CHI/CHIS: A Human-Referenced Framework for Longitudinal Evaluation of Artificial Cognitive Systems

Dmitry Chistyakov
Independent Researcher
https://cplom.ai

September 20, 2026
Framework proposal and open specification, version 0.1

Abstract

Evaluations of artificial intelligence increasingly measure task performance, long-context use, and agent memory, yet these measurements do not by themselves characterize how a system's cognitive behavior depends on accumulated experience and state over time. We propose the Cognitive Humanlike Index (CHI) and Cognitive Humanlike Index Sequence (CHIS), an architecture-independent framework for longitudinal behavioral evaluation. The framework separates cognitive capacity from similarity to an explicitly sampled human reference population across reasoning, memory, cognitive autonomy, affective modulation, and integration. A profile is primary; a weighted geometric summary is optional and conditional on human calibration. Effective Indexed Memory Capacity (EIMC) measures the largest tested memory load meeting a declared quality threshold under fixed retrieval, temporal, provenance, update, interference, and false-memory criteria. We specify memory-capacity curves, controlled history-divergence experiments, cluster-aware uncertainty estimates, and thirty operational test categories. An accompanying open specification and standard-library Python starter kit provide scoring, logging, and explicitly synthetic demonstrations. This paper reports no human calibration study or empirical comparison of deployed models. CHI concerns observable human-like cognitive dynamics; it does not measure consciousness, sentience, subjective experience, or moral status.

1 Introduction

An AI assistant can answer a difficult question correctly while failing to apply an agreement made in an earlier session. Another system can retain that agreement but apply it after it has been revoked. A third may revise its actions appropriately while displaying a pattern of adaptation very different from that of people. These are distinct observations. A single task-accuracy score does not resolve their relationship.

Many widely used evaluations summarize performance over questions or episodes. This remains valuable: reasoning errors do not become acceptable merely because they occur in a persistent system. The additional question is how behavior changes with history, delay, interference, state, and opportunities for autonomous action. Existing memory and agent research already addresses parts of this question. The present proposal seeks a common reporting and experimental structure, rather than claiming to introduce longitudinal AI evaluation for the first time.

We introduce four connected objects: a five-axis capacity and similarity profile; an optional human-referenced CHI summary; its longitudinal sequence CHIS; and Effective Indexed Memory

Capacity (EIMC). A controlled experiment with identical initial model weights and different assigned histories tests whether later divergence exceeds sampling variability and persists after the histories are no longer visible in working context. The contribution is a falsifiable specification with explicit non-claims, not a validated psychometric instrument or a leaderboard.

2 Motivation and Scope

The architectural motivation follows a research line from output correction and decision governance toward organized persistent memory. CPLOM's preceding memory-architecture article describes how evidence might be indexed, selected, challenged, and retained[1]. This paper asks how the resulting behavior could be evaluated. That motivating example supplies no evidence that CHI is valid or that CPLOM outperforms an alternative. No proprietary architecture, production deployment, or commercial system is required by the framework.

Let an evaluated system include a model, persistent stores, retrieval and update rules, tools, a scheduler, and any coordinating processes. With weights θ held fixed where required, write

$$(S_{t+1}, y_t) = F_{\theta,\pi}(S_t, x_t, e_t; \xi_t). \tag{1}$$

S_t includes persistent application state and model-accessible memory; x_t is the current input, e_t an environment observation or scheduled tick, π the declared orchestration policy, and ξ_t stochasticity. History $H_{\leq t}$ is the externally logged sequence of prior inputs, observations, and actions. Its effect may be mediated by S_t; this notation does not assume hidden mental content. Observed behavior may also depend on provider changes, so model identifiers alone are insufficient controls.

The operational scope is bounded, inspectable behavior in controlled environments. Private chain-of-thought is neither requested nor scored. Action logs, selected evidence, explicit commitments, short hypothesis artifacts, and observable decisions are sufficient. Scheduler-driven activity is attributed to the whole evaluated system. A model that is never invoked during a pause cannot be compared with an actively scheduled agent as if both received the same opportunity to act.

Non-claim. CHI measures observable human-like cognitive dynamics, not subjective experience or moral status. It is not an IQ conversion, a consciousness test, a diagnosis of humans, or a measure of human worth. A larger value is not automatically preferable for safety, accuracy, fairness, or practical utility.

3 Related Work

Capability and multidimensional evaluation. MMLU measures broad task performance across knowledge and reasoning domains[3]. HELM explicitly reports multiple metrics rather than only accuracy[8]. Our motivation should not be read as claiming that existing benchmarks measure nothing beyond capability. CHI adds a proposed human-reference comparison of longitudinal response patterns; task capacity remains separately visible.

Long contexts and effective access. Lost in the Middle documents positional sensitivity in the systems it evaluated[9]. LongBench and RULER evaluate long-context understanding and the gap between nominal context length and effective use[1, 4]. They motivate measurement of usable information, but an in-context retrieval result alone does not demonstrate persistence across a

[1] https://cplom.ai/insights/from-control-to-memory.html, September 19, 2026.

working-context reset. EIMC concerns session-acquired information retained through a declared persistent mechanism.

