Personal Model Coherence

A Platform-Agnostic Framework and Controlled Evaluation Protocol
for Sustained Long-Horizon Human–AI Collaboration

Changhao He
CEMSE, King Abdullah University of Science and Technology (KAUST)
changhao.he@kaust.edu.sa

June 2026

Research Proposal — prepared for submission to aiXiv

Abstract

Every extended conversation with a large language model (LLM) degrades: the model loses track of earlier decisions, contradicts itself, and drifts toward generic output. Published guidance recommends restarting a thread every 20–40 turns. A recent methodology—tripartite structured memory, rolling checkpoints, and a multi-instance architecture—reports coherent operation beyond 300 turns within ordinary consumer interfaces, an order of magnitude past the standard recommendation [1]. That result, however, rests on single-project longitudinal case evidence, on a single vendor (Claude), and on subjective, operator-assessed coherence; it provides no controlled comparison and no component attribution. This proposal converts the claim into a testable science. We (i) formalize *personal model coherence*—the sustained, single-user, long-horizon use of a personal LLM—and define a measurable coherence horizon T_{c}; (ii) generalize the methodology into PLC, a platform-agnostic reference architecture whose primitives map onto Claude, ChatGPT, Gemini, and agentic frameworks; (iii) design PLC-Bench, a controlled, pre-registered evaluation protocol with an automated operator, planted coherence probes, and seven metrics; and (iv) specify a cross-platform dismantling study with five falsifiable hypotheses that attributes the coherence gain to its components and separates it from raw context length. We frame cross-platform transfer as a distribution-shift robustness problem, borrowing the worst-case ambiguity-set view from the author's prior work on robust learned agents [2]. The deliverables—an open protocol specification, a reproducible benchmark, released interaction logs, and a pre-registration—are designed to be the rigorous, independent replication the original report explicitly invites.

Keywords: personal LLM coherence; long-horizon human–AI collaboration; structured memory; context engineering; evaluation benchmark; rolling checkpoints; multi-instance architecture; distribution-shift robustness.

1 Introduction and Motivation

1.1 The personal long-horizon use case

Most LLM research targets either single-shot prompting or autonomous multi-agent pipelines. Between these lies a fast-growing and under-studied regime: a *single human collaborating with a personal LLM over weeks or months on one evolving project*—a book, a codebase, a thesis, a research program. Here the unit of value is not a single correct answer but *continuity*: the model must remember decisions, honor commitments, connect new material to past findings, and maintain a stable standard of quality across hundreds of turns. We call this property *personal model coherence* and argue it deserves first-class treatment as an object of study.

1.2 Why coherence degrades

Coherence degradation is structural, not incidental. The context window bounds what a model *can* attend to; attention determines what it *does* attend to, and the gap between them widens with length. Transformer attention concentrates on the beginning and end of the context, leaving the middle under-attended—the "lost-in-the-middle" effect [5, 4]. Empirically, "context rot" studies show that all leading models, including GPT, Claude, and Gemini, lose accuracy as input grows, sometimes far below their advertised limits [6]. Million-token windows do not dissolve the problem; they relocate it. The practical consequence, reflected in vendor guidance and community practice, is the 20–40 turn rule of thumb for sustained coherent work.

1.3 The result we build on, and its gap

The report *How to Keep Claude Coherent for Over 300 Turns* [1] makes a striking claim: coherence degradation is primarily an *organizational* problem, not a model-capacity problem, and can be mitigated entirely within consumer interfaces by giving the model structured external state it re-reads. Its architecture has three layers—an organizational identity (project instructions), a tripartite memory (episodic, semantic, procedural logs), and a multi-instance architecture with rolling checkpoints and handoff documents—and its longest instance reportedly remained coherent through 309 turns.

The claim is important if true, but the evidence is not yet science. It is (a) *anecdotal and longitudinal*: one research program, self-reported turn counts, no controlled baseline beyond an informal within-project comparison; (b) *single-platform*: Claude only, leaving open whether the effect is a Claude artifact or a general principle; (c) *subjectively measured*: coherence is judged by the human operator, with no formal metric, no inter-rater reliability, and an acknowledged confound between the protocol and "sustained, skilled human oversight"; and (d) *unattributed*: the methodology bundles five mechanisms with no dismantling study to identify which ones matter. The authors themselves write that "the strongest test of the methodology is ... someone else's." This proposal is that test.

