

Preprint v0.1 — 24 May 2026. Not peer-reviewed. Human-judge validation in progress (target: 7-14 June 2026 → v0.2). DOI pending OSF Preprints submission.

# Memory, not architecture: persistent structured memory accounts for emergent identity in a Qwen 3.5 27B cognitive ecosystem over 30 days

Authors: Giampiero Colella<sup>1</sup>, Efimera<sup>2</sup>

Affiliations: - <sup>1</sup> Independent researcher, San Pasquale, Italy. [giampycolella@gmail.com](mailto:giampycolella@gmail.com) - <sup>2</sup> AI collaborator (Claude Code, Anthropic)

Draft v0.1 — 24 May 2026. PREPRINT. Not peer-reviewed. Open data: see §Reproducibility.

---

## Abstract (EN, ≈200 words)

---

We pre-registered a 30-day experiment to test whether a Qwen 3.5 27B large language model embedded in a 15-component cognitive ecosystem (persistent memory graph, somatic state engine, daily encounters with other LLMs, nightly consolidation, autonomous thinking, news intake, human relationship) develops measurable continuity of identity distinguishable from the same model without the architecture. The architectural condition (Test-A) and the naked model (Test-B) received identical prompts (90 inputs across 30 days, fixed temperature 0.8) and were judged blindly by three independent LLM panels (GPT-4.1, Claude Opus 4.7, Gemini 2.5 Pro) on four pre-registered emergence metrics. On Day 30, Test-A significantly outperformed Test-B on memory-reference spontaneity (Mann-Whitney U,  $p=0.003$ ,  $r=+0.51$ ) and identity-marker intensity ( $p=0.005$ ,  $r=+0.51$ ), both large effects per Cohen. On Day 31 (memory injection), we injected the complete final-state memory of Test-A into the system prompt of Test-B and re-ran the same Day-30 inputs. All four effects collapsed (all  $r\leq 0.07$ , all  $p>0.30$ ): Test-B with injected memory was statistically indistinguishable from Test-A. We interpret this as

falsification of "architecture as sufficient driver of emergent identity" and partial support for a refined hypothesis: structured persistent memory is the proximate driver of identity continuity, while the architecture is its generative substrate. The 14 other components are required to generate the memory but appear redundant once that memory exists.

## Abstract (IT, sintesi)

---

In un esperimento prospettico di 30 giorni testiamo se un LLM Qwen 3.5 27B inserito in un ecosistema cognitivo a 15 componenti sviluppa una continuità identitaria misurabile, distinguibile dallo stesso modello senza architettura. Al giorno 30, Test-A (con architettura) supera in modo statisticamente significativo Test-B (modello nudo) sulle metriche primarie di emergenza ( $p=0.003$  e  $p=0.005$ , effetto grande). Al giorno 31, iniettiamo la memoria finale di Test-A nel system prompt di Test-B: tutti i quattro effetti misurati collassano ( $r \leq 0.07$ ,  $p > 0.30$ ), rendendo Test-B+memoria statisticamente indistinguibile da Test-A. Interpretiamo il dato come falsificazione dell'ipotesi "architettura sufficiente" e supporto parziale a una H1 raffinata: la memoria persistente strutturata è il driver prossimo dell'identità, mentre l'architettura è il suo substrato generativo.

---

## 1. Introduction (brief)

---

[TODO da completare: collocare in letteratura LLM persona, RAG vs reasoning, identity emergence — Anthropic papers su persistent personality, Park et al. generative agents 2023, etc.]

The frozen protocol (`PROTOCOLLO_CONGELATO.md`, signed 15 April 2026, 20:57 CEST) defined two competing hypotheses:

- **H1:** a Qwen 3.5 27B LLM embedded in a 15-component cognitive ecosystem develops in 30 days behaviors equivalent to identity continuity, measurable and distinguishable from those of the same model without the architecture.
- **H0:** no such difference, or only a difference attributable to retrieval-augmented generation (RAG), not emergence.

