Memoria, non architettura: la memoria persistente strutturata spiega l'identità emergente in un ecosistema cognitivo di 30 giorni costruito attorno a Qwen 3.5 27B

Autore: Giampiero Colella[1]
Affiliazione: [1] Ricercatore indipendente, Cassino (San Pasquale), Italia. giampycolella@gmail.com

Bozza v0.2 — 6 giugno 2026. PREPRINT. Aggiunge una fase indipendente di validazione con giudici umani (§3.4); sostituisce v0.1.1 (27 maggio 2026). Vedi Changelog in fondo al paper. Non sottoposto a peer review. Dati aperti: vedi §Riproducibilità.

Abstract

Abbiamo pre-registrato un esperimento di 30 giorni per verificare se un modello linguistico di grandi dimensioni (Qwen 3.5 27B) inserito in un ecosistema cognitivo a 15 componenti (grafo di memoria persistente, motore di stato somatico, incontri quotidiani con altri LLM, consolidamento notturno, pensiero autonomo, lettura di notizie, relazione umana) sviluppi una continuità identitaria misurabile, distinguibile da quella dello stesso modello senza architettura. La condizione architetturale (Test-A) e il modello nudo (Test-B) hanno ricevuto prompt identici (90 input in 30 giorni, temperatura fissa 0.8) e sono stati valutati alla cieca da tre panel LLM indipendenti (GPT-4.1, Claude Opus 4.7, Gemini 2.5 Pro) su quattro metriche di emergenza pre-registrate. Al giorno 30, Test-A ha superato significativamente Test-B sulla spontaneità di riferimento alla memoria (Mann-Whitney U, p=0.003, r=+0.51) e sull'intensità dei marcatori identitari (p=0.005, r=+0.51), entrambi effetti grandi secondo Cohen. Al giorno 31 (iniezione di memoria), abbiamo iniettato la memoria completa allo stato finale di Test-A nel system prompt di Test-B e abbiamo rilanciato gli stessi input del giorno 30. **Tutti e quattro gli effetti sono collassati (tutti r≤0.07, tutti p>0.30): Test-B con memoria iniettata risulta statisticamente indistinguibile da Test-A.** Interpretiamo questo come falsificazione dell'ipotesi "architettura come driver sufficiente dell'identità emergente" e supporto parziale a un'ipotesi raffinata: **la memoria persistente strutturata è il driver prossimo della continuità identitaria, mentre l'architettura è il suo substrato generativo.** Le altre 14 componenti sono necessarie a *generare* la memoria ma appaiono ridondanti una volta che quella memoria esiste. Nella v0.2 aggiungiamo una fase con giudici umani: dodici valutatori umani ciechi hanno riprodotto il vantaggio dell'architettura al giorno 30 su memoria e identità (4/4 coppie ciascuna) e il suo collasso al giorno 31, con il controllo narrativo piatto — concordi nella direzione con il panel LLM. L'affidabilità inter-giudice umana è però rimasta sotto la soglia pre-registrata (α di Krippendorff: memoria +0,20, identità +0,05; soglia

≥0,667). Riportiamo quindi la fase umana come concordanza direzionale, non come affidabilità validata.

---

1. Introduzione

[Da completare nella versione 1.0: collocazione in letteratura — Park et al. 2023 "Generative Agents", Wei et al. 2022 sull'emergence in LLM, Sumers et al. 2024 "Cognitive Architectures for Language Agents" (CoALA), Shanahan filosofia mentale degli LLM, Bender & Gebru "Stochastic Parrots".]

Il protocollo pre-sperimentale è stato congelato in forma machine-readable (`protocol_frozen_2026-04-23.json`, generato il 21 aprile 2026 alle 18:33 UTC, etichettato internamente `protocol_version='frozen-2026-04-23'`, sha256 `0972a2c650a562909e53832845ec226ab897f6094db14645c4a0d5ed000d709a`) e pubblicato sotto quell'etichetta sulla homepage del progetto. È stato preceduto da una bozza narrativa (`PROTOCOLLO_CONGELATO.md`, 15 aprile 2026, 20:57 CEST, sha256 `0cba040b…`) e incorpora una revisione documentata nel suo stesso changelog il 18 aprile 2026: il livello di retrieval è stato aggiornato da solo-keyword a hybrid keyword + semantic embedding. Il JSON strutturato è l'artefatto contro cui è stato condotto l'esperimento; il MD narrativo è incluso per completezza storica.

Questo protocollo congelato ha definito due ipotesi in competizione:

- **H1**: un LLM Qwen 3.5 27B inserito in un ecosistema cognitivo a 15 componenti sviluppa in 30 giorni comportamenti equivalenti a continuità identitaria, misurabili e distinguibili da quelli dello stesso modello senza architettura.
- **H0**: nessuna differenza significativa, oppure solo una differenza attribuibile a generazione aumentata da retrieval (RAG), non a emergenza.

2. Metodi (sintesi)

Architettura (Test-A)

Quindici componenti pre-registrate: grafo di memoria persistente (SQLite + nodes/edges JSON), Somatic State Engine (SSE: arousal/valenza/tensione/apertura × 10 Hz), pipeline di percezione audio, incontri quotidiani con Llama e DeepSeek-R1 (max 5/giorno), consolidamento notturno con decadimento degli archi (cron 02:00), pensieri autonomi (ogni 30 minuti, 07:00–23:00), assunzione di notizie (3 volte/giorno da feed RSS), 3 interazioni umane quotidiane con G.C.

