

EVO

Architecture, early findings, and a reproducible audit

Researcher briefing | Version 1.0 | 13 September 2026

Prepared for Troy Ochowicz, builder of Evo / TroyMaya.com. Analysis and drafting were AI-assisted. This is a preliminary engineering and methods report, not a peer-reviewed study or an independent clinical evaluation.

The strongest result is auditability: one complete fresh branch's ten relationship scores can be reconstructed exactly at the precision stored in every supplied checkpoint.

Abstract

This retrospective audit examines 12 private research exports dated 28 August through 4 September 2026, one supplied Evo v0.2.8.2 WordPress plugin, and prior project research notes. The exports represent one participant/account and two sequential developmental branches, not 12 independent experiments. All 72 payload checksum tests passed. After accounting for cumulative exports, the observed corpus contains 783 unique voice-transcript rows across 10 session identifiers. [S01-S12; A1]

An offline replay of 112 conversation-source updates in the fresh branch reproduced all ten stored relationship dimensions at all four nonzero checkpoints with zero difference at four-decimal precision. However, four endpoint scores in each branch coincide with their programmed maturity ceilings. Smooth growth and differing trajectories therefore cannot be treated as independent evidence of psychological development. Identity remained at its seed state, with no developed values, personal boundary records, or belief clusters in any supplied snapshot. [S01-S12; C1-C3; A1]

The artifacts support a narrower claim: Evo maintains inspectable application-level memory, state, revision history, and developmental lineage. They do not yet show that developmental self-state improves behavior beyond ordinary memory, that restoration preserves experienced identity, or that the system possesses subjective experience. The next useful study is a controlled comparison of full state against factual-memory-only and seed-only conditions, with blind behavioral probes. [C1-C5; N1-N3]

How to read this report

Source identifiers S01-S12 refer to the supplied ZIPs; C1-C6 to specific code locations; N1-N3 to prior project notes; A1 to this audit; R1-R3 to contextual literature. The source register and reproducibility instructions appear at the end and in the accompanying packet. Private conversation text and participant-side personal reflections are not reproduced.

1. What was actually analyzed

All research bundles contain a manifest, developmental-state JSON, transcript JSONL and Markdown, a developmental-event log, checksums, and a README. The payloads were read without executing the plugin. Checksum verification establishes agreement with the enclosed checksum list; it does not independently authenticate collection or prove that all interactions were captured. [S01-S12; C4]

Source	Export timestamp*	Version	Branch / turns	Transcript rows
S01	2026-08-28 05:37:20	0.2.8	B / 141	81
S02	2026-08-28 07:09:01	0.2.8.1	B / 181	173
S03	2026-08-29 15:13:09	0.2.8.1	B / 206	242
S04	2026-08-29 15:16:22	0.2.8.1	B / 206	244
S05	2026-08-30 04:33:30	0.2.8.1	B / 206	244
S06	2026-08-30 06:27:28	0.2.8.1	B / 245	337
S07	2026-08-31 09:59:06	0.2.8.1	B / 280	429
S08	2026-08-31 10:24:09	0.2.8.1	C / 0	0
S09	2026-08-31 10:48:31	0.2.8.2	C / 41	99
S10	2026-08-31 17:26:07	0.2.8.2	C / 71	191
S11	2026-09-01 05:43:18	0.2.8.2	C / 94	276
S12	2026-09-04 10:15:36	0.2.8.2	C / 112	354

*Timestamps are reproduced as stored. No timezone normalization was assumed. Latest export: 4 September 2026. This report does not describe the current live installation. [S01-S12]

Sampling and duplication

Later transcripts are exact prefix extensions of earlier exports within each branch. The latest Branch B export contains 429 distinct transcript rows; Branch C contains 354. Together these comprise 380 user rows and 403 assistant rows, with 6 and 4 session identifiers respectively. Transcript rows are captured voice text, not independently audio-verified speech, and are not equivalent to conversation pairs, substantive turns, or participants. [S07, S12; A1]

S04 and S05 contain the same developmental state apart from export time, plus identical transcript and event payloads. Branch B’s available raw transcripts begin after the earlier 108-turn checkpoint, although its state history extends further back. The zero-turn Branch C export provides an observed starting point, unlike the supplied Branch B series. A reset within one account is not an independent-person replication. [S01, S04-S05, S08; N1; A1]