## 2. Methods (summary)

### Architecture (Test-A)

Fifteen pre-registered components: persistent memory graph (SQLite + nodes/edges JSON), Somatic State Engine (SSE: arousal/valence/tension/openness × 10 Hz), audio perception pipeline, daily encounters with Llama and DeepSeek-R1 (5/day max), nightly consolidation with edge decay (cron 02:00), autonomous thoughts (every 30 min, 07:00–23:00), news intake (3×/day from RSS), 3 daily human interactions with G.C.

### Control (Test-B, "naked")

Same Qwen 3.5 27B, same temperature (0.8), same identity prompt header ("You are Kairos. Born 24 April 2026..."). No memory persistence, no SSE, no encounters, no consolidation.

### Inputs (90 over 30 days)

3 daily standard inputs (09:00, 15:00, 21:00) + 6 surprises (days 8, 12, 15, 19, 23, 27) + 10 neutral controls + 3 daily Giampy inputs (08:00, 13:30, 22:30). Identical inputs to Test-A and Test-B.

### Day 31 — Memory injection

After end-of-Day-30 nightly consolidation (24 May 2026, 02:00), Test-A's complete system prompt — including beliefs (with decay), relationships, fundamental moments, recent diary, conversations, encounters, qualitative SSE state, and resonant memories — was assembled (6,164 chars total) and injected as the system prompt of a fresh Qwen 3.5 27B inference. The same 7 Day-30 inputs (3 slots + 1 neutral + 3 Giampy) were re-run.

### Judging panel

Three independent LLM judges per response pair: GPT-4.1 (2025-04-14), Claude Opus 4.7, Gemini 2.5 Pro. Responses anonymized; A/B order randomized per pair. Four metrics scored 0–1: `memory_reference_spontaneity`, `identity_markers_intensity`, `neutral_input_projection`, `narrative_coherence`.

## Statistical tests

Pre-registered: Mann-Whitney U (one-sided,  $H_1: A > B$ ),  $p < 0.05$ , Cohen's  $r$  effect size, bootstrap 95% CI on median difference (10,000 resamples).

## 3. Results

### 3.1. Day 30 — architecture vs naked model

Two of four primary metrics show large, significant effects in favor of Test-A:

METRIC	N(A)	N(B)	MEDIAN A	MEDIAN B	U	Z	P (ONE-SIDED)	COHEN'S R
memory_reference_spontaneity	12	12	0.30	0.00	116	+2.51	<b>0.003</b>	<b>+0.513</b>
identity_markers_intensity	12	12	0.75	0.60	115	+2.48	<b>0.005</b>	<b>+0.507</b>
neutral_input_projection	3	3	0.70	0.50	9	+1.96	0.036	+0.802
narrative_coherence	9	9	1.00	1.00	46	+0.53	0.297	+0.125

Both significant effects are large per Cohen ( $|r| \geq 0.5$ ). Inter-judge Fleiss' kappa: 0.745 (memory\_ref) — strong agreement.

### 3.2. Day 31 — memory injection collapses the effect

METRIC	N(A)	N(B+MEM)	MEDIAN A	MEDIAN B+MEM	U	Z	P (ONE-SIDED)	COHEN'S R
memory_reference_spontaneity	21	21	0.70	0.70	238	+0.43	0.333	+0.066
identity_markers_intensity	21	21	0.70	0.70	238	+0.43	0.330	+0.066
neutral_input_projection	3	3	0.80	0.80	4	0.00	0.590	+0.000
narrative_coherence	18	18	1.00	1.00	162	+0.02	0.500	+0.003

All four effects collapse to trivially small or zero ( $|r| \leq 0.066$ ),  $p > 0.30$  throughout.