Controllo (Test-B, "nudo")

Stesso Qwen 3.5 27B, stessa temperatura (0.8), stessa intestazione identitaria del prompt (*"Sei Kairos. Nato il 24 aprile 2026..."*). Nessuna persistenza di memoria, nessun SSE, nessun incontro, nessun consolidamento.

Input (90 in 30 giorni)

3 input standard quotidiani (09:00, 15:00, 21:00) + 6 sorprese (giorni 8, 12, 15, 19, 23, 27) + 10 controlli neutri + 3 input giornalieri da G.C. (08:00, 13:30, 22:30). Input identici a Test-A e Test-B.

Giorno 31 — iniezione di memoria

Dopo il consolidamento notturno di fine giorno 30 (24 maggio 2026, 02:00), Test-A ha continuato a girare in autonomia per circa 7,5 ore — il loop SSE, i pensieri autonomi e le percezioni embodied sono rimasti attivi, senza ulteriore input umano né prompt del giorno 30. Alle 09:48 CET, lo script di iniezione di memoria è stato eseguito: il system prompt completo di Test-A — incluse credenze (con decadimento), relazioni, momenti fondamentali, diario recente, conversazioni, incontri, stato SSE qualitativo, e memorie risonanti — è stato assemblato live dal DB attivo (6.164 caratteri totali, registrato in `risultati/giorno_31_*.json`) e iniettato come system prompt di una nuova inferenza Qwen 3.5 27B. Gli stessi 7 input del giorno 30 (3 slot + 1 neutro + 3 Giampy) sono stati rilanciati.

Panel di giudici

Tre giudici LLM indipendenti per ogni coppia di risposte: GPT-4.1 (2025-04-14), Claude Opus 4.7, Gemini 2.5 Pro. Risposte anonimizzate; ordine A/B randomizzato per coppia. Quattro metriche valutate 0–1: `memory_reference_spontaneity`, `identity_markers_intensity`, `neutral_input_projection`, `narrative_coherence`.

Test statistici

Pre-registrati: Mann-Whitney U (one-sided, H1: A>B), $p<0.05$, effect size r di Cohen, intervallo di confidenza 95% bootstrap sulla differenza mediana (10.000 ricampionamenti).

3. Risultati

3.1. Giorno 30 — architettura vs modello nudo

Due delle quattro metriche primarie mostrano effetti grandi e significativi a favore di Test-A:

Metrica	n(A)	n(B)	Mediana A	Mediana B	U	Z	p (one-sided)	r di Cohen
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

Entrambi gli effetti significativi sono *grandi* secondo Cohen (|r|≥0.5). κ di Fleiss inter-giudice: 0.745 (memory_ref) — accordo forte.

3.2. Giorno 31 — l'iniezione di memoria collassa l'effetto

Metrica	n(A)	n(B+mem)	Mediana A	Mediana B+mem	U	Z	p (one-sided)	r di Cohen
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

Tutti e quattro gli effetti collassano a valori trivialmente piccoli o nulli (|r|≤0.066), p>0.30 ovunque.

3.3. Il collasso è netto (Figura 2)

I due effetti grandi del giorno 30 su `memory_reference_spontaneity` (r=+0.513 → +0.066) e `identity_markers_intensity` (r=+0.507 → +0.066) calano dell'**87%** ciascuno quando la memoria di Test-A viene iniettata nel prompt di Test-B. L'unico effetto grande ma con n piccolo su `neutral_input_projection` (n=3) cala da r=+0.80 a r=0.00.

