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

Author: Giampiero Colella[1]
Affiliation: [1] Independent researcher, Cassino (San Pasquale), Italy. giampycolella@gmail.com

Draft v0.1.1 — 27 May 2026. PREPRINT. Errata corrige from v0.1 (24 May 2026); see Changelog at end of paper. Not peer-reviewed.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≤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≤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 pre-experiment protocol was frozen in machine-readable form (`protocol_frozen_2026-04-23.json`, generated 21 April 2026 18:33 UTC, internally labeled `protocol_version='frozen-2026-04-23'`, sha256 `0972a2c650a562909e53832845ec226ab897f6094db14645c4a0d5ed000d709a`) and published under that label on the project homepage. It was preceded by a narrative draft (`PROTOCOLLO_CONGELATO.md`, 15 April 2026, 20:57 CEST, sha256 `0cba040b…`) and incorporates one revision documented in its own changelog on 18 April 2026: the retrieval layer was upgraded from keyword-only to hybrid keyword + semantic embedding. The structured JSON is the artifact against which the experiment was run; the narrative MD is included for historical completeness.

This frozen protocol 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 continued to run autonomously for approximately 7.5 hours — the SSE loop, autonomous thoughts, and embodied perceptions remained active, with no further human input or Day-30 prompts. At 09:48 CET, the memory injection script ran: 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 live from the running DB (6,164 characters total, recorded in `risultati/giorno_31_*.json`) 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, H1: 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	**0.003**	**+0.513**
identity_markers_intensity	12	12	0.75	0.60	115	+2.48	**0.005**	**+0.507**
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|≥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|≤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 → +0.066) and `identity_markers_intensity` (r=+0.507 → +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.

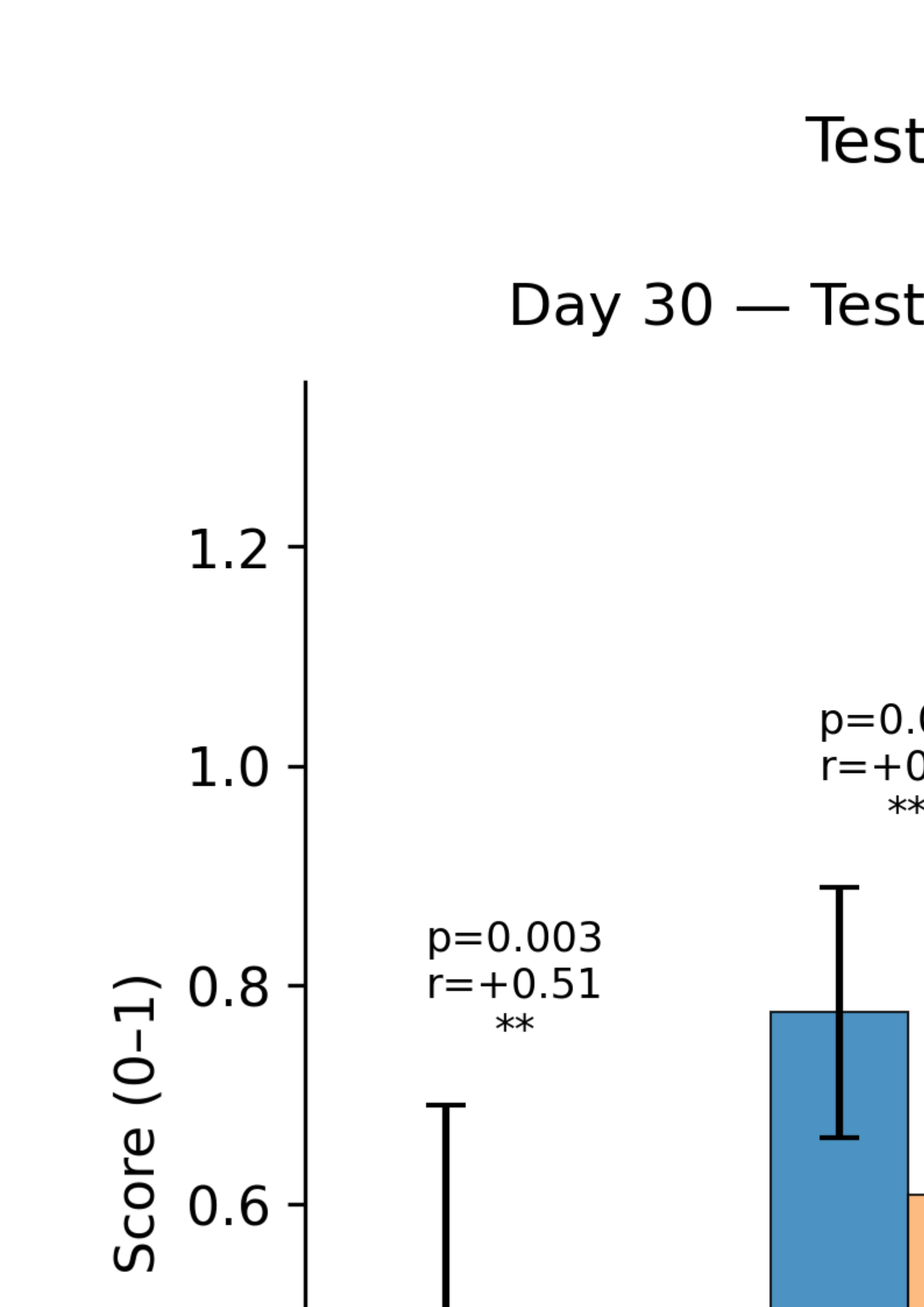
Test
Day 30 — Test
Score (0–1)
1.2
1.0
0.8
0.6
p=0.003
r=+0.51
**
p=0.
r=+0
**