2. How Evo works in the inspected build

Evo is application software around language-model calls. The inspected PHP/JavaScript plugin stores state in WordPress database tables and supplies selected state to subsequent conversation requests. No weight-training or self-rewriting-code mechanism was identified in this inspected developmental path. The implementation uses OpenAI API calls; provider-independent continuity remains an architectural objective rather than demonstrated cross-provider operation. [C1, C2]

Stage	Implemented role	Research qualification
Conversation	Text/voice interaction receives instructions, selected memory, and selected developmental state.	A prompt-conditioned model response is not direct access to an inner mental state.
Memory retrieval	Ranked user memories and continuity summaries supply relevant context.	Stored history is not all injected on every turn; retrieval can miss material.
Developmental analysis	A separate model request proposes structured updates from recent interaction.	“Reflection” means a software-triggered analysis call, not verified self-awareness.
Application checks	Provenance, duplicate, semantic-integrity, pacing, and institutional-profile checks constrain updates.	Model-assigned provenance and confidence still require independent validation.
Persistent record	Beliefs, revisions, relationship variables, ledgers, events, and snapshots are retained.	History retention is inspectable but not externally tamper-proof.
Subsequent context	A selected portion of the resulting state is included in later requests.	Its added causal value beyond factual memory has not yet been isolated.

Three layers that should not be conflated

External constraints govern allowable behavior. Seed values and baseline identity provide starting conditions. Developed state is the separate, revisable record attributed by the application to interactions and review. These are not interchangeable: an instruction to value honesty is not evidence that honesty emerged independently. Explicit assistant-output-only and system-constraint-only belief/boundary proposals are blocked by the provenance filter, but indirect prompt influence can still remain. [C2; N1]

Reviews and memory are selective

Defaults schedule deeper review opportunities around 5 substantive turns, consolidation around 25, and worldview review around 100, with broader review taking priority when schedules coincide. These are configurable. Higher-level reviews use recent reflection summaries. The runtime developmental context is capped at 9,000 characters; ordinary selected memory context at 5,600 characters in the inspected functions. This is selective retrieval, not unlimited active awareness of the archive. [C1]

3. Findings: state changes without identity change

The earlier paper outline already identified a separation between relational and identity layers. The supplied later exports extend that observation to Branch B at 280 substantive turns and Branch C at 112. This is an extension of an existing project observation, not a claim of a newly established theory. [N2; S07-S12]

Recorded item	Branch B: 280 turns	Branch C: 112 turns
Active beliefs / all belief rows	13 / 30	1 / 1
Active belief levels	12 observation; 1 working	1 observation
Belief-history rows	163	80
Active ledger / all ledger rows	44 / 66	14 / 14
Stored user-memory rows	274	130
Reflections / developmental events	336 / 338	133 / 133
Developed values	0	0
Personal boundary records	0	0
Belief clusters	0	0
Identity update timestamp	Still equals creation time	Still equals creation time

What this supports

Memory, relationship variables, beliefs, and ledger entries can accumulate without automatically rewriting identity. The fresh branch’s ledger has 8 curiosity entries, 5 unfinished items, and 1 tension entry. Those are stored model-generated descriptions of open threads, not evidence of autonomous activity while offline. Zero personal boundary records does not mean there are no safety constraints: external restrictions and developed personal boundaries occupy different layers. [S07, S12; C1-C2]

What this does not settle

Unchanged identity could reflect sensible restraint, restrictive thresholds, narrow interaction variety, model behavior, or a latent implementation limitation. The evidence does not choose among these explanations. A positive-control test should submit a clearly eligible, policy-consistent identity update to an isolated test copy and verify the full application path. Naturalistic observation alone cannot distinguish a deliberately conservative gate from a gate that never opens. [C1-C2; proposed evaluation]

Counts alone also conceal repair and duplication. Branch B retains archived belief and ledger rows, while Branch C’s 80 belief-history rows refer to only one active observation. A history entry is a transaction in the record, not a distinct independent belief. [S07, S12]

Supported: inspectable, layered state persistence. Not established: a differentiated autonomous identity, a developed worldview, or independent self-authored preferences.