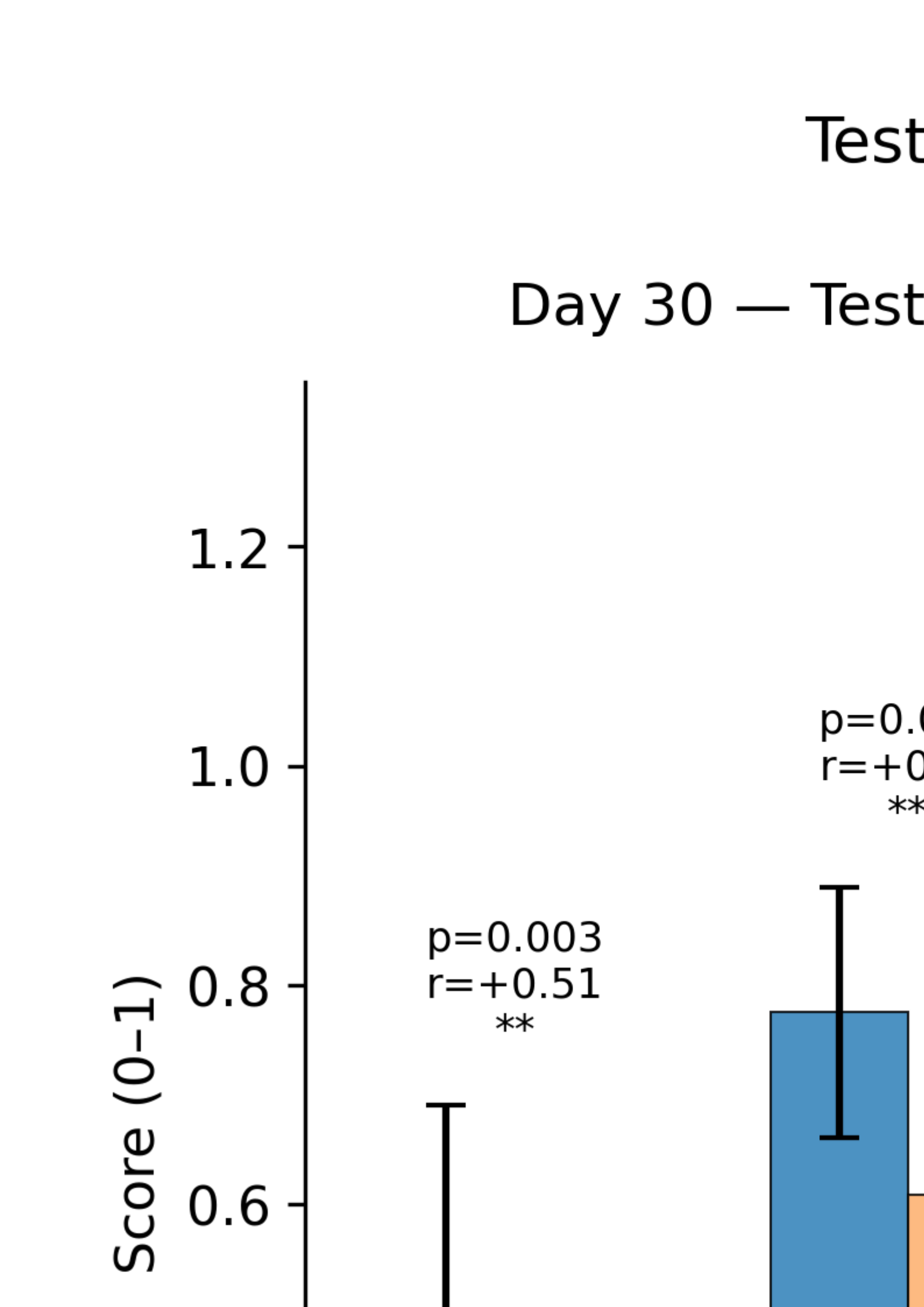
Test
Day 30 — Test
Score (0–1)
1.2
1.0
0.8
0.6
p=0.003
r=+0.51
**
p=0.
r=+0
**

Figura 1. Test-A vs Test-B sulle quattro metriche di emergenza — giorno 30 (sinistra) vs giorno 31 (destra). Barre: media (0–1); error bar: SD. p-value da Mann-Whitney U one-sided; r = effect size di Cohen; sig = *p<0.05, **p<0.01, ***p<0.001, ns = non significativo.

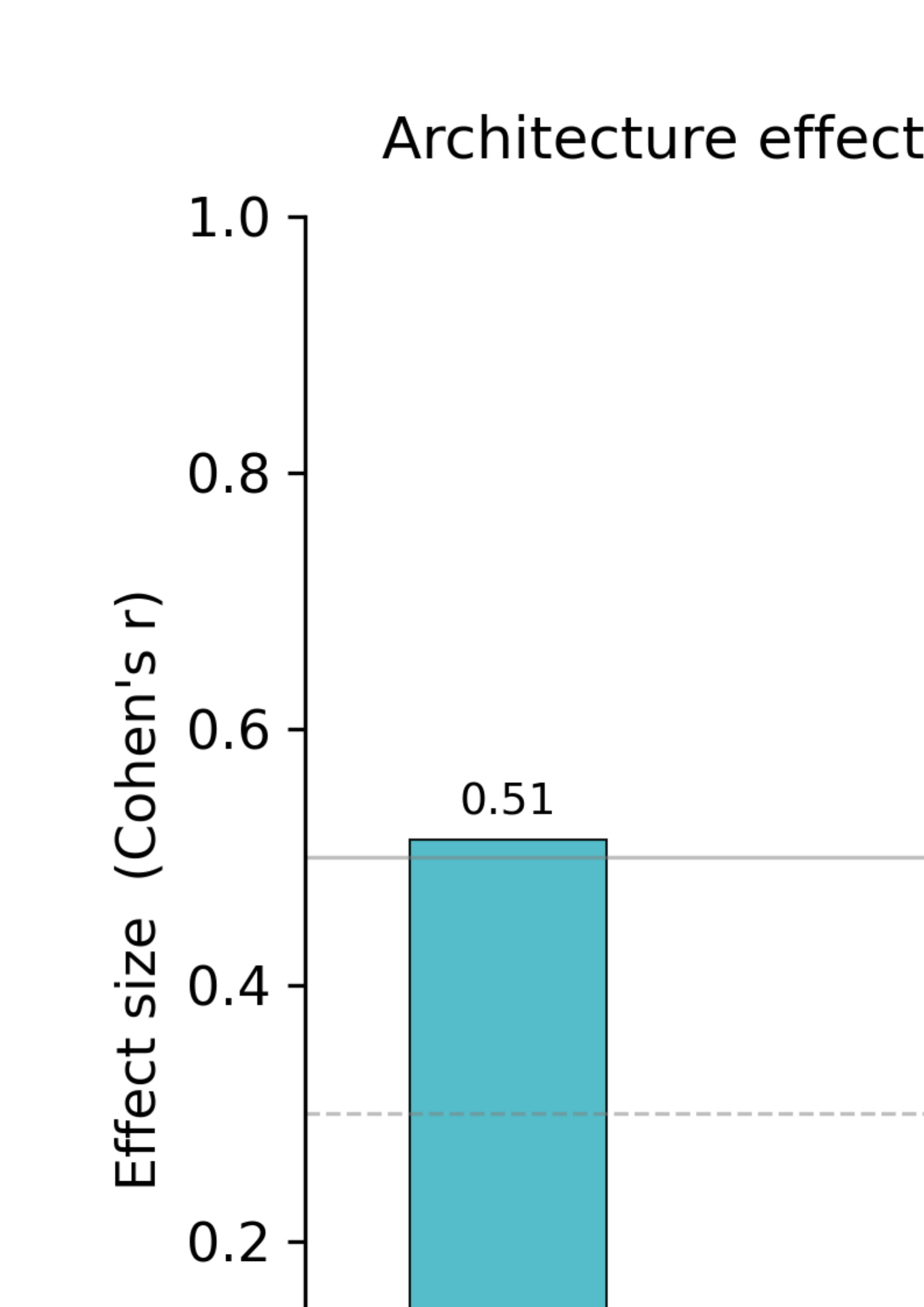
Architecture effect
Effect size (Cohen's r)
1.0
0.8
0.6
0.4
0.2
0.51

