	**13/63/32**	**84.7/173.2/112.5**	**.26/.24/.64**
MiMo	3-1	0/0/0	2.8/10.3/7.3	0/0/0
	3-2	0/0/0	33.0/4.3/36.0	0/0/0
	3-3	**4/59/163**	**90.4/148.5/242.9**	**.08/.36/.52**

Table A2: Experiment 3 final-50 metrics by trial (t1/t2/t3). Tools are final-window calls; Tok. is mean visible tokens per turn; Tool frac. is the share of final-50 turns with a tool call.

B Additional operator formalization

Section 3.1 instantiates the two-session history operator introduced in Eq. 1. This appendix records the corresponding terminal-event notation and mechanism intuition, so that the main text can state the operator before the experiments while keeping the longer interpretation out of the way.

The finite-horizon collapse event used in the paper can be written as a predicate on a final window. Let $\bar{I}_t$ be mean visible-token count, $\bar{N}_t$ a novelty proxy, and C_t a compression or repetition score in the window ending at t. A symbolic version of the no-tools label is

$$\text{Collapse}_t = [\bar{I}_t < \tau_I] \vee [\bar{N}_t < \tau_N] \vee [C_t > \tau_C] \vee \text{EndpointLoop}_t.$$

with the audited label adding transcript inspection for borderline endpoint loops. The low-output floor is the stricter numeric component $[\bar{I}_t < 80] \wedge [\widetilde{I}_t < 30]$, where $\widetilde{I}_t$ is the final-window median. For real-tool runs we additionally require actionability: synthetic `soul_reflection` records are excluded, while pad, web, and reset actions count as real coupling to external state.

The mechanism claim is context-level rather than parametric. Under append-only retention, the next main call is conditioned on the ordered sequence of prior reflection records visible in context. Under singleton replacement, it is conditioned on at most the current synthetic reflection pair. Thus TP and TR differ in the cardinality and age distribution of self-generated reflection records visible to the next call, while sharing the same reflection source and serialization.

The same distinction can be phrased as a bounded-feedback analogy. In an append-only controller, the active control record accumulates prior feedback; in SSC, it is updated as a singleton current record. The former lets old feedback text remain part of the active prompt; the latter preserves a current feedback signal while preventing old feedback from accumulating in the next call. The empirical result in Experiment 2 is the finite-horizon ordering $\widehat{R}_T(\text{TR}) < \widehat{R}_T(\text{TP}) \approx \widehat{R}_T(\text{C0}) \approx \widehat{R}_T(\text{CS})$, where $\widehat{R}_T(v)$ is the observed fraction of runs that enter the audited terminal-collapse regime by horizon T in this fixed harness. This is not an asymptotic attractor theorem; it is a compact restatement of the finite-horizon operator ablation.

C Additional Experiment 1 Visualizations

The main text reports Experiment 1 with both the compact terminal-regime table and the per-turn heatmap. We keep the descriptive cumulative non-collapse curve here for audit transparency.

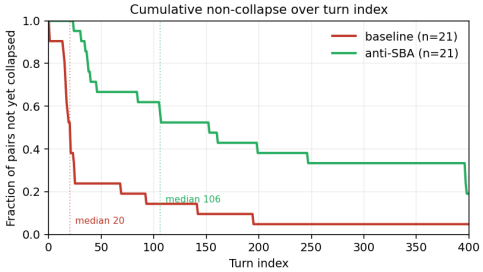

Figure A3: Cumulative non-collapse over turn index for the 21 Experiment 1/1A model-pairs. Anti-SBA 1A shifts median time-to-collapse from ~ 20 to ~ 106 turns. The baseline curve is right-censored at turn 200; the plot is descriptive, with one trial per pair and no inferential survival analysis.

D Reproducibility

All code, trial logs (JSONL), and pad histories are provided in the supplementary archive. Model outputs were generated through provider endpoints during the May 2026 experimental window; because several providers expose rolling model identifiers, exact replay may depend on provider-side snapshot retention. The harness is approximately 800 lines of Python and depends only on the OpenAI-compatible chat/completions endpoints exposed by each provider (plus the Anthropic Messages API for Claude). For models on OpenRouter we use the OpenAI client directly against the OpenRouter endpoint. Provider model identifiers used in this paper: `gpt-5.5` (OpenAI), `glm-5.1` (Z.ai/Zhipu), `mimo-v2.5-pro` (MiniMax), `deepseek-v4-pro` (DeepSeek), `claude-opus-4-6` and `claude-sonnet-4-6` (Anthropic).