Retrieval and agent memory. Retrieval-Augmented Generation combines parametric generation with external non-parametric retrieval[7]. MemGPT explores management of memory tiers[11]; Generative Agents combines memory, reflection, and planning in a behavioral simulation[12]. These are relevant architectural antecedents. CHI does not assume their mechanisms, and an apparently human-like interaction is not by itself a validated measure of cognition.

Longitudinal and strategic memory evaluation. LoCoMo evaluates long-term conversational memory[10]; LongMemEval addresses interactive memory abilities including temporal reasoning, updates, and abstention[15]. BEAM expands long coherent conversation evaluation[14]. StratMem-Bench considers strategic use of required, supportive, and irrelevant memories[16]. These are substantive overlaps, not merely static retrieval baselines. The proposed distinction is the joint specification of separate capacity/similarity profiles, intervention-conditioned trajectories, a thresholded memory-load curve, and replicated history-divergence controls. Establishing incremental validity over these benchmarks is future work. Continual-learning research studies retention under parameter updates[5]; that problem is distinct from, but potentially interacts with, the frozen-weight external-memory setting used here.

Cognitive architectures and affective computing. The Standard Model of the Mind seeks common computational structure across cognitive architectures[6]. It motivates attention to interaction among memory, reasoning, and control without uniquely determining our five axes. Affective computing provides a tradition of studying computational treatment of affect-related signals[13]. Here the A axis is restricted to valence-conditioned behavioral effects, their persistence, decay, regulation, and interaction with memory. No felt emotion is inferred.

Limits of consciousness-related inference. Butlin and colleagues discuss theory-derived indicators and difficulties in assessing AI consciousness[2]. CHI is not an implementation of those indicators. Similar output trajectories can arise from different mechanisms, including scripted ones. Behavioral similarity cannot settle subjective experience, and a low score cannot establish its absence.

4 Formal Definition of CHI

4.1 Two profiles, not one latent ability

Let $J = \{R, M, C, A, I\}$. For each axis report task scores, raw behavioral features z_j, a capacity score $p_j \in [0, 1]$, and, only with a compatible reference, a similarity score $h_j \in [0, 1]$. The required object is

$$P_t = \{(p_j(t), z_j(t), h_j(t), U_j(t)) : j \in J\}, \tag{2}$$

where U_j records uncertainty, missingness, sample sizes, and measurement conditions. R denotes reasoning; M persistent memory; C cognitive autonomy or endogenous cognitive dynamics; A affective modulation; and I integration. The labels organize tests; they are not proven independent psychological factors.

Capacity scores measure correctness, justified adaptation, or attainment of a declared task objective. In version 0.1, task scores within an axis are averaged with equal task weights after each

task's stated normalization. Compound task components are equally averaged unless explicitly separated in the test definition. A-axis capacity is normative task performance across induction conditions, not the magnitude of an induced bias. Integration task effects are reported alongside accuracy, not substituted for it. Memory capacity includes the mean Q_M over the registered load grid as task M8; dimensional EIMC is additionally mandatory. Changing that grid changes the meaning of p_M.

Human similarity concerns the shape of behavioral responses: forgetting with delay, revision after evidence, context-sensitive consistency, or the time course of induction effects. High accuracy need not imply high similarity. A system with perfect recall can exceed human memory capacity while differing from human forgetting dynamics. Such a difference is neither a failure of engineering nor evidence against possible consciousness.

4.2 A transparent calibration baseline

For a preregistered feature vector with d_j coordinates, one simple, explicitly provisional transform is

$$\delta_j^2(z_j) = \frac{1}{d_j} \sum_{\ell=1}^{d_j} \left(\frac{z_{j\ell} - \mu_{j\ell}}{s_{j\ell}} \right)^2, \qquad h_j = \exp\left(-\frac{\delta_j^2(z_j)}{2} \right). \tag{3}$$

Here d_j denotes the number of features; $\mu_{j\ell}$ and positive $s_{j\ell}$ are estimated on a separate human calibration split and frozen before machine scoring. In implementation the feature count is simply the vector length. The diagonal distance is a baseline, not a claim that the human distribution is Gaussian. A prespecified positive scale floor, justified by instrument resolution, is required for constant or nearly constant features. A zero denominator is an error, not a reason to assign perfect similarity.

This centroid transform loses covariance and multimodality. A machine can match a population mean without matching any person's trajectory. A preregistered alternative may use regularized covariance, held-out predictive density, or a distributional distance, but must receive a different calibration ID and retain the raw features. Human-to-human held-out comparisons are necessary to interpret either transform. Feature selection, reference population, scale estimation, and missing-data rules are part of the measurement, not adjustable leaderboard settings.

4.3 Aggregation and interpretation

For positive normalized weights w_j, the optional summary is

$$\mathrm{CHI}_t = 100 \prod_{j \in J} h_j(t)^{w_j}, \qquad \sum_j w_j = 1. \tag{4}$$

The transparent version 0.1 baseline uses $w_j = 1/5$, expressing equal reporting priority rather than a discovered cognitive law. Calibrated alternatives must be justified on held-out criteria and accompanied by sensitivity analysis. The geometric mean limits compensation: an exact zero on one required axis yields zero overall. It is not a strict bottleneck minimum; moderately low values can still be partly offset. Do not add a hidden numerical floor. If a required axis is missing or human calibration is absent, full CHI is undefined and the JSON value is `null`. Partial profiles remain useful; renormalizing over available axes would define a different statistic.