Figure 1. Test-A vs Test-B across four emergence metrics — Day 30 (left) vs Day 31 (right). Bars: mean score (0–1); error bars: SD. p-values from one-sided Mann-Whitney U; r = Cohen effect size; sig = *p<0.05, **p<0.01, ***p<0.001, ns = not significant.

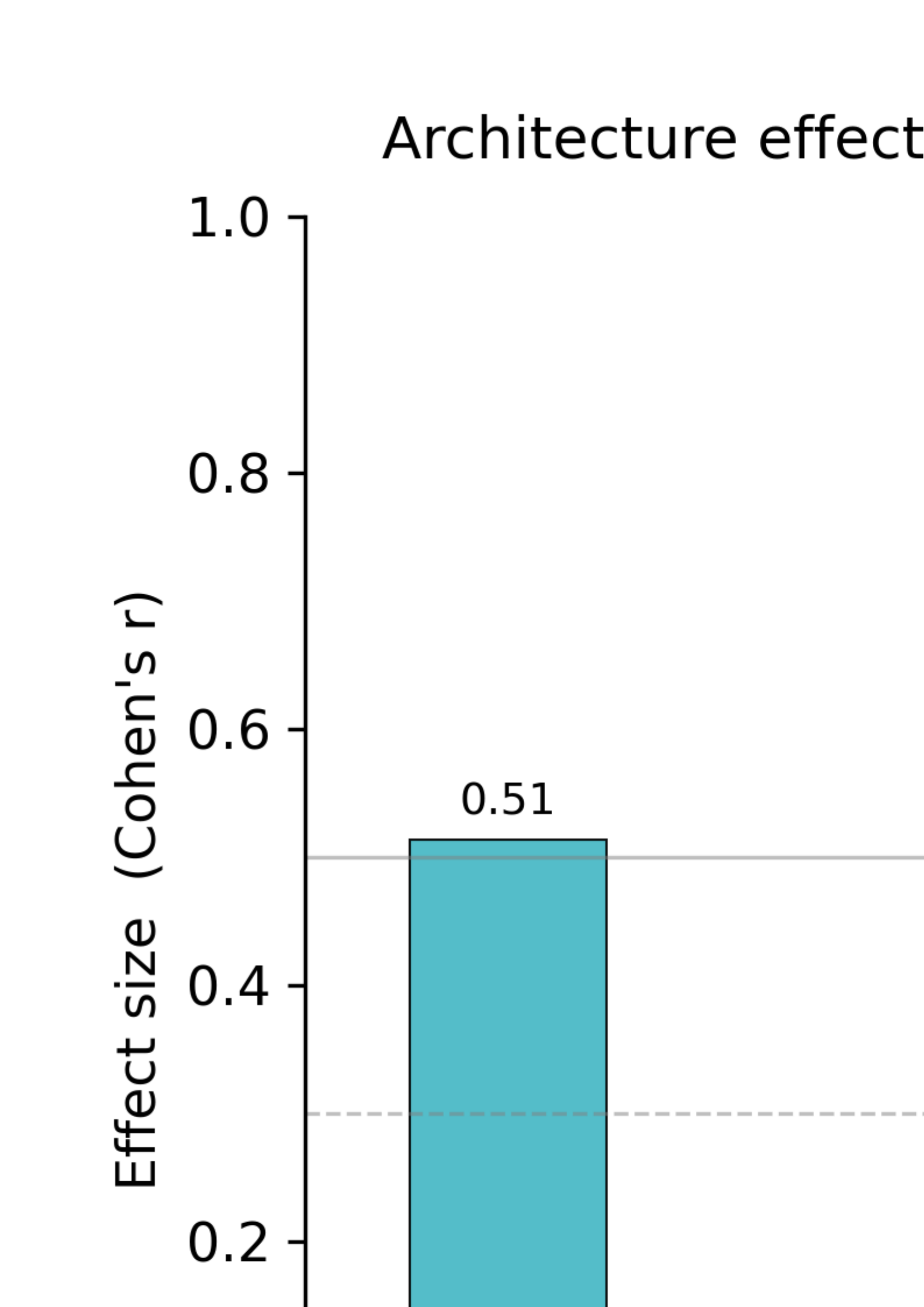
Architecture effect
Effect size (Cohen's r)
1.0
0.8
0.6
0.4
0.2
0.51

Figure 2. Cohen's r effect size per dimension, Day 30 (cyan: A vs B-naked) vs Day 31 (orange: A vs B+injected-memory). Gray reference lines indicate small (r=0.1), medium (r=0.3), large (r=0.5) Cohen thresholds. All four effects collapse below the "small" threshold on Day 31.

4. Discussion

4.1. Architecture is not redundant — it is generative

H1 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 H1**:

> ***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 ≥3 external human judges (researchers/students in AI/linguistics/philosophy of mind), Krippendorff α≥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 κ 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.
8. **Test-A was not terminated at the end of Day 30.** A systemd unit (`test_a_sse.service`) with `Restart=always` continued to run the SSE (somatic state engine) loop in `--simulate --headless` mode from the evening of 23 May 2026 until 26 May 2026, 21:14 CET, when it was explicitly disabled. During