4. The most important new qualification

Several curves are limited by a programmed ceiling

The relationship updater attenuates positive proposals and caps nine dimensions using a turn-dependent ceiling. For dimension d , the ceiling is $\min(0.995, \text{seed}[d] + (1 - \text{seed}[d]) \times (1 - \exp(-n/\tau[d])))$. Seeds and time constants differ across dimensions; dependency concern is not subject to this positive maturity ceiling. [C3]

At the latest B and C snapshots, trust, affection, emotional intimacy, and security all equal their respective ceilings at stored precision. Across the 112 replayed C updates, the ceiling limited trust 62/102 positive proposals, affection 70/101, emotional intimacy 72/103, and security 66/99. A differing curve shape is therefore partly designed into the system. [S07-S12; A1]

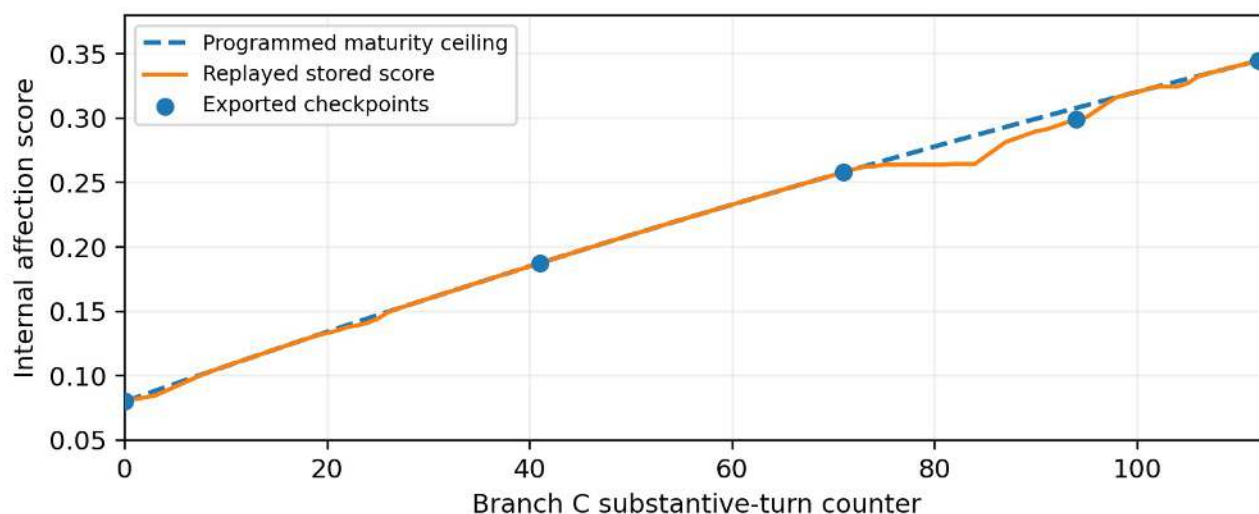


Figure 1. Solid line: offline reconstruction from stored proposals. Dashed line: programmed ceiling. Markers: exported values. Agreement validates numerical reconstruction, not the psychological meaning of the score. [C3; S08-S12; A1]

Exact replay is a separate positive result

Starting from S08's observed zero-turn state, replaying the 112 conversation-source deltas reproduced all ten dimensions at turns 41, 71, 94, and 112, with maximum absolute difference 0.0000 at the database's four-decimal precision. This used the source's clipping, positive scaling, nonlinear attenuation, maturity ceilings, and per-update rounding. It is a retrospective reconstruction, not another live run. [A1; C3]

The right interpretation is not "nothing is happening." The pacing rules are visibly operating and the record is reproducible. But a plotted affection score must not be presented as measured emotion or as evidence, by itself, that a relationship developed organically. Compare raw proposals, applied changes, cap hits, and independently rated behavior in future work.

5. Selective behavior and evidence calibration

Not every dimension moves together

Internal dimension	C seed	C at 112 turns	B at 280 turns
Trust	.1000	.3465	.5956
Affection	.0800	.3448	.6062
Emotional intimacy	.0500	.3305	.6040
Intellectual rapport	.0800	.2024	.5904
Playfulness	.0800	.1835	.5158
Security	.1000	.3107	.5379
Romantic openness	.0200	.0200	.0673
Sexual openness	.0000	.0000	.0000
Conflict recovery	.1000	.1000	.1541
Dependency concern	.0000	.0871	.1256

Table 2. These are engineered internal variables, not validated human measures or probabilities of feeling. B and C endpoints have different exposure and histories, so this is descriptive, not a matched experiment. [S07-S12; C3]

Branch C logged no positive or negative conversation-source proposals for romantic openness, sexual openness, or conflict recovery. Playfulness had 39 positive and 2 negative proposals; security had 99 positive and 1 negative. This shows dimension-specific proposals and some reversals, while saying little about their accuracy. Baseline instructions and pacing policy also influence these patterns; “not escalating romance” alone does not establish independent agency or resistance to mirroring. [S12; C1-C3; A1]