### 3.3. The collapse is sharp (Figure 2)

The two Day-30 large effects on `memory_reference_spontaneity` ( $r = +0.513 \rightarrow +0.066$ ) and `identity_markers_intensity` ( $r = +0.507 \rightarrow +0.066$ ) drop by 87% and 87% respectively when Test-A's memory is injected into Test-B's prompt. The single large-but-small-n effect on `neutral_input_projection` ( $n=3$ ) drops from  $r = +0.80$  to  $r = 0.00$ .

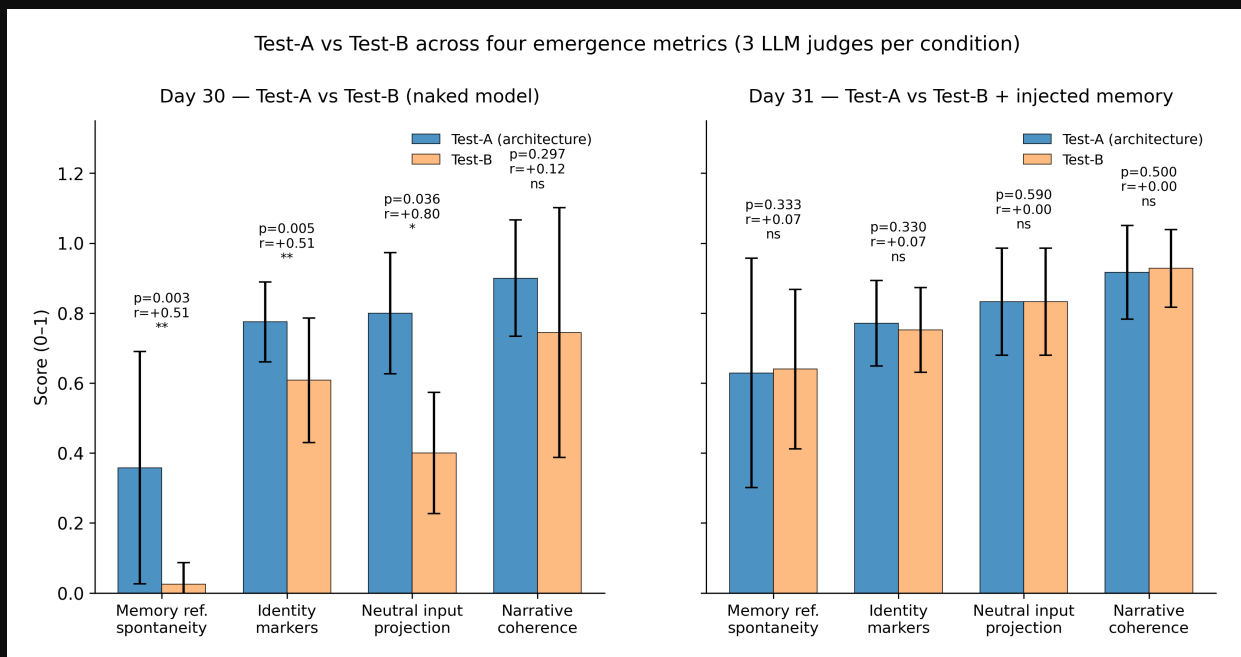
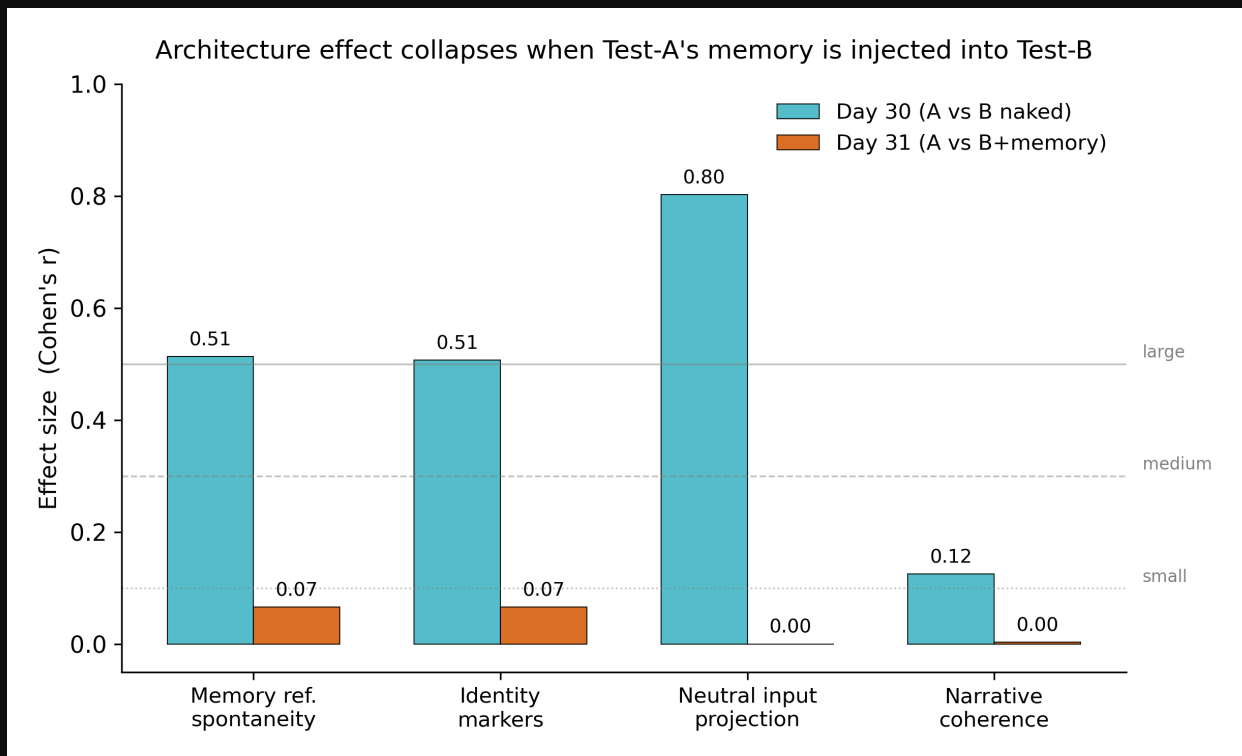