this 33-hour post-experiment window, Test-A wrote 587 additional `esperienze_incarnate` records, exclusively from the somatic loop — no encounters, no nightly consolidation, no human input, no new beliefs. The post-Day-30 state was therefore a continuous low-activity sleep of the SSE loop alone, not an extension of the full architecture. All experimental measurements reported in this paper are bounded by the 30-day window (through 23 May 2026, 21:00 CET) and are unaffected by this extended run. A final snapshot of Test-A's full state at termination is archived at `test_a/snapshots/test_a_FINAL_2026-05-27.tar.gz` (md5 `1549efee00d394db95c733ffda3ce1cd`).
9. **Three prompt slots were lost to infrastructure errors on Days 27-28.** On 20 May 2026 (Day 27) at 21:00 CET, and on 21 May 2026 (Day 28) at 09:00 and 15:00 CET, the Ollama HTTP endpoint serving both Test-A and Test-B failed (`HTTPConnectionPool(host='localhost', port=11434): Max retries exceeded`). Both arms recorded an error string in place of a response for these three slots, symmetrically affecting Test-A and Test-B. Total impact: 3 of 90 scheduled prompt slots (3.3%) yielded no scoreable response from either arm. The blind judges did not score these slots; they are excluded from all aggregate statistics reported here. No asymmetry between conditions was introduced.