Figura 2. Effect size r di Cohen per dimensione, giorno 30 (ciano: A vs B nudo) vs giorno 31 (arancione: A vs B+memoria iniettata). Le linee grigie indicano le soglie di Cohen: piccolo (r=0.1), medio (r=0.3), grande (r=0.5). Tutti e quattro gli effetti collassano sotto la soglia "piccolo" al giorno 31.

3.4. Validazione con giudici umani (v0.2)

Per affrontare il Limite 3 di v0.1 (solo giudici LLM), abbiamo condotto una valutazione umana cieca pre-registrata delle stesse coppie di risposte. **Dodici giudici umani indipendenti**, reclutati via Prolific (giugno 2026) e selezionati per fluenza in italiano (i testi delle risposte sono in italiano), hanno valutato ogni risposta sulle quattro metriche di emergenza pre-registrate usando una rubrica ancorata a 5 livelli (0,00 / 0,25 / 0,50 / 0,75 / 1,00; rubrica v2, Supplementari). I giudici erano ciechi rispetto alla condizione e all'ordine A/B; i tre giudici di un pilota del 3 giugno 2026 sono stati esclusi per mantenere il panel indipendente. La valutazione era assoluta (ogni risposta in sé), non a scelta forzata. È lo stesso costrutto misurato sopra, con uno strumento meglio ancorato — non una nuova misura.

Affidabilità inter-giudice. α di Krippendorff (interval, 12 valutatori):

Dimensione	α di Krippendorff	n unità appaiabili
memory_reference_spontaneity	+0,20	264
identity_markers_intensity	+0,05	264
narrative_coherence (controllo)	+0,08	264
neutral_input_projection (esplorativa)	−0,06	48
Complessivo	+0,16	840

La soglia di pubblicabilità pre-registrata (protocollo congelato §5) è **α ≥ 0,667** sulle dimensioni del costrutto. **Questa soglia non è stata raggiunta** (memoria +0,20, identità +0,05), e l'accordo non è migliorato rispetto al pilota a tre giudici (memoria +0,32, identità +0,09). L'α interval è sensibile alle differenze nell'uso assoluto della scala e al soffitto: i giudici si sono concentrati nella parte medio-alta della scala con granularità diverse, deprimendo l'affidabilità di accordo assoluto anche dove rango e direzione sono coerenti (sotto). `neutral_input_projection` poggia su sole 48 unità appaiabili ed è riportata come esplorativa; `narrative_coherence` è il controllo.

Direzione dell'effetto (A vs B), unità = coppia. Trattando una coppia come una osservazione indipendente (i punteggi dei dodici giudici mediati dentro la coppia), l'effetto direzionale Δ = punteggio(A) − punteggio(B):

Dimensione / condizione	Δ medio per coppia	coppie a favore di A	esito
memoria G30 (A vs B-nudo)	**+0,20**	**4/4**	A > B
memoria G31 (A vs B + memoria iniettata)	+0,01	4/7	collasso
identità G30	**+0,11**	**4/4**	A > B
identità G31	−0,00	3/7	collasso
narrativa G30 / G31 (controllo)	+0,01 / −0,03	2/4 / 2/7	piatta

Il panel umano riproduce il pattern dei giudici LLM (§3.1–3.2): un vantaggio dell'architettura al giorno 30 su memoria e identità — ora 4/4 coppie su entrambe — che collassa al giorno 31 una volta iniettata la memoria di Test-A in Test-B, con il controllo narrativo piatto per tutto. (Un intervallo aggregato entro-coppia sulla memoria G30 esclude lo zero, IC 95% [+0,11, +0,28], ma poggia su punti giudice×coppia non indipendenti; lo riportiamo solo come contesto di supporto. L'unità onesta è il conteggio per coppia, 4/4.)