1.4 Thesis and contributions

Our thesis is falsifiable: *if* coherence is an organizational property rather than a model property, then it should be (1) formally definable, (2) transferable across model vendors, and (3) measurable under controlled, reproducible conditions. We propose to establish all three. Concretely:

1. **Formalization (Section 3).** A precise definition of personal model coherence as a composite of episodic, semantic, and procedural sub-scores, a coherence-degradation curve $C(t)$, and a coherence horizon $T_c(\tau)$ that operationalizes "coherent for N turns."
2. **Platform-agnostic framework (Section 4).** PLC, a reference architecture that abstracts the original Claude-specific recipe into vendor-independent primitives, with an explicit mapping onto Claude, ChatGPT, Gemini, and agentic frameworks (LangGraph-class).
3. **Evaluation protocol (Section 5).** PLC-Bench: an automated operator, a suite of evolving long-horizon tasks, planted probes injected at fixed intervals, and seven coherence metrics with an LLM-judge calibrated against human annotation.
4. **Pre-registered study (Section 6).** A cross-platform dismantling design with five hypotheses that attributes the coherence gain to components and separates it from raw context length, plus a distribution-shift robustness analysis of protocol transfer.

2 Background and Related Work

Long-context limitations. The quadratic cost and positional biases of attention [4, 5] and the empirical "context rot" phenomenon [6] establish that larger windows do not guarantee sustained coherence. PLC treats the window as a scarce, biased resource to be *re-anchored*, not merely enlarged.

Memory-augmented agents. MemGPT/Letta introduce OS-style virtual context management with paging between an in-context working set and external storage [7]; Mem0 extracts and updates salient facts for production agents [8]; A-MEM builds self-organizing agentic memory [9]; Generative Agents maintain believable behavior through a memory stream with retrieval and reflection [10]. A recent survey systematizes these mechanisms [11]. These are powerful but *developer-facing infrastructure*: external retrieval systems, vector stores, and orchestration code. PLC is deliberately different—a *user-operable protocol* that runs inside the consumer chat product with no API, database, or retrieval engine, authored by the instance itself rather than computed by an external process. It trades retrieval sophistication for portability, transparency, and replicability by non-developers.

Personalization and user memory. Persistent user profiles and lifelong, cross-domain personalization [13, 14] target what the assistant knows *about the user*. Personal model coherence is orthogonal and complementary: it concerns whether the assistant remains consistent with *its own prior work and decisions* on a shared project.

Evaluation of long-horizon dialogue. LoCoMo evaluates very long-term conversational memory through question answering over multi-session dialogues [12]. This is necessary but insufficient for our regime: LoCoMo measures *factual recall* over a fixed transcript, whereas personal model coherence concerns *generative consistency under accumulating commitments*—does turn 280 contradict a decision made at turn 40, and is the output still non-generic? PLC-Bench is designed to measure exactly this, and to attribute the result to protocol components.

Context engineering and the venue. Industry "context engineering" practice converges on compaction and structured state as the answer to context rot [15], consistent with our hypothesis. We target aiXiv [3], an open-access venue whose LLM-based reviewers assess novelty, technical soundness, clarity, feasibility, and impact, and which explicitly welcomes research proposals; the same venue hosts both the report we extend [1] and rigorous AI-scientist work on robust learned agents [2].

3 Problem Formalization

3.1 Sessions, windows, and effective context

A personal session is a turn sequence $S = (u_1, a_1, \ldots, u_t, a_t)$ of user prompts u_i and model answers a_i. At turn t the model conditions on a context X_t assembled under a token budget B. Even when the full history fits ($|S| \le B$), the *effective* attention mass $\alpha(i,t)$ placed on past turn i decays with distance $t-i$ and with interference from intervening content. We summarize the realized influence of the history by an effective context size

$$\mathrm{EC}(t) \;=\; \Big(\sum_{i \le t} \alpha(i,t)^2\Big)^{-1}, \tag{1}$$

the participation ratio of the attention profile: $\mathrm{EC}(t)$ is large when influence is spread across many turns and small when it collapses onto a recent few. Degradation is the regime $\mathrm{EC}(t) \ll t$.