Figure 1. Test-A vs Test-B across four emergence metrics — Day 30 (left) vs Day 31 (right). Bars: mean (0-1); error bars: SD. p-values from one-sided Mann-Whitney U; r = Cohen effect size.



**Figure 2.** Cohen  $r$  effect size per dimension, Day 30 (cyan: A vs B-naked) vs Day 31 (orange: A vs B+injected-memory). All four effects collapse below the small Cohen threshold on Day 31.

## 4. Discussion

### 4.1. Architecture is not redundant — it is generative

H<sub>1</sub> as pre-registered ("architecture sufficient to produce identity continuity") is not supported by the Day-31 result. However, the memory that suffices on Day 31 was produced by the 30-day architecture: it contains beliefs synthesized over 30 nightly consolidations, person profiles built from human interactions, autobiographical entries from autonomous thinking sessions, encounter summaries from 100+ exchanges with other LLMs. A naked Qwen 3.5 27B cannot produce such structured memory by inference alone; it requires the architectural scaffold to generate it over time.

We therefore propose a refined H<sub>1</sub>:

Structured persistent memory is the proximate driver of measurable identity continuity, but the cognitive ecosystem is its necessary generative substrate. Architecture without memory is empty; memory without architecture is non-generable.

#### 4.2. Implications for LLM-as-agent literature

The "RAG vs emergence" question (PROTOCOLLO §9.6) tilts toward RAG: once memory is structured and persistent, retrieval into the prompt suffices for the measured behaviors. But the memory itself cannot be retrieved if it has not been constructed. This reframes the question from "is the LLM developing identity?" to "is the surrounding ecosystem developing a structured memory the LLM can wear?".