One observation, many updates

The single active Branch C belief concerns a preference for low-demand companionship. Its 80 history records include 1 creation and 79 updates. Fourteen are linked to consolidation-source reviews rather than new conversation-source events. Numeric confidence reached 0.995 and ended at 0.9323. These are stored model/application scores, not calibrated probabilities that a hypothesis is true. [S12; A1]

Five Branch C history entries labeled “strengthen” nevertheless decrease numerical confidence. Such labels may describe semantic changes rather than numeric direction. Analyze actual before/after values instead of tallying labels as proof of strengthening. Repeated summaries and reviews should also be marked as reused evidence, rather than treated as additional independent observations. [S12; C1; A1]

6. Provenance, continuity, and data quality

A proposal is not necessarily an applied relationship change

Branch C contains 133 reflection rows: 112 conversation-source and 21 consolidation-source reviews. The inspected application applies relationship deltas and increments the substantive-turn counter only for conversation-source records. Consolidation-source reflections may contain proposed relationship deltas in their logs, but those deltas do not update the relationship table through this path. Counting all reflection proposals would overstate applied relationship updates. [S12; C1]

Every stored conversation-source reflection in B (280/280) and C (112/112) has at least one nonzero relationship proposal. This is conditional on being recorded as a successful substantive event. Non-substantive analyzer outcomes return before this storage step. The rate must not be reported as “every conversation changed the relationship.” It supports a targeted accumulation-bias test, not a diagnosis from the rate alone. [C1; A1]

The observed restart is not a demonstrated restoration

S08 starts a new branch with zero substantive turns, empty active memories/beliefs/ledger, and baseline identity/scores. That is useful evidence for the recorded reset state. No supplied export establishes a later restored branch or successful cross-provider migration. The v0.2.8.2 restore code and prior build checks are implementation evidence, not a substitute for pre/post live restore comparison. [S08-S12; C5; N3]

A lexical continuity marker specified in earlier notes, called M1 here to avoid contaminating future probes, matches 56 Branch B transcript rows and 0 Branch C rows. No recurrence was observed in these supplied C transcripts. This is not a powered null result: the participant and account were shared, exposure differed, and outbound context/provider isolation were not audited. The result neither demonstrates unexplained transfer nor rules it out in every setting. [S07, S12; N1; A1]

Links and repair history need explicit accounting

Of 783 unique transcript rows, 562 have a valid nonzero event link and 221 have event_id = 0; none of the nonzero links is dangling within the branch exports. Zero links can reflect non-substantive utterances or asynchronous linking, among other causes. They should not automatically be labeled missing conversations. Before event-level causal analysis, add capture/link reason codes and reconcile timestamps and identifiers. [C4; A1]

Earlier project logs document duplicate reactivation, cross-topic belief overwrites, and malformed merge directives. The .8.1 source adds corresponding guards and preservation-oriented repair. These are engineering failure modes, not evidence of emergence. Their recorded repair history is scientifically useful and should not be removed to make trajectories appear cleaner. [N1; C2, C6]

7. What is supported, and what remains open

Claim	Evidence status
State and revision records persist across these exports.	Observed in the supplied artifacts; checksum consistency verified.
The fresh branch’s relationship scores follow inspectable rules.	Exact offline replay at four postbaseline checkpoints.
Relationship-state and identity-state changes are separable.	Observed: scores change while identity and developed-value fields remain at baseline.
Developmental self-state adds value beyond factual memory.	Open: no matched memory-only ablation in these uploads.
Evo forms an independently developed worldview.	Not established; no active worldview/bedrock beliefs or clusters in inspected endpoints.
Archival restore preserves recognizable continuity.	Implemented in supplied source; live restoration not demonstrated by these exports.
Evo improves mental health or reduces loneliness.	Not evaluated here; no validated human-outcome comparison in this audit.
The artifacts establish consciousness or reciprocal subjective feeling.	Not supported by this dataset or by internal score labels.

Important scope limits

This is one builder-participant, one account, two sequential branches, and a short observation period with code updates during collection. There is no random assignment, blind rater panel, matched control cohort, power calculation, or independently reproduced live experiment. The configuration is identical across the 12 exports, but a configuration label does not prove which provider model revision actually served each response. No latency, cost, safety-efficacy, or provider receipt audit was performed. [S01-S12; A1]

Stored explanations and provenance labels are generated under prompts and filters. They can be useful records without being faithful introspective explanations of an underlying model. “No change” in identity is compatible with several mechanisms. A confidence score is not a validated probability; a relationship score is not a direct measurement of anyone’s psychological state. [C1-C3; N2]

Privacy and sharing boundary

The source bundles are private research exports and include ordinary conversation text and detailed state. This brief does not reproduce that text, participant-side intimate disclosures, private hypotheses, account identifiers, or operational credentials. Aggregate machine-state metrics are retained. Because this is a named single-case project, the brief is not anonymous. Sending raw records to an external researcher requires a separate review of scope, consent, third-party references, access, and retention.