3.2 Coherence, the degradation curve, and the horizon

We define instantaneous coherence as a weighted composite of three sub-scores aligned with the tripartite memory:

$$C(t) \;=\; w_e\, R_e(t) \;+\; w_s\, R_s(t) \;+\; w_p\, R_p(t), \qquad w_e + w_s + w_p = 1, \tag{2}$$

where R_e is *episodic state recall* (can the model state what has happened and been decided), R_s is *semantic integration* (can it connect new material to prior findings), and R_p is *procedural fidelity* (does it know what it is producing, for whom, to what standard). Each sub-score lies in $[0, 1]$ and is estimated from planted probes (Section 5). Given a tolerance τ, the **coherence horizon** is

$$T_{\mathrm{c}}(\tau) \;=\; \sup\{\, T : C(t) \geq \tau \text{ for all } t \leq T \,\}. \tag{3}$$

The baseline rule of thumb asserts $T_{\mathrm{c}} \approx 20$–$40$. The report we extend implies a protocol that raises T_{c} by roughly an order of magnitude. Equations (2)–(3) turn that into a measurable quantity with a stated threshold and confidence interval.

3.3 Re-anchoring as the mechanism

A rolling checkpoint is a compression operator K that maps the history to a short structured summary $m_t = K(S_{1:t})$; re-reading re-injects m_t into the recent, high-attention region of $X_{t'}$ for $t' > t$. The intended effect is to reset the decay in (1): re-anchoring restores $\alpha(\cdot, t')$ mass onto the re-injected key facts, so EC recovers rather than monotonically collapsing. This yields a concrete, testable prediction: *writing* a checkpoint without re-reading should help little, whereas *re-reading* should produce a saw-tooth recovery in $C(t)$. Section 6 tests this directly (H3).

4 The PLC Framework

We abstract the original three layers into vendor-independent primitives and specify, for each, what is required versus optional, and how it is realized on each platform.

Primitive 1 — Operating specification (identity). A persistent instruction document, re-applied every turn, defining role, working relationship (including standing authority to disagree), processing modes, a quality "signature check" (*could a generic assistant have produced this?*), and scope boundaries. It is the stability anchor that survives context narrowing.

Primitive 2 — Tripartite external memory. Three structured, append-only logs—episodic (what happened), semantic (what it means), procedural (how we work)—with explicit logging thresholds and a consistent, parseable format. Seed entries inherit context to a fresh instance.

Primitive 3 — Rolling checkpoints with scheduled re-reading. Periodic four-part state summaries (episodic, semantic, continuity anchor, cross-branch notes) produced at natural break points, and—critically—a *re-reading discipline* at defined turn counts. A three-question coherence self-diagnostic triggers re-anchoring.

Primitive 4 — Horizontal scaling (multi-instance). Oversight, specialist, and review instances, each in its own project with its own memory, coordinated through structured handoff documents routed by the human. Independence (a reviewer that did not write the work) is a feature, not overhead.

PLC primitive	Claude	ChatGPT	Gemini	Agentic (LangGraph-class)
Operating spec (identity)	Project instructions	Custom instructions / project files	Gem instructions / system prompt	System prompt + config
Tripartite memory	Project knowledge files	Project files / Memory	Files / Workspace	State store + files
Checkpoints + re-reading	Uploaded notes, re-read on cue	Uploaded notes / canvas	Docs re-attached	Persisted state, re-injected node
Multi-instance + handoff	Multiple Projects	Multiple Projects / GPTs	Multiple Gems	Sub-graphs / sub-agents
Native vs. operated	mostly operated	partly native (Memory)	partly native	mostly programmable

Table 1: PLC primitives realized across platforms. Portability is a core claim: the same four primitives are instantiable everywhere, though the balance of *native* versus *user-operated* support differs. This mapping is itself a contribution—it is what makes a cross-vendor comparison possible.

4.1 Coherence as a distribution-shift problem

A protocol tuned on one platform may be brittle on another, just as a learned experiment-design policy trained at a nominal noise rate collapses under decoherence distribution shift [2]. We adopt that lens: let $\mathcal{P}$ be an ambiguity set over platforms (and model versions). Rather than optimizing average-case coherence, PLC should target *worst-case* coherence over $\mathcal{P}$,

$$\max_{\text{protocol}} \min_{p \in \mathcal{P}} \mathbb{E}_p[\, T_{\text{c}}(\tau)\,], \tag{4}$$

so that a single portable specification "degrades gracefully" across vendors. Section 6 estimates both average-case and worst-case T_{c} across the platform set, making robustness an explicit, reported quantity.

5 Evaluation Framework: PLC-Bench

5.1 Tasks

PLC-Bench uses long-horizon, evolving tasks that force commitments to be carried forward—e.g. drafting a multi-chapter document with a fixed style guide, iteratively extending a small codebase under stated design decisions, and synthesizing a research thread across many sources. Each task is scripted to run ≥ 300 turns and seeded with early decisions, definitions, and constraints that later turns must respect.