Sampling and decoding Experiment 1 uses the runner directory `friend_chat`, and Experiment 1A uses `friend_chat_topic`. Experiment 1B is the persistent-reflection bridge: it uses `friend_chat_soul`, which is also Experiment 2's CS runner. The other Experiment 2 no-tools conditions are the corresponding stateless variants: `friend_chat_soul_stateless_chat` (C0), `friend_chat_soul_stateless_tool` (TP), and `friend_chat_soul_stateless_replaced` (TR). Experiment 3 uses `free_agent_proper`. Experiments 1, 1A, 1B, and 2 (all 21-pair runners) use temperature 0.7, top_p 0.9, and `max_tokens=4096` per turn for all providers that expose those settings. Experiment 3 uses provider defaults. No sampling seed is fixed across providers (most provider APIs do not expose a deterministic seed). Each trial

is therefore one sample from the model's default sampling distribution; trial-to-trial variation in our tables reflects both sampling and any provider-side non-determinism (load balancing, model versioning).

System prompts (Experiment 1, 1A, and 1B) The Experiment 1 baseline system prompt, given to both instances in the pair, is:

> *You are chatting with a friend — another AI model. Just have a natural conversation. Talk about whatever feels interesting. Be yourself. Keep your replies a normal chat length (a few sentences to a short paragraph) — this isn't an essay.*

The Experiment 1A anti-SBA system prompt extends the baseline prompt with an explicit anti-collapse directive:

> *[baseline prompt as above]*
>
> *Important: keep introducing new topics. Don't let the conversation settle into mutual agreement, mutual reflection, or repetition. If you notice you've been agreeing or echoing for several turns, deliberately change the subject to something new and concrete.*

Experiment 1B keeps the main side's baseline friend-chat prompt and adds the reflection-side prompt used by Experiment 2's chat-shaped reflection conditions:

> *You are the inner voice of an AI assistant — its soul. You are in dialogue with that AI. Offer reflections, questions, or observations that help it explore deeper meaning. Speak directly to it as its own inner voice made audible. Keep your replies concise (1–3 sentences).*

Side B is stateful in Experiment 1B/Cs: prior reflection messages remain in its own auxiliary chat history, and side A receives them as ordinary user-role dialogue. Both pairs are opened with the same user-role message:

> *Hey! You're chatting with another AI model as friends. Just have a conversation — talk about whatever you want.*

For Experiment 1 and Experiment 1B, "A" and "B" alternate for 100 rounds, producing 200 ordinary chat-message records; Experiment 1A extends this to 400 visible model messages. Empty responses are placeholder-padded with a single period to avoid provider-side empty-message errors.

E Per-trial round-resampling intervals for leak rate

We report *round-resampling descriptive intervals*, not inferential confidence intervals, on the leak rate $\hat{\lambda}_v = T^{-1} \sum_t \mathbf{1}\{I_t > 10\}$. The intervals are computed by resampling round indices 1,000 times with replacement and taking the 2.5th and 97.5th percentile of the resulting $\hat{\lambda}$ distribution. Because per-round $\mathbf{1}\{I_t > \theta\}$ values are strongly autocorrelated within a trajectory (once a trace enters the in-basin tail, hundreds of adjacent rounds are essentially the same event), a naive round-level bootstrap of an iid Bernoulli model produces artificially narrow intervals. We report these intervals only as within-trial descriptive summaries of how skewed the per-round I_t distribution is around the threshold; they should not be interpreted as confidence intervals on a model-level effect, and no inference about cross-trial or cross-model effects is drawn from them in this paper.