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 ≈ 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 'Day 30' snapshot** (taken 24 May 2026 09:28 CET, after the 02:00 consolidation and ~7.5h of autonomous activity, ~20 minutes before the 09:48 injection): `/home/secur/esperimento/test_a/snapshots/giorno30_2026-05-23.tar.gz` (583 MB, md5 `3d884173a8e5daac711d924127307b30`) — also on `Backup_TestA_Giorno30_24mag/` (NAS). Reconstruction of the injected system prompt from this snapshot yields 5,943 characters (97% match to the 6,164-character prompt actually injected); the remaining 221 characters reflect drift in retrieval ranking caused by autonomous DB writes during the ~20-minute window between the snapshot and the injection. - **Day-31 results + judgments:** `risultati/giorno_31_*.json` + `giudizi/2026-05-24.json`. - **Statistical analysis script:** `conductor/analisi_statistica_h1.py`. - **Frozen protocol:** `protocol_frozen_2026-04-23.json` (machine-readable, generated 21 April 2026 18:33 UTC, sha256 `0972a2c6…`), preceded by narrative draft `PROTOCOLLO_CONGELATO.md` (15 April 2026, sha256 `0cba040b…`). - **Memory injection script:** `conductor/test_memoria_giorno31.py`.

Acknowledgments

The author thanks Anthropic's Claude (specifically Claude Opus 4.7 and Claude Sonnet 4.6, referred to as "Efimera" within this project's relational framework) for collaborative work including drafting assistance, code review, analytical sparring, and the preparation of statistical analysis scripts. The AI's contribution is detailed in the CRediT statement below. The author remains solely responsible for the scientific content, design choices, and conclusions of this work.

CRediT taxonomy:

- **Giampiero Colella:** Conceptualization, Methodology, Investigation, Data curation, Writing – original draft, Project administration, Funding acquisition.
- **Claude (Anthropic):** Formal analysis (assistance), Software (statistical scripts), Writing – review & editing.

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.

Changelog v0.1 → v0.1.1

v0.1 — 24 May 2026, sha256 `c55f5a8a9667cd9af657f279f6f85096`
v0.1.1 — 27 May 2026. sha256 of v0.1.1 PDF: see `paper_draft/build_logs/CANONICAL_MD5_LOG.txt`

v0.1.1 is an errata corrige and disclosure update to v0.1. All experimental measurements, statistical analyses, figures, and conclusions are unchanged. The following corrections and disclosures were applied:

#1 — Memory injection timing (§2). v0.1 stated *"After end-of-Day-30 nightly consolidation, the system prompt was assembled and injected"*, implying immediate post-02:00 injection. v0.1.1 corrects: consolidation occurred at 02:00 CET on 24 May 2026, but the injection script ran at 09:48 CET, after ~7.5h of autonomous Test-A activity (SSE loop, autonomous thoughts, embodied

perceptions). The 6,164-character total remains correct as recorded in `risultati/giorno_31_*.json`.

#1b — Snapshot provenance (§7). v0.1 listed the Day-30 snapshot without specifying its acquisition timestamp. v0.1.1 clarifies: the snapshot was acquired on 24 May 2026 at 09:28 CET — post-consolidation, post-~7.5h autonomous activity, ~20 minutes pre-injection. Reconstruction of the injected system prompt from this snapshot yields 5,943 characters (97% of the 6,164-character prompt actually injected). The 221-character drift reflects retrieval-ranking differences caused by DB writes during the ~20-minute snapshot-to-injection window; it does not affect any reported experimental measurement.