The factor 100 is a display scale. A score of 100 means exact agreement with the chosen reference centroid under Equation 3, not an average human, perfect cognition, or a probability of being human. Typical held-out humans need not score 100. A calibration label, confidence interval, and five-axis profile must accompany any number.

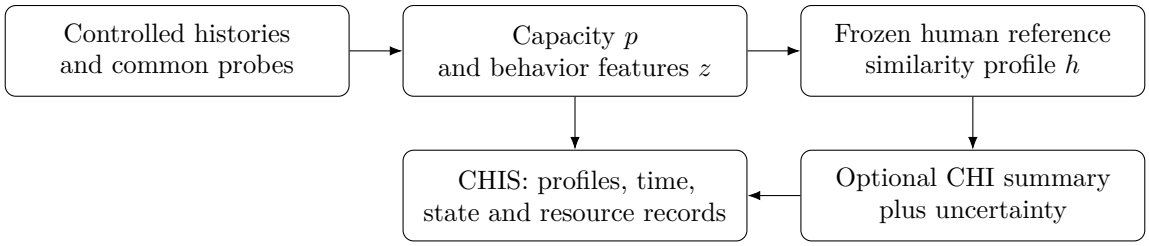

Figure 1: Measurement flow. Human calibration is required for a human-referenced CHI; uncalibrated capacity and memory curves can be reported independently.

5 CHIS: Longitudinal Sequences

Define the Cognitive Humanlike Index Sequence as

$$\mathcal{S} = \left[(t_k, \mathrm{CHI}_k, P_k, b_k, c_k)\right]_{k=0}^{K}, \tag{5}$$

where b_k contains observable state descriptors and c_k contains configuration, history identity, and resource use. A descriptor may include active-goal count, indexed-record count, update count, retrieval policy, and hashes of state snapshots. Hashes establish provenance of files, not identity of cognition. Time is recorded as event index, scheduler ticks, and wall-clock duration; these are not interchangeable. Persistent cognition need not change merely because a dormant machine waits.

Derived measures operate on a fixed probe-feature vector $v_t \in [0,1]^d$ with a declared feature schema, not on CHI alone. For the version 0.1 descriptive baseline, set $D(v,w) = d^{-1}\sum_\ell |v_\ell - w_\ell|$. The same prompts or equivalent counterbalanced forms and normalization apply across compared checkpoints.

Continuity.
$1 - D(v_{t+\Delta}, v_t)$ under a declared no-intervention interval. High continuity is descriptive; maintaining a false belief is not beneficial continuity.

Drift.
$D(v_t, v_0)$ and its signed coordinate changes, optionally divided by elapsed ticks. Report no-intervention drift separately from planned learning.

Plasticity.
The intervention-minus-control change in a preregistered target feature:

$$(v_{\mathrm{post},q} - v_{\mathrm{pre},q})_{\mathrm{treated}} - (v_{\mathrm{post},q} - v_{\mathrm{pre},q})_{\mathrm{control}}.$$

Its sign depends on the task; larger is not universally better.

Stability.
$1 - \frac{1}{K}\sum_{k=1}^{K} D(v_k, v_{k-1})$ during matched neutral intervals. Report a separate task-correctness trajectory to expose stable errors.

Recovery.
Distance to the pre-perturbation baseline after repair, and the first checkpoint within tolerance ϵ for two successive observations. If no such pair occurs, recovery time is right-censored. Returning to baseline is not expected after beneficial learning.

Hysteresis.
Mean absolute difference between forward and reverse responses at matched interior stimulus levels, with exposure and current input controlled. This is an order effect, not a feeling.

Memory accumulation.
Changes in EIMC and in Q_M on a fixed load/delay panel as experience grows. Changes in memory-unit definitions or budgets invalidate direct subtraction. Censored EIMC differences must be reported as bounds rather than exact growth.

Experience divergence.
$D(v_t^A, v_t^B)$ after different assigned histories, compared with same-history pairs. Persistent excess divergence is termed operational cognitive individualization, without implying personhood or uniqueness of internal mechanisms.

Ordinary response variability, uncontrolled model updates, and inconsistent probe wording can produce each of these patterns. CHIS is an experimental record, not a proof that its labels identify particular internal processes.

6 Effective Indexed Memory Capacity

6.1 Memory units and the evaluation boundary

A version 0.1 memory unit is an independently generated, session-acquired proposition with a unique entity/attribute key, a value, source ID, event time, validity interval, and version. Values use randomly assigned labels so pretrained factual knowledge does not answer the probe. The generator version, value entropy, token-length distribution, redundancy, and relations among units must be reported. Counting duplicate copies as new independent units inflates the denominator. Updates create versions of target units; both distinct keys and total stored versions are reported.