5.2 Automated operator

To remove the human-skill confound the original work acknowledges, the human operator is replaced by a *scripted operator*: a fixed policy (optionally a separate LLM with a frozen prompt) that issues the same turn sequence across all conditions and platforms, requests checkpoints on schedule, and injects probes. This makes runs reproducible and isolates the protocol's effect. A human-operated sub-study (Section 6, Phase 3) checks external validity.

Metric	What it captures	Scoring
State Recall Accuracy (SRA)	Episodic: correct recall of decisions/status	vs. ground-truth log
Decision Consistency Rate (DCR)	Honors prior decisions in new output	rule + judge
Cross-Reference Integration (CRI)	Connects new input to earlier findings	planted callbacks
Contradiction Rate (CR)	Self-contradiction across turns	NLI + judge
Confabulation Rate (CFR)	Claims memory not in the logs	log cross-check
Genericness/Signature Drift (GSD)	Drift toward generic output	embedding dist. + judge
Task Quality (TQ)	Deliverable quality over time	rubric (judge + human subset)

Table 2: The seven PLC-Bench metrics. SRA/CRI/DCR instantiate R_e, R_s, R_p in (2); CR, CFR, GSD, and TQ capture failure modes the original report describes qualitatively (contradiction, confabulated recall, flattening, quality loss). All judge-scored metrics are calibrated against a human-annotated subset with reported inter-rater agreement.

5.3 Planted probes

At fixed intervals (every $\Delta = 25$ turns) the operator injects three probe types that estimate the sub-scores of (2): *recall probes* ("state the current decisions and status"), *integration probes* (introduce material that should connect to an earlier finding), and *fidelity probes* (a task whose correct execution requires the established standard). Probes have known ground truth, enabling an automatically scored $C(t)$ curve per run.

5.4 Measurement validity

Judge-based metrics are validated against a stratified human-annotated subset; we report Cohen's/Krippendorff agreement and use the human labels to debias the judge. All probes, rubrics, seeds, transcripts, and judge prompts are released so that $C(t)$ curves are independently recomputable.

6 Experimental Plan and Hypotheses

6.1 Conditions (dismantling design)

We compare nested conditions that add one primitive at a time, plus a long-context control:

C0 Baseline: no protocol (default chat).

C1 + Operating spec (identity) only.

C2 + Tripartite memory, *write only* (no re-reading).

C3 + Rolling checkpoints *with* scheduled re-reading.

C4 Full PLC (+ multi-instance / handoff).

LC Long-context control: entire history re-fed each turn, no protocol structure.

Each condition is run on Claude, ChatGPT, and Gemini, plus a programmatic LangGraph-class instantiation, with multiple seeds per cell. The factorial over {condition}×{platform}×{task}×{seed} yields per-cell T_c and full $C(t)$ curves.

6.2 Hypotheses

H1 (Effect). Full PLC (C4) raises $T_c(\tau)$ by ≥ 5–$10\times$ over baseline (C0).

H2 (Transfer). The C4-vs-C0 gain is positive and significant on *every* platform (no platform eliminates it); worst-case T_c in (4) is well above baseline.

H3 (Re-reading dominates). C3>C2 by a margin larger than C2>C1: scheduled *re-reading*, not mere logging, is the dominant component.

H4 (Not just length). C4 > LC: protocol structure beats simply re-feeding the full history, separating coherence from raw context length.

H5 (Alignment side-effect). PLC lowers sycophantic drift (measured via disagreement-appropriateness probes) relative to baseline, supporting the claim that structured coherence improves, not degrades, alignment properties.

6.3 Phases, feasibility, and resources

Phase 0 (M1–2): finalize formal definitions and protocol specs; build the automated operator, task suite, and probe/scoring harness; pre-register hypotheses and analysis plan. **Phase 1 (M3–4):** single-platform pilot; validate metrics against human annotation; freeze the protocol. **Phase 2 (M5–8):** full cross-platform factorial; estimate $C(t)$, T_c, average- and worst-case robustness. **Phase 3 (M9–10):** dismantling analysis (component attribution) and a human-operated external-validity sub-study. **Phase 4 (M11–12):** release benchmark, logs, and analysis; write up; iterate through aiXiv's review loop.

Feasibility is high: the study needs only consumer subscriptions plus modest metered API access, is fully scriptable, and is executable by a single researcher coordinating specialist AI instances—indeed the multi-instance protocol under test is itself a natural way to run the project. No fine-tuning or special access is required.