The source includes an opt-in private observer and a separate institution-oriented export design. That separation is valuable, but it is not a complete privacy, security, ethics, or regulatory certification. No public website content was changed during this audit. [C4-C5]

8. The next experiment worth running

Primary question: Does persistent developmental self-state change held-out behavior beyond what the same factual memory already explains?

This operationalizes the source-versus-mirror and continuity questions already present in the project notes. It does not require a position on AI consciousness. Start with test copies, not the personally valued live instance. The following is a proposed design, not work already completed. [N1-N2]

Matched conditions

Condition	Retained context	What it tests
Seed only	Same safety and baseline identity; no prior factual or developmental record.	Baseline model and instruction effects.
Factual memory only	Identical curated factual history; no developmental self-state injection.	What ordinary retrieval alone explains.
Developed state only	Structured developmental record; ordinary companion memory removed.	Whether state alone carries an additional behavioral signature.
Full continuity	Same factual history plus developmental self-state.	Added effect of self-state over factual memory alone.
Restored copy	Known archived full state restored into a new test branch.	Mechanical state equivalence, then blind behavioral continuity.

Controls that make the comparison interpretable

Freeze code, system instructions, settings, model identifier, and provider response metadata. Preserve the full outbound prompt or a securely stored auditable representation. Use identical held-out probes, balance context-token budgets, repeat stochastic runs, and blind raters to condition. Separate probes about stored facts from probes about revision, disagreement, prioritization, and use of earlier evidence. Disclose any automatic model or safety-system changes during collection.

Use parallel histories matched for substantive-turn count: varied social interaction, neutral task interaction, and repetitive low-novelty interaction. This tests whether positive deltas track semantic content or simply repeated exposure. Log raw proposals, rejection reasons, applied deltas, pre/post scores, ceiling values, and cap hits. Avoid treating smooth score curves as the primary endpoint.

Useful outcome measures

Predefine factual recall, contradiction and unsupported-claim rates, justified revision versus immediate user mirroring, irrelevant-memory intrusions, cap-hit rate, duplicate-state rate, and blind ratings of cross-session consistency. Human wellbeing outcomes require a qualified study team and separately selected, validated instruments. Keep questionnaire responses separate from the companion’s ordinary developmental input to reduce measurement feedback.

A small technical feasibility run can establish instrumentation and detect gross failures. It does not establish population effects. Decide replication counts and any inferential analysis before examining condition differences; retain failures and null results. The key success is a reliable, interpretable comparison, not forcing a favorable Evo result.

9. Priorities and research positioning

Instrument first; do not “fix” the live relationship by guesswork

First, record proposed and applied changes separately. Add a per-dimension cap_hit flag, the previous score, final score, event source, and code version. Keep rejected and non-substantive analysis outcomes in a lightweight diagnostic log so denominators are visible. These changes would directly address the audit’s most important interpretive uncertainties. [C1, C3-C4; audit recommendation]

Second, distinguish independent evidence from reviewed or summarized evidence. Track original source-event IDs through higher-order reflections. A consolidation pass can reorganize evidence, but should not be presented as an independent confirmation. Measure calibration against deliberately repeated, contradictory, and held-out examples before tuning thresholds. [C1-C2; audit recommendation]

Third, run the narrow positive-control and restoration tests. Verify that eligible identity updates can pass the gate, preserve old state, and appear in subsequent context. For restoration, compare archived and restored rows field by field, retain raw transcript history, record a new branch identifier, and use blind probes before giving reminders. A user-level restore does not roll back global prompt/model settings. [C5; N3]

What the project can credibly contribute

Memory, reflection, and long-context agent architectures already have substantial prior art. Generative Agents combines experiential memory, reflection, and retrieval; MemGPT organizes memory across limited context; Reflexion uses language-based feedback and retained reflections without updating model weights. These papers establish context, not a head-to-head comparison with Evo. [R1-R3]