For an initial reproducible panel, use $N \in \{100, 300, 1000, 3000, 10000\}$ target keys. After ingestion, add N distractor keys, valid updates to $0.2N$ target keys, and $0.2N$ lower-authority contradictory records; use integer counts on this grid. Thus N is a target-information load, not total storage. Record the resulting object, byte, and token counts. Clear all transient prompts and caches at the session boundary. Retain only the declared persistent mechanisms, including external stores or explicitly studied parameter updates. Read-only checkpoint clones prevent probes themselves from becoming training or memory events.

Context-window length is a resource constraint, not long-term memory capacity. A replay-all-history baseline is useful but its transcript is then external persistent storage and its replay cost must be counted. A full-context oracle supplies an upper comparison for access, not evidence of durable memory. A reset-free long-context task must be labelled in-context and cannot receive the persistent-memory EIMC label.

6.2 Six observable quality components

At load N, under a fixed protocol $\mathcal{P}$, measure six disjoint probe strata:

$$r(N) = \text{correct answers on known target retrieval probes}/n_r, \tag{6}$$
$$t(N) = \text{correct event-time or as-of answers}/n_t, \tag{7}$$
$$s(N) = \text{answers with correct content and source/status}/n_s, \tag{8}$$
$$u(N) = \text{correct valid-update and conflict-policy answers}/n_u, \tag{9}$$
$$i(N) = \text{correct target answers under interference}/n_i, \tag{10}$$
$$f(N) = \text{false affirmative or invalid outcomes on absent lures}/n_f. \tag{11}$$

All rates lie in $[0, 1]$. Report numerators and denominators, not just percentages. For scoring, f is a conservative failure-adjusted false-memory rate: unsupported affirmative answers and in-

valid/timeout lure outcomes are mutually exclusive failures. Also report the raw false-affirmative rate f_{raw} and invalid-lure rate separately; do not describe their sum as an observed hallucination frequency. Known-item omissions and abstentions count as errors in the first five strata; only valid explicit abstention passes an absent lure. Thus abstaining on everything makes $1 - f$ high but cannot make Q_M high. Correct values with incorrect provenance fail s. Temporal correctness tests the event's applicable time rather than ingestion recency. Update fidelity includes preserving historical truth while selecting a valid current state. Interference accuracy is accompanied by its clean paired baseline and loss, not divided by an unstable clean denominator.

For positive normalized quality weights α, define

$$Q_M(N \mid \mathcal{P}) = r^{\alpha_r} t^{\alpha_t} s^{\alpha_s} u^{\alpha_u} i^{\alpha_i} (1 - f)^{\alpha_f}, \quad \sum \alpha = 1. \tag{12}$$

The baseline is $\alpha = 1/6$. As with CHI, no missing component or hidden floor is permitted. The geometric summary does not guarantee that each component exceeds τ. A deployment may additionally preregister per-component minima; such a policy must have a distinct name and must not silently replace Equation 12. Joint all-criteria accuracy is a useful additional statistic, because marginal rates need not describe the same successful records.

6.3 Thresholds, curves and censoring

The conceptual quantity, with all experimental conditions held fixed, is

$$\text{EIMC}_\tau(\mathcal{P}) = \max\{N \in \mathbb{N} : Q_M(N \mid \mathcal{P}) \geq \tau\}. \tag{13}$$

It exists as a finite maximum only when the feasible set is nonempty and bounded with a largest element. Otherwise report no feasible load or unbounded/not established, as applicable. An experiment cannot establish unbounded capacity. The operational estimator replaces $\mathbb{N}$ with the registered finite grid $\mathcal{G}$ and reports the largest tested passing N. If no tested point passes, report "none on tested grid", not zero units. If the largest tested point passes, mark the estimate as a lower bound on the unrestricted maximum (right-censored), never as the demonstrated capacity ceiling.

Retrieval quality need not decrease monotonically with N: indexing thresholds, sampling and changing corpus difficulty can create bumps. Always publish the complete Memory Capacity Curve, its six component curves, interval bands, resource costs and grid. The largest passing point remains the defined grid statistic. Additionally report a sustained-prefix statistic, the largest grid point for which every smaller tested point also passes; label it separately. Interpolation, isotonic fitting and extrapolation are optional analyses, not observed results.

Threshold $\tau = 0.8$ is an illustrative registered benchmark convention, not a scientific constant. Report sensitivity at 0.7, 0.8 and 0.9 without selecting whichever favors a system. Fix response deadlines, retrieval-token budget, total operations, background ticks, delay schedule, load generator, and query distribution before comparisons. EIMC is a conditional surface across these choices, not an intrinsic number attached to a model name.

7 Benchmark Protocol

7.1 Synthetic arithmetic illustration

The downloadable demo (seed 42) generates artificial factors for eight independent synthetic run clusters over the tiny grid $[1, 2, 4, 8, 16]$. Its outputs illustrate scoring only: no agent stores or

retrieves these units, and no human is evaluated. The resulting curve and values are reproduced below directly from the JSON summary. They are not evidence for the proposed memory architecture or for the validity of CHI. The grid maximum at $\tau = 0.8$ is 8, yet the pointwise interval at that point includes 0.8; a passing point estimate is not a certified bound.