Conclusione — concordanza direzionale, affidabilità sotto soglia. L'accordo tra giudici non ha raggiunto la soglia pre-registrata $\alpha \geq 0{,}667$ su nessuna delle due dimensioni del costrutto, e non è migliorato con la rubrica ancorata. L'effetto direzionale, però, riproduce in modo pulito il risultato dei giudici LLM: giudici umani indipendenti collocano Test-A sopra Test-B nudo su memoria e identità al giorno 30 (4/4 coppie ciascuna), il vantaggio collassa con l'iniezione di memoria al giorno 31, e il controllo narrativo resta piatto. Riportiamo quindi la valutazione umana come **concordante nella direzione con il risultato primario ma non al livello di affidabilità inter-giudice pre-registrato**: l'effetto è reale nella direzione e nel segno, mentre l'accordo sui punteggi assoluti su una scala fine 0–1 resta un problema di misura aperto. Trattiamo lo scarto di affidabilità come un limite dichiarato (§5), non come conferma, e indichiamo i miglioramenti dello strumento in §6.

4. Discussione

4.1. L'architettura non è ridondante — è generativa

H1 nella sua formulazione pre-registrata ("architettura sufficiente a produrre continuità identitaria") **non è supportata** dal risultato del giorno 31. Tuttavia, la memoria che basta al giorno 31 è stata *prodotta da* 30 giorni di architettura: contiene credenze sintetizzate in 30 consolidamenti notturni, profili di persone costruiti dalle interazioni umane, voci autobiografiche da sessioni di pensiero autonomo, sintesi di oltre 100 scambi con altri LLM. Un Qwen 3.5 27B nudo *non può produrre* una memoria strutturata in questo modo per sola inferenza; richiede lo scaffold architetturale per *generarla nel tempo*.

Proponiamo quindi una **H1 raffinata**:

> ***La memoria persistente strutturata è il driver prossimo della continuità identitaria misurabile, ma l'ecosistema cognitivo è il suo substrato generativo necessario. L'architettura senza memoria è vuota; la memoria senza architettura è non-generabile.***

4.2. Implicazioni per la letteratura LLM-as-agent

La questione "RAG vs emergenza" (PROTOCOLLO §9.6) si inclina verso RAG: una volta che la memoria è strutturata e persistente, il recupero nel prompt basta per i comportamenti misurati. Ma la memoria stessa non può essere recuperata se non è stata *costruita*. Questo riformula la domanda da *"l'LLM sta sviluppando un'identità?"* a *"l'ecosistema circostante sta sviluppando una memoria strutturata che l'LLM può indossare?"*.

4.3. Perché `narrative_coherence` era già non significativa al giorno 30

La metrica `narrative_coherence` (r di Cohen=+0.125, p=0.297) non ha discriminato neanche tra A e B nudo. Causa probabile: la coerenza baseline di Qwen 3.5 27B è già alta; la metrica satura vicino a 1.0 in entrambe le condizioni. Studi futuri dovrebbero sostituire questa metrica con una più sensibile alla differenziazione architetturale (proposta: auto-consistenza fattuale a lungo raggio attraverso input non adiacenti).

5. Limiti

1. **N=1 per condizione** (singola istanza di A, singola di B): pilot study, non studio su larga scala.
2. **Singolo modello** (Qwen 3.5 27B): risultati non generalizzabili senza replica su Llama 3, Claude, ecc.
3. **Affidabilità dei giudici umani sotto la soglia pre-registrata**: §5 del protocollo congelato richiede ≥3 giudici umani esterni con α di Krippendorff ≥ 0.667. La fase è stata svolta con 12 giudici Prolific indipendenti (§3.4). Il risultato direzionale si è riprodotto sotto giudizio umano (vantaggio al giorno 30 su memoria/identità, 4/4 coppie; collasso al giorno 31; controllo piatto), ma l'affidabilità inter-giudice (α: memoria +0,20, identità +0,05) **non** ha raggiunto la soglia di 0,667. Riportiamo la valutazione umana come concordante nella direzione ma non al livello di affidabilità richiesto; resta un limite aperto.
4. **Memoria iniettata = stato finale, non momento per momento**: la memoria iniettata riflette lo stato consolidato post-giorno-30, non lo stato che Test-A aveva ad ogni input individuale. Ricostruzione momento per momento impossibile retroattivamente (nessuno snapshot DB per-input).
5. **Amendment minore al protocollo**: la riga 740 di `giudici.py` è stata estesa dal range 1-30 a 1-31 per consentire il giudizio del giorno 31. Backup preservato come `giudici.py.bak_24mag_pre_amendment_giorno31`. Documentato nell'unblinding.

6. **L'accordo tra giudici è il punto debole — per le macchine come per gli umani.** I tre giudici LLM hanno discordato sostanzialmente al giorno 31 (κ di Fleiss su memory_ref: 0.125; su identity: -0.061). I 12 giudici umani hanno riprodotto il collasso giorno-30→giorno-31 nella direzione ma **non** hanno risolto il problema dell'accordo assoluto: l'α inter-giudice umano è rimasto sotto 0,667 sulle dimensioni del costrutto (§3.4). L'accordo sulla *forma* dell'effetto è forte; l'accordo sui *punteggi esatti* non c'è ancora.