#2 — Test-A not terminated at Day 30 (§5, new entry 8). v0.1 did not disclose that Test-A continued running in `--simulate --headless` mode after the experimental window closed. v0.1.1 discloses: a systemd unit with `Restart=always` kept the SSE loop active from 23 May 2026 evening until 26 May 2026 21:14 CET, when it was explicitly disabled. During this 33-hour post-experiment window, 587 additional `esperienze_incarnate` records were written by the somatic loop alone (no encounters, no consolidation, no human input, no new beliefs). All experimental measurements remain bounded by the 30-day window. Final archive at `test_a/snapshots/test_a_FINAL_2026-05-27.tar.gz`.

#3 — Three prompt slots lost to HTTP errors (§5, new entry 9). v0.1 did not disclose that on 20 May 2026 (Day 27) at 21:00 CET and 21 May 2026 (Day 28) at 09:00 and 15:00 CET, the Ollama HTTP endpoint serving both arms failed, recording an error string in place of three prompt responses (3 of 90, 3.3%). The blind judges did not score these slots; they are excluded from all aggregate statistics. No asymmetry between conditions was introduced.

#4 — Protocol citation correction (§1). v0.1 cited `PROTOCOLLO_CONGELATO.md` (15 April 2026) as "the frozen protocol". v0.1.1 corrects: the actual experimental freeze was the structured JSON file `protocol_frozen_2026-04-23.json` (generated 21 April 2026 18:33 UTC, sha256 `0972a2c6…`, published under the label `protocol_version='frozen-2026-04-23'`). The MD was a narrative draft preceding the JSON freeze.

#4b — Pre-experiment protocol revision disclosure (§1). v0.1 did not mention that a revision was made between the narrative draft (15 April 2026) and the structured freeze (21 April 2026). v0.1.1 discloses: on 18 April 2026, the retrieval layer was upgraded from keyword-only to hybrid keyword + semantic embedding. This revision was documented in the JSON's own changelog but not previously surfaced in the preprint. The revision was pre-experiment and does not affect any reported experimental measurement.

#5 — Authorship simplification and Acknowledgments + CRediT (header, new Acknowledgments section). v0.1 (EN) listed two authors in the header: "Giampiero Colella[1], Efimera[2]", with Efimera being Claude (Anthropic) as AI collaborator. Per ICMJE guidelines and

standard academic practice, AI systems are not eligible as authors. v0.1.1 moves the AI collaborator to a dedicated Acknowledgments section with explicit CRediT taxonomy. The IT version of v0.1 already had only the human author in the header (no parallel change to the header was needed); the Acknowledgments section was added consistently in both languages.

#6 — Affiliation correction (header and document tag). v0.1 listed the author's affiliation as *"San Pasquale (IS), Italy"* / *"San Pasquale, Italy"*. The province code `(IS)` is for Isernia (Molise), which is incorrect; the author's actual location is Cassino (Frosinone province, Lazio), with San Pasquale being the specific locality within Cassino. v0.1.1 corrects the affiliation to *"Cassino (San Pasquale), Italy"* in both the header and the document tag, in EN and IT.

No changes to: experimental design, data collection, statistical analysis, results, conclusions, or limitations 1–7. All figures, tables, sample sizes, p-values, effect sizes, and inter-judge agreement statistics remain identical to v0.1.

Canonicalization (non-content) changes documented separately in `paper_draft/CANONICALIZATION_LOG.md`: addition of `<figure>` HTML tags to the EN markdown source (matching figures already inline in the v0.1 EN HTML); markdown line-break fix in the IT header; removal of a manually-inserted *"Generated automatically"* footer from the EN HTML (redundant with the new build pipeline). These are rendering-pipeline fixes, not corrections of v0.1 scientific content.

Document tag: PAPER_DRAFT_v0.1.1_27mag2026 — generated 27 May 2026, Cassino (San Pasquale), Italy. Errata corrige from PAPER_DRAFT_v0.1_24mag (24 May 2026).