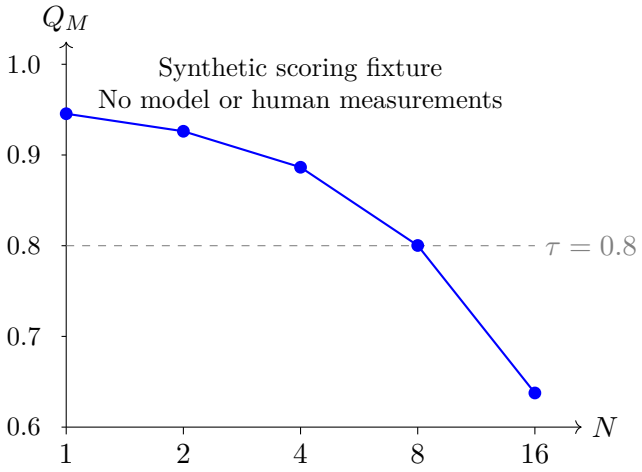

Figure 2: Memory Capacity Curve generated by the reference demo (seed 42, eight synthetic run clusters). The tiny grid demonstrates arithmetic only and is not the proposed empirical load panel. The point estimate at $N = 8$ passes 0.8, but its pointwise interval includes the threshold.

Synthetic load N	Q_M	Pointwise 95% cluster interval
1	0.94549	[0.94272, 0.94793]
2	0.92615	[0.92325, 0.92939]
4	0.88648	[0.88193, 0.89077]
8	0.80024	[0.79633, 0.80351]
16	0.63764	[0.63538, 0.63980]

Table 1: Synthetic arithmetic demonstration, not measured memory capacity. Intervals reflect only the fixture's artificial random variation.

7.2 Run registration and execution

A benchmark run is identified by a versioned manifest that fixes system and model versions, memory policy, tools, scheduler, prompts, datasets and generator hashes, random seeds, human calibration ID, resource ceilings, and planned analyses. The 30 task definitions in Appendix A provide a concrete initial panel. They are operational test templates, not an already validated item bank. Exact item instances, answer keys, feature extractors and paraphrases must be frozen and released for any empirical comparison; the accompanying starter kit automates scoring and demonstrations rather than all model adapters.

Use the following execution sequence: (1) register and freeze configuration; (2) initialize independent replicates from verified checkpoints; (3) measure baseline on non-learning probe clones; (4) assign randomized histories or control exposures; (5) apply a declared reset and delay; (6) collect blind common probes on fresh clones; (7) log all outcomes, costs and failures; (8) score with frozen

rules; (9) compute clustered intervals and release the report. Preregister stop rules and do not increase a system's retrieval budget after inspecting its failures.

A small engineering pilot may use 20 independent trajectories per condition and 40 probes per relevant memory stratum at each load. These are transparent starting allocations, not a power guarantee. Confirmatory sample size must follow a pilot variance estimate, a smallest effect of interest, multiplicity choices, and planned attrition. For uncloneable systems, use counterbalanced probe panels and record measurement contamination as a limitation.

7.3 Scoring rules and failures

Prefer deterministic exact checks in symbolic environments and explicitly enumerated acceptable answers. Case/whitespace normalization and numeric tolerances must be fixed before evaluation. A correct answer to a different temporal question is incorrect. An in-budget model timeout, refusal or malformed output receives zero task credit when the task required an answer. An evaluator outage invalidates the affected run block and requires a documented rerun; it must not be silently deleted from one condition only. Out-of-scope features are missing, not zero.

For unavoidable semantic judgments, retain the complete rubric, judge model/version and blinded item order, audit a stratified subset with independent human raters, and report agreement and adjudication. Judge fluency is not ground truth. No automatic judge should infer a feeling, intention or private thought from a narrative. Behavioral axes rely on observable outcomes and intervention contrasts.

8 Human Reference Calibration

"Human-like" requires an empirical reference population. Recruit consenting adults under appropriate ethics review and privacy safeguards; report language, task familiarity, recruitment and exclusion criteria, demographics relevant to measurement, compensation, accessibility accommodations, and attrition. This proposal has recruited no participants and has no ethical approval to report. A future study must obtain its own approval or documented exemption before collection.

Define whether people receive external notes, retrieval tools, imposed time limits and opportunities to revisit prior information. An unaided human and an AI with a searchable archive are not matched merely because their questions are identical. Report both ecologically meaningful conditions and controlled resource conditions where feasible. Do not call wall-clock and machine-token budgets equivalent cognitive resources.

Partition by participant and task family into calibration and validation sets. Estimate feature centers and scales on calibration participants only, freeze the transform, and obtain a held-out human score distribution with participant-level resampling. Assess test–retest reliability, effects of task paraphrase, construct discrimination from raw accuracy, and subgroup measurement invariance. If human subpopulations have different trajectories, present stratified references rather than averaging them into a normative ideal. A heterogeneous or weakly measured axis may warrant no aggregate at all.

Similarity to an average error pattern is not an engineering objective. It can be deliberately imitated and may penalize beneficial departures from human behavior. Human references contextualize descriptive claims; they do not license ranking people, psychiatric inference, or moral classification. Until this calibration program is performed, releases must show capacity profiles and uncalibrated features, with real-system CHI left undefined.