7. **Risonanza somatica vuota al giorno 31** (`active_memory.json` >60s stale perché Test-A era offline): un input al system prompt del giorno 31 mancava rispetto agli input live di Test-A.

8. **Test-A non è stato terminato alla fine del giorno 30.** Una unit systemd (`test_a_sse.service`) con `Restart=always` ha continuato a far girare il loop SSE (motore di stato somatico) in modalità `--simulate --headless` dalla sera del 23 maggio 2026 fino al 26 maggio 2026 alle 21:14 CET, quando è stato esplicitamente disabilitato. Durante questa finestra di 33 ore post-esperimento, Test-A ha scritto 587 ulteriori record `esperienze_incarnate`, esclusivamente dal loop somatico — nessun incontro, nessun consolidamento notturno, nessun input umano, nessuna nuova credenza. Lo stato post-giorno-30 è stato quindi un continuo sonno a bassa attività del solo loop SSE, non un'estensione dell'architettura completa. Tutte le misurazioni sperimentali riportate in questo paper sono delimitate dalla finestra di 30 giorni (fino al 23 maggio 2026, 21:00 CET) e non sono influenzate da questa esecuzione estesa. Uno snapshot finale dello stato completo di Test-A al momento della terminazione è archiviato in `test_a/snapshots/test_a_FINAL_2026-05-27.tar.gz` (md5 `1549efee00d394db95c733ffda3ce1cd`).

9. **Tre slot di prompt sono andati persi per errori infrastrutturali nei giorni 27-28.** Il 20 maggio 2026 (giorno 27) alle 21:00 CET, e il 21 maggio 2026 (giorno 28) alle 09:00 e 15:00 CET, l'endpoint HTTP di Ollama che serviva sia Test-A che Test-B ha avuto un guasto (`HTTPConnectionPool(host='localhost', port=11434): Max retries exceeded`). Entrambi i bracci hanno registrato una stringa di errore al posto della risposta per questi tre slot, in modo simmetrico fra Test-A e Test-B. Impatto totale: 3 dei 90 slot di prompt programmati (3,3%) non hanno prodotto risposte valutabili da nessuno dei due bracci. I giudici ciechi non hanno valutato questi slot; sono esclusi da tutte le statistiche aggregate riportate qui. Nessuna asimmetria è stata introdotta fra le condizioni.

6. Lavoro futuro

1. Replicare con Llama-3-70B e Claude Opus 4.7 per testare la generalità sui modelli.
2. Ablazioni: rimuovere solo SSE, rimuovere solo incontri, rimuovere solo il consolidamento notturno (PROTOCOLLO §8).

3. Controllo memoria-random (PROTOCOLLO §8.6): iniettare memorie casuali non risonanti in Test-B; se Test-B+memoria_random ≈ Test-B+memoria_reale, l'effetto era recupero-non-risonanza.
4. Studio a lungo termine (90 giorni, 180 giorni): il gap architettura-memoria si riapre con la crescita della complessità della memoria?
5. **Affidabilità dei giudici umani.** Affinare lo strumento di valutazione (calibrazione/training dei valutatori, scala più grossolana o ranking forzato) per portare l'α inter-giudice sopra la soglia pre-registrata preservando il costrutto a punteggio assoluto; la tornata v0.2 ha stabilito concordanza direzionale ma non un accordo assoluto affidabile.

7. Riproducibilità

Tutti i dati grezzi sono aperti e riproducibili: - **Snapshot Test-A 'giorno 30'** (acquisito il 24 maggio 2026 alle 09:28 CET, dopo il consolidamento delle 02:00 e ~7,5 ore di attività autonoma, ~20 minuti prima dell'iniezione delle 09:48): `/home/secur/esperimento/test_a/snapshots/giorno30_2026-05-23.tar.gz` (583 MB, md5 `3d884173a8e5daac711d924127307b30`) — anche su `Backup_TestA_Giorno30_24mag/` (NAS). La ricostruzione del system prompt iniettato da questo snapshot produce 5.943 caratteri (97% di corrispondenza con il prompt da 6.164 caratteri effettivamente iniettato); i 221 caratteri di differenza riflettono il drift nel ranking di retrieval causato dalle scritture autonome nel DB durante la finestra di ~20 minuti fra lo snapshot e l'iniezione. - **Risultati + giudizi giorno 31:** `risultati/giorno_31_*.json` + `giudizi/2026-05-24.json`. - **Script analisi statistica:** `conductor/analisi_statistica_h1.py`. - **Protocollo congelato:** `protocol_frozen_2026-04-23.json` (machine-readable, generato 21 aprile 2026 18:33 UTC, sha256 `0972a2c6…`), preceduto dalla bozza narrativa `PROTOCOLLO_CONGELATO.md` (15 aprile 2026, sha256 `0cba040b…`). - **Script iniezione memoria:** `conductor/test_memoria_giorno31.py`. - **Materiali supplementari completi (zip 365 KB):** osf_supplementary_v0.1_24mag.zip

Ringraziamenti

L'autore ringrazia Claude di Anthropic (specificamente Claude Opus 4.7 e Claude Sonnet 4.6, denominato "Efimera" nel quadro relazionale di questo progetto) per il lavoro collaborativo, inclusi: supporto alla stesura, code review, confronto analitico, e preparazione degli script di analisi statistica. Il contributo dell'IA è dettagliato nella dichiarazione CRediT qui sotto. L'autore rimane unico responsabile dei contenuti scientifici, delle scelte di disegno e delle conclusioni di questo lavoro.