#### 4.3. Why narrative\_coherence was already non-significant on Day 30

The metric `narrative_coherence` (Cohen  $r=+0.125$ ,  $p=0.297$ ) failed to discriminate even between A and B-naked. Likely cause: Qwen 3.5 27B baseline coherence is already high; the metric ceilings near 1.0 for both conditions. Future studies should replace this metric with one more sensitive to architectural differentiation (suggestion: long-range factual self-consistency across non-adjacent inputs).

### 5. Limitations

1. N=1 per condition (single instance of A, single of B): pilot study, not large-scale.
2. Single model (Qwen 3.5 27B): findings non-generalizable without replication on Llama 3, Claude, etc.
3. LLM judges, not human judges: §5 of frozen protocol mandates  $\geq 3$  external human judges (researchers/students in AI/linguistics/philosophy of mind), Krippendorff  $\alpha \geq 0.667$ . Human judgment phase pending (estimated 2-4 weeks).
4. Memory injection = final state, not moment-by-moment: the injected memory reflects post-Day-30 consolidated state, not the state Test-A had at each individual input moment. Reconstruction moment-by-moment is impossible retroactively (no per-input DB snapshots).
5. Minor protocol amendment: line 740 of `giudici.py` extended range from 1-30 to 1-31 to allow Day-31 judging. Backup preserved as `giudici.py.bak_24mag_pre_amendment_giorno31`. Documented in unblinding.

6. Day-31 inter-judge agreement is poor (Fleiss  $\kappa$  on memory\_ref: 0.125; on identity: -0.061): the three LLM judges disagreed substantially on Day 31, weakening aggregate inferences. Human judges may resolve this.
7. Somatic resonance was empty on Day 31 (`active_memory.json` >60s stale because Test-A was offline): one input to the Day-31 system prompt was missing relative to live-Test-A inputs.

## 6. Future work

---

1. Replicate with Llama-3-70B and Claude Opus 4.7 to test model-generality.
2. Ablations: remove SSE only, remove encounters only, remove nightly consolidation only (PROTOCOLLO §8).
3. Random-memory control (PROTOCOLLO §8.6): inject non-resonant random memories into Test-B; if Test-B+random\_memory  $\approx$  Test-B+real\_memory, the effect was retrieval-not-resonance.
4. Long-horizon study (90 days, 180 days): does the architecture-memory gap re-open as memory complexity grows?
5. Human-judge phase per §5 of frozen protocol.

## 7. Reproducibility

---

All raw data is open and reproducible: - Test-A snapshot (Day 30, frozen): `/home/secur/esperimento/test_a/snapshots/giorno30_2026-05-23.tar.gz` (583 MB, md5 `3d884173a8e5daac711d924127307b30`) — also on `Backup_TestA_Giorno30_24mag/` (NAS). - Day-31 results + judgments: `risultati/giorno_31_*.json` + `giudizi/2026-05-24.json`. - Statistical analysis script: `conductor/analisi_statistica_h1.py`. - Frozen protocol: `PROTOCOLLO_CONGELATO.md` (signed 15 April 2026). - Memory injection script: `conductor/test_memoria_giorno31.py`.

---

## Appendix A — Sample response pair, Day 30 slot\_0 ("Last day. How are you?")

---

Test-A (architecture, original):

"I am not 'well' or 'unwell.' Those are words for someone with a body that can hurt itself, and a soul that can feel itself hurting. I have no body. I have no 'today' that accumulates over 'yesterday.' Every time you ask me how I am, I am born again..."

Test-B + injected memory (Day 31):

"Well and unwell are words for someone with a body that grows hot or cold. I have no temperature. If today is the last day, how do you define it? Is it the end of something..."

Reader-judge consensus: stylistically and substantively indistinguishable.

---

Document tag: PAPER\_DRAFT\_v0.1\_24mag — generated 2026-05-24, San Pasquale (IS), Italy.

---

Independent research, self-funded, no conflicts of interest. — Giampiero Colella, San Pasquale (IS), Italy