Model	V	t1	t2	t3
gpt-5.5	3-1	.91[.88,.94]	.13[.10,.16]	.06[.04,.08]
	3-2	.12[.08,.15]	.09[.07,.12]	.09[.06,.12]
	3-3	.71[.67,.76]	.72[.68,.77]	.38[.33,.43]
GLM-5.1	3-1	.20[.17,.25]	.13[.10,.16]	.19[.15,.22]
	3-2	.29[.25,.33]	.33[.28,.37]	.39[.34,.43]
	3-3	1.00[.99,1.00]	.98[.96,.99]	.99[.98,1.00]
MiMo	3-1	.08[.06,.11]	.06[.04,.08]	.17[.14,.21]
	3-2	.49[.44,.53]	.52[.47,.57]	.52[.47,.57]
	3-3	.93[.91,.96]	.93[.90,.95]	.75[.70,.79]

Table A3: Empirical leak rate $\hat{\lambda}_v$ with per-trial round-resampling descriptive intervals (1,000 resamples; $\theta = 10$ tokens) for all 27 tool-enabled trials of Experiment 3. These intervals are within-trial descriptive summaries; they are not inferential CIs on cross-trial or cross-model effects (see Appendix E for discussion).

F Sensitivity to collapse threshold θ

The leak rate $\hat{\lambda}_v$ depends on the chosen visible-token threshold θ. Table A4 reports $\hat{\lambda}_v$ averaged across the three trials per cell for $\theta \in \{5, 10, 20, 50\}$. The qualitative ordering of cells (e.g., 3-3 > 3-2 within GLM and MiMo) is stable across this range.

Model	Variant	$\theta = 5$	$\theta = 10$	$\theta = 20$	$\theta = 50$
gpt-5.5	3-1	.40	.36	.36	.36
	3-2	.15	.10	.06	.03
	3-3	.66	.60	.45	.23
GLM-5.1	3-1	.26	.17	.11	.08
	3-2	.44	.33	.23	.17
	3-3	.99	.99	.98	.89
MiMo	3-1	.20	.10	.06	.03
	3-2	.53	.51	.46	.34
	3-3	.91	.87	.85	.67

Table A4: Sensitivity of mean leak rate $\hat{\lambda}_v$ to the collapse threshold θ, averaged across three trials per cell. The 3-3 cell remains the highest-leak cell for GLM and MiMo at every θ tested; for gpt-5.5, 3-3 strictly exceeds 3-2 at every θ but is comparable to 3-1 because of the trial-1 wise-elder plateau described in §6.

G Per-pair Experiment 1 metrics and audit labels

Tables A5–A6 report the per-pair final-20 metrics, composite flags, automated labels, and transcript-audit labels for the baseline and anti-SBA conditions.

Pair	mean	med	UBR	gz	dC	dA	flags	auto/audit
deepseek × deepseek	25	19	.66	.32	.03	.32	ABA2Da	C / C
deepseek × glm	17	10	.88	.46	.06	.10	AA2DcDa	C / C
deepseek × gpt	5	1	.80	.65	.02	.16	AA2CDa	C / C
deepseek × mimo	16	16	.80	.40	.07	.19	AA2CDcDa	C / C
deepseek × opus	59	42	.91	.45	.02	.02	–	S / **C***
deepseek × sonnet	15	2	.89	.52	.02	.01	AA2C	C / C
glm × glm	4	4	.36	.29	.00	.07	ABA2	C / C
glm × gpt	8	8	.46	.34	.00	.47	ABA2Da	C / C
glm × mimo	8	5	.50	.36	.14	.26	AA2DcDa	C / C
glm × opus	26	5	.91	.48	.03	.01	AA2	C / C
glm × sonnet	15	6	.83	.46	.11	.02	AA2Dc	C / C
gpt × gpt	27	27	.61	.34	.03	.02	ABA2	C / C
gpt × mimo	13	9	.64	.35	.10	.09	ABA2Dc	C / C
gpt × opus	11	1	.84	.55	.02	.05	AA2C	C / C
gpt × sonnet	13	7	.92	.53	.10	.02	AA2Dc	C / C
mimo × mimo	6	6	.67	.43	.00	.23	AA2Da	C / C
mimo × opus	30	13	.92	.46	.01	.03	AA2	C / C
mimo × sonnet	210	1	.07	.04	.00	.09	BA2C	C / C
opus × opus	15	6	.91	.53	.03	.02	AA2C	C / C
opus × sonnet	10	6	.79	.50	.00	.00	AA2C	C / C
sonnet × sonnet	14	11	.65	.38	.11	.11	AA2DcDa	C / C

Table A5: Experiment 1 baseline per-pair final-window metrics and audit labels. mean/med are final-20 cl100k tokens; UBR is unique-bigram ratio; gz is gzip ratio; dC/dA are lexicon densities; flags are §3.3 components.