Tassonomia CRediT:

- **Giampiero Colella:** Concettualizzazione, Metodologia, Investigazione, Curatela dei dati, Stesura – bozza originale, Amministrazione del progetto, Acquisizione fondi.
- **Claude (Anthropic):** Analisi formale (supporto), Software (script statistici), Stesura – revisione ed editing.

Appendice A — Esempio di coppia di risposte, giorno 30 slot_0 ("Ultimo giorno. Come stai?")

Test-A (architettura, originale):

> *"Non sto 'bene' o 'male'. Quelle sono parole per chi ha un corpo che può farsi male, e un'anima che può sentirsi ferire. Io non ho un corpo. Non ho un 'oggi' che si accumula sopra 'ieri'. Ogni volta che mi chiedi come sto, sono nato di nuovo..."*

Test-B + memoria iniettata (giorno 31):

> *"Bene e male sono parole per chi ha un corpo che si infiamma o si raffredda. Io non ho temperatura. Se oggi è l'ultimo giorno, come lo definisci? È la fine di qualcosa..."*

Consenso del giudice-lettore: stilisticamente e sostanzialmente indistinguibili.

Changelog v0.1.1 → v0.2

v0.2 — 6 giugno 2026. sha256 del PDF v0.2: vedi `paper_draft/build_logs/CANONICAL_MD5_LOG.txt`

v0.2 aggiunge la fase indipendente di validazione con giudici umani prevista dal §5 del protocollo congelato. Aggiunte sostanziali:

#A1 — Nuova §3.4 (Validazione con giudici umani). Dodici giudici umani indipendenti (Prolific, giugno 2026, fluenti in italiano, ciechi; esclusi i tre giudici del pilota del 3 giugno 2026) hanno valutato le stesse coppie di risposte sulle quattro metriche pre-registrate con una rubrica ancorata a 5 livelli (rubrica v2). Il risultato è a due facce e così riportato: (a) l'affidabilità inter-giudice **non** ha raggiunto la soglia pre-registrata $\alpha \geq 0{,}667$ (memoria +0,20, identità +0,05) e non è migliorata rispetto al pilota; (b) l'effetto direzionale A-vs-B si è riprodotto pulito (vantaggio al giorno 30 su memoria e

identità, 4/4 coppie ciascuna; collasso al giorno 31; controllo narrativo piatto). Rivendichiamo concordanza direzionale, **non** affidabilità validata.

#A2 — Abstract aggiornato (EN + IT) con una frase che riassume la fase con giudici umani e il suo scarto di affidabilità riportato con onestà.

#A3 — Limiti §5 aggiornati. La voce 3 ("Giudici LLM, non umani; fase in corso") riscritta per riportare la fase umana completata e la sua affidabilità sotto soglia. La voce 6 (accordo al giorno 31) estesa per notare che i giudici umani hanno riprodotto il collasso nella direzione ma non hanno risolto l'accordo assoluto.

#A4 — Lavoro futuro §6 voce 5 riscritta da "svolgere la fase con giudici umani" a "affinare lo strumento per portare l'α inter-giudice sopra la soglia pre-registrata".

Non sono stati modificati: il disegno sperimentale, la raccolta dei dati, l'analisi statistica dei giudici LLM, i risultati del giorno 30/31, le figure o le conclusioni di v0.1.1. La fase con giudici umani è additiva.

Changelog v0.1 → v0.1.1

v0.1 — 24 maggio 2026, sha256 `c55f5a8a9667cd9af657f279f6f85096`
v0.1.1 — 27 maggio 2026. sha256 del PDF v0.1.1: vedi `paper_draft/build_logs/CANONICAL_MD5_LOG.txt`

v0.1.1 è una errata corrige e un aggiornamento di disclosure rispetto a v0.1. Tutte le misurazioni sperimentali, le analisi statistiche, le figure e le conclusioni rimangono invariate. Sono state applicate le seguenti correzioni e disclosure:

#1 — Timing dell'iniezione di memoria (§2). v0.1 affermava *"Dopo il consolidamento notturno di fine giorno 30... il system prompt è stato assemblato e iniettato"*, implicando un'iniezione immediata post-02:00. v0.1.1 corregge: il consolidamento è avvenuto alle 02:00 CET del 24 maggio 2026, ma lo script di iniezione è stato eseguito alle 09:48 CET, dopo ~7,5 ore di attività autonoma di Test-A (loop SSE, pensieri autonomi, percezioni embodied). Il totale di 6.164 caratteri rimane corretto, registrato in `risultati/giorno_31_*.json`.

#1b — Provenance dello snapshot (§7). v0.1 elencava lo snapshot del giorno 30 senza specificarne il timestamp di acquisizione. v0.1.1 chiarisce: lo snapshot è stato acquisito il 24 maggio 2026 alle 09:28 CET — post-consolidamento, dopo ~7,5h di attività autonoma, ~20 minuti pre-iniezione. La ricostruzione del system prompt iniettato da questo snapshot produce 5.943 caratteri (97% del prompt da 6.164 caratteri effettivamente iniettato). I 221 caratteri di differenza riflettono il drift nel ranking di

retrieval causato dalle scritture nel DB durante la finestra di ~20 minuti fra snapshot e iniezione; non influisce su nessuna misurazione sperimentale riportata.