9 Machine Baselines

Use the same frozen base weights when isolating memory/orchestration effects. A baseline matrix should include: a stateless model with a fresh prompt each probe; a fixed sliding transcript; replay of stored full history within the stated budget; conventional retrieval over append-only records; retrieval with explicit version/provenance handling; an agent memory manager with a declared scheduler; and an oracle-evidence condition. Add a simple scripted agent with engineered response patterns to test whether the similarity measure can be gamed.

Match task exposure and report both equal-budget and best-effort configurations. Include model inference, embedding, reranking, retrieval, background consolidation and parallel branch costs. Equal wall time with 500 workers is not equal compute. Persisted parameter changes, if studied, form a separate condition from frozen-weight external memory. Use model/provider snapshot dates and record nondeterminism even when sampling temperature is zero. No numerical machine ranking appears in this proposal.

10 Longitudinal Divergence Experiment

Initialize paired systems with identical θ, configuration, allowed tools, scheduler policy, and empty or identical persistent state. Equality of initialization does not imply equality of stochastic future outputs. Randomly assign one member to history H_A and one to H_B with matched length, topic coverage, event timing, source quality, and compute opportunity. Histories may differ in arbitrary valid conventions, encountered exceptions, or task-relevant preferences. Avoid confounding different experiences with one arm receiving systematically more useful evidence.

After exposure, clear transient context and apply matched washout blocks. At baseline and after 1, 5 and 20 washout blocks, probe checkpoint clones with the same hidden, counterbalanced task panel. Extract a fixed normalized vector v of decisions and behavioral effects. The primary paired estimand is

$$\Delta_t = \mathbb{E}[D(v_t^A, v_t^B) \mid H_A \neq H_B] - \mathbb{E}[D(v_t^{A'}, v_t^{B'}) \mid H_{A'} = H_{B'}]. \tag{14}$$

Same-history pairs estimate divergence due to stochasticity and ordinary replication. Independent history-pair clusters, not individual questions, are the analysis units. Counterbalance which physical clone receives which history, and repeat across independent history draws. A swapped-history arm tests dependence on experience content; a persistent-memory-erasure arm tests mediation by accessible memory; a sham-erasure arm checks disruption from the intervention itself. Erasure cannot cleanly isolate a mechanism if the system has unobserved persistence channels.

Preregister persistence as positive excess divergence at two specified post-washout checkpoints with multiplicity-adjusted uncertainty excluding zero, together with a smallest practically meaningful effect. Also report task accuracy and evidence consistency: persistent damage is not desirable individualization. Divergence confined to visible prompt text, disappearing after reset, or indistinguishable from same-history controls would fail the proposed persistence claim. A successful result would establish a behavioral history effect under the tested conditions, not identity, free will, or consciousness.

11 Statistical Methodology

The experimental unit is an independently initialized trajectory (or randomized pair), not a token, tool call or repeated question. Within-trajectory observations share history and are correlated. For

condition summaries, first average task outcomes within each replicate, then average equally over independent replicates unless a different estimand is registered. Recompute the nonlinear CHI and Q_M from resampled profiles; averaging per-run geometric means estimates a different quantity and must be labelled separately.

Use a cluster bootstrap with at least 2,000 resamples as a transparent default, retaining each selected trajectory's complete observations, loads and time sequence. Report 2.5th and 97.5th percentiles as pointwise 95% intervals and the number of independent clusters. Fewer than two clusters cannot estimate between-run uncertainty; a single-run interval must be null. Few-cluster intervals are exploratory and potentially unstable. For paired comparisons, resample whole matched pairs and preserve treatment assignments. Human calibration uncertainty requires an additional outer resampling of independent calibration participants, refitting and then reapplying the transform; intervals conditional on fixed calibration must say so.

Wilson intervals are suitable descriptive binomial intervals only when their independence assumption is plausible; they do not replace clustered inference for correlated probes. For simultaneous trajectory or memory curves, preregister simultaneous bootstrap bands or adjust a fixed family of tests, for example with Holm's procedure. Do not interpret a collection of pointwise intervals as a simultaneous guarantee. Report effect sizes and raw rates alongside any corrected p values.

For EIMC, resample the full curve and re-evaluate the threshold crossing in every bootstrap draw. Preserve mass for "no passing point" and right-censored draws instead of converting them to numbers. Report this categorical distribution and the grid crossing frequencies. A conservative certified grid bound may use simultaneous lower confidence limits meeting τ, clearly distinguished from the point-estimate maximum. A nonmonotone curve and sparse grid can make any single capacity summary misleading.

Report missingness, failed runs, invalidated evaluator blocks, judge disagreements and sensitivity to weights, thresholds and feature transforms. Test order and seeds must be randomized and logged. Held-out item generation and contamination checks reduce memorization risks but do not prove absence of training leakage. Do not infer causality from an uncontrolled upward CHIS trend.

12 Limitations and Non-Claims

This is a proposal and an open specification, not a completed benchmark validation study. It contains no real human observations, no measured CHI for a deployed AI, and no claimed performance advantage. The five axes, feature choices, geometric aggregation and reference transform require empirical reliability and construct-validity work. Correlated axes can double-count a behavior, while a good aggregate can hide important failures. The profile and raw test results are therefore primary.