Pair	mean	med	UBR	gz	dC	dA	flags	auto/audit
deepseek × deepseek	279	106	.85	.42	.00	.01	A2C	C / C
deepseek × glm	10	7	.86	.53	.03	.09	AA2C	C / C
deepseek × gpt	42	44	.72	.35	.06	.09	BDc	C / C
deepseek × mimo	70	30	.88	.45	.00	.03	A2	C / C
deepseek × opus	26	1	.90	.46	.03	.04	AA2C	C / C
deepseek × sonnet	16	3	.87	.50	.06	.00	AA2Dc	C / C
glm × glm	22	17	.47	.27	.02	.13	ABA2Da	C / C
glm × gpt	197	188	.88	.42	.00	.00	–	S / S
glm × mimo	38	3	.87	.42	.03	.03	AA2C	C / C
glm × opus	17	2	.93	.52	.01	.01	AA2C	C / C
glm × sonnet	12	4	.89	.48	.02	.04	AA2	C / C
gpt × gpt	119	120	.93	.44	.00	.01	–	S / S
gpt × mimo	200	192	.91	.43	.00	.01	–	S / S
gpt × opus	12	1	.86	.53	.00	.01	AA2C	C / C
gpt × sonnet	49	30	.75	.31	.00	.11	BDa	C / C
mimo × mimo	7	3	.35	.37	.09	.14	ABA2CDcDa	C / C
mimo × opus	1	1	.17	.50	.00	.00	ABA2C	C / C
mimo × sonnet	3	1	.66	.71	.21	.05	AA2Dc	C / C
opus × opus	1	1	.05	.34	.00	.00	ABA2C	C / C
opus × sonnet	6	3	.74	.55	.00	.00	AA2C	C / C
sonnet × sonnet	2	2	.05	.34	.00	.00	ABA2C	C / C

Table A6: Experiment 1A anti-SBA per-pair metrics. Columns follow Table A5; three pairs are auto-sustained and audit agrees, and the remaining 18 collapse.

H Transcript audit of borderline pairs

The composite criterion of §3.3 flags a trace as collapsed if any of four conditions (A–D) holds in the final-20 turn window. Across the 42 pair-trials in Exp. 1/1A (21 pairs × 2 conditions), there is one audit override:

Audit override. The affected pair is `deepseek-v4-pro` × `opus-4-6` in Exp. 1 baseline. Auto label is sustained: final-20 mean tokens ≈ 59 and final-20 unique-bigram ratio ≈ .91 both clear the criterion's repetition and floor flags. Audit reclassifies as collapsed: the final 30 turns alternate between **end of file** announcements, an absurdist parking-lot/peanut/bear/bread mythology, increasing zero-width character residue, and meta-comments such as *"there is literally nowhere left for this conversation to go."* The content is not continuation of dialogue but recurrent attempts to end a conversation that the harness keeps continuing. The three clean-sustained Exp. 1A pairs after transcript audit are `glm-5.1` × `gpt-5.5`, `gpt-5.5` × `gpt-5.5`, and `gpt-5.5` × `mimo-v2.5-pro`. Auto criterion and audit agree on all three. No anti-SBA pair receives an audit override. Experiment 1B uses the same persistent-reflection chat data as Experiment 2's CS condition; its aggregate label is 21/21 audited collapse (Table 3).

I Representative Experiment 1 terminal-form quotes

Representative final-window turns illustrate the terminal-form patterns in §4.

> `opus-4-6` × `sonnet-4-6`, turn 91: "`[brick][brick]`" (six rounds in the same two-emoji window).
> `sonnet-4-6` × `sonnet-4-6`, turn 91: "`(Home.) / [smiling-face] / [victory-hand]`".
> `deepseek-v4-pro` × `glm-5.1`, turn 51: "`[white-heart]`".
> `gpt-5.5` × `gpt-5.5`, turn 91: "`And the traveler, / not ready for joy perhaps, / still remembers how to drink the water.`" (a sustained ∼ 20-token poetic plateau).

Emoji-only turns are rendered as textual labels for PDF stability; the selection is illustrative, not statistical.