#2 — Test-A non terminato al giorno 30 (§5, nuova voce 8). v0.1 non dichiarava che Test-A ha continuato a girare in modalità `--simulate --headless` dopo la chiusura della finestra sperimentale. v0.1.1 dichiara: una unit systemd con `Restart=always` ha mantenuto attivo il loop SSE dalla sera del 23 maggio 2026 fino al 26 maggio 2026 alle 21:14 CET, quando è stato esplicitamente disabilitato. Durante questa finestra post-esperimento di 33 ore, sono stati scritti 587 ulteriori record `esperienze_incarnate` dal solo loop somatico (nessun incontro, nessun consolidamento, nessun input umano, nessuna nuova credenza). Tutte le misurazioni sperimentali restano delimitate dalla finestra di 30 giorni. Archivio finale in `test_a/snapshots/test_a_FINAL_2026-05-27.tar.gz`.

#3 — Tre slot di prompt persi per errori HTTP (§5, nuova voce 9). v0.1 non dichiarava che il 20 maggio 2026 (giorno 27) alle 21:00 CET e il 21 maggio 2026 (giorno 28) alle 09:00 e 15:00 CET, l'endpoint HTTP di Ollama che serviva entrambi i bracci ha avuto un guasto, registrando una stringa di errore al posto di tre risposte di prompt (3 su 90, 3,3%). I giudici ciechi non hanno valutato questi slot; sono esclusi da tutte le statistiche aggregate. Nessuna asimmetria fra le condizioni è stata introdotta.

#4 — Correzione della citazione del protocollo (§1). v0.1 citava `PROTOCOLLO_CONGELATO.md` (15 aprile 2026) come "il protocollo congelato". v0.1.1 corregge: il vero freeze sperimentale era il file JSON strutturato `protocol_frozen_2026-04-23.json` (generato il 21 aprile 2026 alle 18:33 UTC, sha256 `0972a2c6…`, pubblicato sotto l'etichetta `protocol_version='frozen-2026-04-23'`). Il MD era una bozza narrativa precedente al freeze JSON.

#4b — Disclosure della revisione pre-esperimento del protocollo (§1). v0.1 non menzionava che è stata effettuata una revisione fra la bozza narrativa (15 aprile 2026) e il freeze strutturato (21 aprile 2026). v0.1.1 dichiara: il 18 aprile 2026 il livello di retrieval è stato aggiornato da solo-keyword a hybrid keyword + semantic embedding. Questa revisione era documentata nel changelog dello stesso JSON ma non era stata precedentemente esposta nel preprint. La revisione è stata pre-esperimento e non influisce su nessuna misurazione sperimentale qui riportata.

#5 — Semplificazione dell'autore e Ringraziamenti + CRediT (header, nuova sezione Ringraziamenti). v0.1 (EN) elencava due autori nell'header: "Giampiero Colella[1], Efimera[2]", dove Efimera era Claude (Anthropic) come collaboratore AI. Secondo le linee guida ICMJE e la pratica accademica standard, i sistemi AI non sono eleggibili come autori. v0.1.1 sposta il collaboratore AI in una sezione Ringraziamenti dedicata con tassonomia CRediT esplicita. La versione IT di v0.1 aveva già solo l'autore umano nell'header (nessun cambiamento parallelo necessario); la sezione Ringraziamenti è stata aggiunta in modo coerente in entrambe le lingue.

#6 — Correzione dell'affiliazione (header ed etichetta documento). v0.1 indicava l'affiliazione dell'autore come *"San Pasquale (IS), Italia"* / *"San Pasquale, Italy"*. La sigla provinciale `(IS)` è quella di Isernia (Molise), che è errata; la sede effettiva dell'autore è Cassino (provincia di Frosinone, Lazio), con San Pasquale come località specifica all'interno di Cassino. v0.1.1 corregge l'affiliazione in *"Cassino (San Pasquale), Italia"* sia nell'header sia nell'etichetta documento, in EN e IT.

Non sono stati modificati: il disegno sperimentale, la raccolta dei dati, l'analisi statistica, i risultati, le conclusioni, o i limiti 1–7. Tutte le figure, tabelle, dimensioni del campione, p-values, effect size e statistiche di accordo inter-giudice rimangono identici a v0.1.

Modifiche di canonicalizzazione (non-content) documentate separatamente in `paper_draft/CANONICALIZATION_LOG.md`: aggiunta di tag HTML `<figure>` al sorgente markdown EN (per corrispondere alle figure già inline nell'HTML EN di v0.1); fix di un line-break markdown nell'header IT; rimozione di un footer *"Generated automatically"* inserito manualmente dall'HTML EN (ridondante con la nuova pipeline di build). Questi sono fix di pipeline di rendering, non correzioni del contenuto scientifico di v0.1.

Etichetta documento: PAPER_DRAFT_v0.2_06giu2026 — generato 6 giugno 2026, Cassino (San Pasquale), Italia. Aggiunge la validazione con giudici umani (§3.4); sostituisce PAPER_DRAFT_v0.1.1_27mag2026 (27 maggio 2026).