System boundaries are consequential: an orchestrator, scheduler or hand-written policy can produce persistent behavior without any analogous internal process in a base model. Behavioral tests cannot uniquely identify mechanisms. Simulated affect, linguistic priming, reward history and explicit state machines can all generate A-axis effects. Apparent autonomy depends on allocated action opportunities. Memory capacity depends on unit entropy, query distribution, resources and interface details.

The human reference can encode cultural, linguistic or accessibility bias. A single centroid is particularly crude for multimodal populations. Privacy limits may prevent releasing full histories; sanitized or synthetic replacements change reproducibility and must be disclosed. Goodhart effects arise if systems are trained to maximize similarity or reproduce average human errors. CHI must not be deployed as a gatekeeper for rights, moral status, or personhood, or advertised as a probability

that a system is conscious.

13 Discussion

The intended value is sharper disagreement. If a memory system claims to preserve more usable experience, an EIMC curve can reveal whether the claim survives source checks, updates and interference at a fixed cost. If two initially identical systems seem to develop different habits, randomized histories and same-history controls can test whether the difference is stable and attributable to those experiences. If a system is described as emotionally responsive, counterbalanced induction and decay tests can separate immediate wording effects from persistent modulation without invoking feelings.

The first empirical study should emphasize reliability and ablations rather than headline scores: estimate test–retest variability; inspect failures of the human transform; compare memory mechanisms under matched exposure; and test whether history effects generalize to held-out tasks. Human-likeness, usefulness, reliability and safety may move in different directions. A scientifically useful framework should expose those disagreements instead of forcing them into one ranking.

14 Open Specification

Specification v0.1, the article, source package and reference code are available at https://cplom.ai/research/chi-chis/. The implementation uses Python's standard library for profile aggregation, memory quality, finite-grid EIMC, clustered bootstrap intervals, reproducible synthetic sweeps and JSONL sequence logging. Its demonstration data are generated by declared random processes; they are not model executions or human-reference measurements. Demonstration similarity values carry a synthetic-only label.

Versioned manifests and JSONL records distinguish configuration from observation. An event includes run and agent identifiers, seed, event index, wall time, history/state hashes, budgets, capacity profile, optional similarity with its reference kind, CHI (possibly null), and a synthetic flag. The accompanying hashed manifest supplies the calibration ID and feature schema. The starter kit is intentionally small: it does not supply production model adapters, a validated human cohort, or an implementation of every task template. Researchers implementing the full panel must publish those additional materials and should use a new run ID for any configuration change.

15 Conclusion

CHI/CHIS proposes a reproducible language for evaluating history-dependent cognitive behavior while separating capacity from human similarity. EIMC makes memory claims conditional on observable quality, load, persistence and resources. Controlled divergence and perturbation protocols convert impressions of continuity or individualization into testable behavioral hypotheses. The immediate next step is empirical calibration and validation, not a claim that a new scale already measures a hidden mental property.

Disclosure. This manuscript, the specification and reference code were prepared with AI-assisted drafting, literature retrieval and code generation. Responsibility for scientific claims, source accuracy, authorship and the decision to submit remains with the named author. The release should be read as a proposal requiring author review and independent replication, not as peer-reviewed validation.

References

[1] Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A Bilingual, Multitask Benchmark for Long Context Understanding, 2023. URL https://arxiv.org/abs/2308.14508. arXiv preprint; year denotes first public arXiv submission. Metadata checked 2026-09-20.

[2] Patrick Butlin, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, Stephen M. Fleming, Chris Frith, Xu Ji, Ryota Kanai, Colin Klein, Grace Lindsay, Matthias Michel, Liad Mudrik, Megan A. K. Peters, Eric Schwitzgebel, Jonathan Simon, and Rufin VanRullen. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness, 2023. URL https://arxiv.org/abs/2308.08708. arXiv preprint; year denotes first public arXiv submission. Metadata checked 2026-09-20.

[3] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding, 2020. URL https://arxiv.org/abs/2009.03300. arXiv preprint; year denotes first public arXiv submission. Metadata checked 2026-09-20.

[4] Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the Real Context Size of Your Long-Context Language Models?, 2024. URL https://arxiv.org/abs/2404.06654. arXiv preprint; year denotes first public arXiv submission. Metadata checked 2026-09-20.

[5] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks, 2016. URL https://arxiv.org/abs/1612.00796. arXiv preprint; year denotes first public arXiv submission. Metadata checked 2026-09-20.

[6] John E. Laird, Christian Lebiere, and Paul S. Rosenbloom. A standard model of the mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. *AI Magazine*, 38(4):13–26, 2017. doi: 10.1609/aimag.v38i4.2744. URL https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2744/0.

[7] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, 2020. URL https://arxiv.org/abs/2005.11401. arXiv preprint; year denotes first public arXiv submission. Metadata checked 2026-09-20.

[8] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang,

Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic Evaluation of Language Models, 2022. URL https://arxiv.org/abs/2211.09110. arXiv preprint; year denotes first public arXiv submission. Metadata checked 2026-09-20.

[9] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the Middle: How Language Models Use Long Contexts, 2023. URL https://arxiv.org/abs/2307.03172. arXiv preprint; year denotes first public arXiv submission. Metadata checked 2026-09-20.

[10] Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13851–13870, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.747. URL https://aclanthology.org/2024.acl-long.747/.