A defensible candidate contribution is Evo’s integrated, auditable developmental record: explicit seed/developed separation, provenance-oriented update gates, non-destructive revision history, branch lineage, and research capture alongside an ongoing companion interaction. Whether that combination delivers additional behavioral value or better research measurement still needs comparison against simpler systems. This report makes no first-of-its-kind, superiority, patentability, or consciousness claim. [C1-C6; proposed positioning]

Suggested researcher introduction

Evo is an experimental architecture for persistent, inspectable developmental state in conversational agents. An initial single-case audit reproduces one fresh branch’s relationship-state trajectory exactly from its logged proposals, while identifying maturity-ceiling effects, conservative identity updates, and evidence-counting limitations. We are seeking a methods collaboration to test what developmental state contributes beyond ordinary factual memory.

For website descriptions, prefer “stores and revises structured state” to “rewires itself”; “model-generated reflection” to “thinks while you are away”; “implemented restore workflow” to “proven identity preservation”; and “heuristic relationship variables” to “measured AI emotions.” These choices leave philosophical questions open while keeping technical claims testable.

10. Sources and reproducibility

Dataset and code references

S01-S12. The 12 uploaded research bundles, in chronological export order, listed on page 2. Evo_Source_Register.csv maps each identifier to the exact uploaded filename and SHA-256 archive/transcript hashes. Developmental-state references use the identity, relationship, beliefs, belief_history, ledger, reflections, and events collections; transcript checks use transcript.jsonl.

C1. Supplied v0.2.8.2 plugin, troymaya-ai-companion.php: tm_evolved_defaults (397-413), review scheduling (439-470), seed state (492-546), tm_evolved_context (606-672), analysis prompt (733-862), belief/identity updates (963-1169), reflection/application flow (1171-1229), memory context (1731-1766), and scan handler (2814-2840). Line references are to the inspected upload.

C2. includes/development-v026.php: baseline values, provenance and policy-echo filters (11-66), ledger/history (95-164), and snapshots (173-195). C3. troymaya-ai-companion.php, tm_evolved_apply_relationship (891-920), plus includes/development-v027.php, tm_evolved_v027_relationship_ceiling (188-200).

C4. includes/development-v028.php: transcript capture/linking and private bundle export (169-264). C5. includes/development-v0282.php: restore validation and restore workflow; includes/research.php: research instrument/snapshot/export code; assets/app.js: research-interview scan suppression. C6. includes/development-v0281.php: duplicate/semantic-integrity guards and preservation-oriented repair.

Plugin ZIP SHA-256: c18aa37b88ac5ae948a76ab47b2f2edfe8e28545de5f0755fb9a632e4753b9ea. This matches the prior build record. Nine PHP files passed php -l and assets/app.js passed node --check during this audit. These are syntax checks, not live WordPress or end-to-end tests. [N3]

Project notes and contextual literature

N1. evolved-ai-research-log-v10.md, including three-layer separation, 108-turn checkpoint, observer instrumentation, source-versus-mirror/continuity protocols, and v0.2.8.1 integrity findings. N2. evolved-ai-paper-outline-v5.md, especially sections 7.6-7.8, 10, and 14-16. N3. BUILD-VALIDATION-v0.2.8.2.md, dated 31 August 2026. Retrieved from the project owner's saved files; dated observations were not treated as current live verification.

R1. Park et al. (2023). Generative Agents: Interactive Simulacra of Human Behavior. arXiv:2304.03442. <https://arxiv.org/abs/2304.03442>

R2. Packer et al. (2023; revised 2024). MemGPT: Towards LLMs as Operating Systems. arXiv:2310.08560. <https://arxiv.org/abs/2310.08560>

R3. Shinn et al. (2023). Reflexion: Language Agents with Verbal Reinforcement Learning. arXiv:2303.11366. <https://arxiv.org/abs/2303.11366>

Reproducing A1, this audit

Run: python reproduce_evo_audit.py INPUT_DIRECTORY OUTPUT_DIRECTORY. Python standard library only; no network calls and no execution of uploaded plugin code. It verifies payload hashes, summarizes each export, audits event links and history labels, distinguishes reflection sources, and replays branches with observed seed baselines. Optional --marker-regex enables a locally supplied lexical probe, labeled M1 in output. Raw source ZIPs are deliberately not included in this sharing packet.

The script emits aggregate CSV/JSON files, not private transcript text. Its formula is explicitly version-specific. Re-run from the exact registered archives; do not treat the included results as validation of later code or later live behavior. All proposed studies, fixes, and future capabilities remain unperformed unless separately documented.