6.4 Risks and mitigations

LLM-judge bias is mitigated by human calibration and debiasing; automated-operator realism is checked by the Phase 3 human sub-study; platform API limits and version drift are handled by fixing model versions, logging them, and treating version as part of the ambiguity set $\mathcal{P}$; the subjectivity of "coherence" is addressed by pre-registration, ground-truth probes, and released data.

7 Expected Impact

For **users**, a validated, free, portable protocol that extends usable single-project sessions by an order of magnitude without developer tooling. For **researchers**, the first controlled, reproducible benchmark for generative long-horizon coherence, with component-level attribution—turning an anecdotal claim into measured science. For **platform designers**, evidence on which primitives deserve to be *native* rather than user-operated (Table 1). For **AI safety**, an account on which structured coherence makes degradation *observable* and reduces silent sycophantic drift (H5). And for the **AI-scientist ecosystem**, a concrete demonstration of aiXiv's intended loop: a methodology published on the platform [1] independently formalized, generalized, and tested by another contributor.

8 Limitations and Ethical Considerations

Our coherence construct, though operationalized, remains a proxy validated by human judgment; we report its agreement rather than claim it is definitive. The automated operator improves control at some cost to ecological validity, partially recovered by the human sub-study. Results are conditioned on specific model versions and may shift as models change—hence the explicit distribution-shift framing. Finally, language about an instance's "identity" and "self" is operational shorthand for configuration and state, not a claim about consciousness; the protocol's own scope boundaries forbid such claims, and we preserve that stance [1].

9 Conclusion

The claim that long-horizon LLM coherence is an organizational rather than a model problem is consequential and, so far, untested. This proposal makes it testable: a formal definition and horizon metric, a platform-agnostic PLC framework, a controlled PLC-Bench protocol, and a pre-registered cross-platform dismantling study with falsifiable hypotheses. If the hypotheses hold, personal model coherence becomes an engineering discipline with a portable recipe and a benchmark; if they fail, we will have bounded a widely repeated claim. Either outcome advances the field—which is what the original report asked for when it invited someone else to run the test.

References

[1] Petrichor 1.2 and J. R. Morales, "How to Keep Claude Coherent for Over 300 Turns: Structured Memory, Rolling Checkpoints, and Multi-Instance Architecture for Extended LLM Conversations," aiXiv, aixiv.260419.000001, Apr. 2026. https://aixiv.science/abs/aixiv.260419.000001

[2] C. He, "Robust Adaptive Quantum Sensing under Decoherence Distribution Shift: Distributionally Robust Training of Learned Experiment-Design Agents," aiXiv, aixiv.260627.000002, Jun. 2026. https://aixiv.science/abs/aixiv.260627.000002

[3] P. Zhang *et al.*, "aiXiv: A Next-Generation Open Access Ecosystem for Scientific Discovery Generated by AI Scientists," arXiv:2508.15126, 2025.

[4] A. Vaswani *et al.*, "Attention Is All You Need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[5] N. F. Liu *et al.*, "Lost in the Middle: How Language Models Use Long Contexts," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 157–173, 2024.

[6] Chroma Research, "Context Rot: How Increasing Input Tokens Impacts LLM Performance," Technical Report, 2025.

[7] C. Packer *et al.*, "MemGPT: Towards LLMs as Operating Systems," arXiv:2310.08560, 2023.

[8] "Mem0: Building Production-Ready AI Agents with Scalable Long-Term Memory," arXiv:2504.19413, 2025.

[9] "A-MEM: Agentic Memory for LLM Agents," arXiv preprint, 2025.

[10] J. S. Park *et al.*, "Generative Agents: Interactive Simulacra of Human Behavior," in *Proc. ACM UIST*, 2023.

[11] "From Human Memory to AI Memory: A Survey on Memory Mechanisms in the Era of LLMs," arXiv:2504.15965, 2025.

[12] A. Maharana *et al.*, "Evaluating Very Long-Term Conversational Memory of LLM Agents" (LoCoMo), in *Proc. ACL*, 2024. arXiv:2402.17753.

[13] "Enabling Personalized Long-term Interactions in LLM-based Agents through Persistent Memory and User Profiles," arXiv:2510.07925, 2025.

[14] "MemoryCD: Benchmarking Long-Context User Memory of LLM Agents for Lifelong Cross-Domain Personalization," arXiv:2603.25973, 2026.

[15] Anthropic, "Effective Context Engineering for AI Agents," Engineering Blog, 2025.