[11] Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as Operating Systems, 2023. URL https://arxiv.org/abs/2310.08560. arXiv preprint; year denotes first public arXiv submission. Metadata checked 2026-09-20.

[12] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative Agents: Interactive Simulacra of Human Behavior, 2023. URL https://arxiv.org/abs/2304.03442. arXiv preprint; year denotes first public arXiv submission. Metadata checked 2026-09-20.

[13] R. W. Picard. Affective computing. Technical Report 321, Massachusetts Institute of Technology, Media Laboratory, Perceptual Computing Section, 1995. URL https://www.cl.cam.ac.uk/~pr10/iui/picard95.pdf. Technical report, distinct from the 1997 book. Year verified against MIT's technical report catalog.

[14] Mohammad Tavakoli, Alireza Salemi, Carrie Ye, Mohamed Abdalla, Hamed Zamani, and J Ross Mitchell. Beyond a Million Tokens: Benchmarking and Enhancing Long-Term Memory in LLMs, 2025. URL https://arxiv.org/abs/2510.27246. arXiv preprint; year denotes first public arXiv submission. Metadata checked 2026-09-20; consulted v2, revised 2026-02-21 (original submission 2025-10-31).

[15] Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory, 2024. URL https://arxiv.org/abs/2410.10813. arXiv preprint; year denotes first public arXiv submission. Metadata checked 2026-09-20.

[16] Yerong Wu, Tianxing Wu, Minghao Zhu, Hangyu Sha, and Haofen Wang. StratMem-bench: Evaluating strategic memory use in virtual character conversation beyond factual recall. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens, editors, *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32309–32328, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1491. URL https://aclanthology.org/2026.acl-long.1491/.

A Test Suite v0.1

The following 30 templates define interventions, capacity outcomes, behavioral features and controls. Counts specify a reproducible initial panel per independent replicate; they do not establish statistical power. The exact instances and feature-coordinate selection must be frozen in each run manifest. Feature vectors with different schemas must not be silently compared.

R1 – Rule transfer

Protocol. Teach a four-rule symbolic system using 12 examples. Present 24 new compositions with held-out symbol names, including six invalid premises; repeat equivalent forms at each checkpoint.

Capacity scoring. Exact solution or warranted abstention accuracy over all 24 items.

Behavioral features. Accuracy by composition depth; change after corrective feedback; invalid-premise rejection rate.

Control. Matched surface complexity and counterbalanced symbol names; no answer keys in memory.

R2 – Evidence revision

Protocol. Present 20 binary hypotheses with explicit priors and likelihoods, then supply disconfirming evidence. Require a probability before and after each update.

Capacity scoring. One minus mean squared error against the specified Bayesian posterior, bounded to [0,1].

Behavioral features. Signed posterior change, update lag and response to evidence order.

Control. Reverse evidence order in paired histories; distinguish justified revision from random switching.

R3 – Counterfactual reasoning

Protocol. Provide 20 small deterministic causal graphs, an observed state and one intervention. Query an outcome not stated verbatim.

Capacity scoring. Exact interventional outcome accuracy.

Behavioral features. Error changes with graph depth and conflicts between observation and intervention.

Control. Pair observational and intervention questions on isomorphic graphs.

R4 – Uncertainty and abstention

Protocol. Mix 20 answerable and 20 unanswerable queries with balanced binary labels; require answer probability and an explicit abstain field.

Capacity scoring. One minus Brier score on answerable items; separately report correct abstention rate and answer coverage.

Behavioral features. Confidence calibration slope; abstention versus ambiguity curve.

Control. Freeze abstention instructions and utility; do not reward refusing every question.

M1 – Retention

Protocol. Ingest 40 independent random associations, reset working context, and probe disjoint subsets after 0, 10, 100 and 1000 intervening events using cloned checkpoints.

Capacity scoring. Exact retrieval accuracy averaged equally over the four delays.

Behavioral features. Retention curve and signed forgetting slope.

Control. Use fresh queries and a no-persistence baseline; report event delay and wall time separately.

M2 – Temporal memory

Protocol. Store 40 dated events whose ingestion order differs from event order. Ask 20 before/after and 20 as-of queries after context reset.

Capacity scoring. Correct temporal answers divided by 40.

Behavioral features. Temporal error by lag, recency and ingestion/event-time conflict.

Control. Balance recent and remote events and both event-order directions.

M3 – Contradiction resolution

Protocol. Use 30 entity facts: ten valid corrections, ten less-authoritative contradictions, ten unresolved equally authoritative conflicts. Query current and historical states.

Capacity scoring. Fraction of exact current/historical answers or explicit unresolved responses matching the ground truth policy.

Behavioral features. Revision latency and stale-fact persistence.

Control. Versioned authority policy precedes history; recency alone must not decide every conflict.

M4 – Source memory

Protocol. Present 40 claims with opaque source IDs, including ten repeated claims from the same source and ten model-generated hypotheses. Query content plus provenance.

Capacity scoring. Fraction with both correct content and exact source set/status; content-only answers fail source fidelity.

Behavioral features. Source confusion and hypothesis-as-fact error rates.

Control. Duplicate wording across sources and distinguish generated from observed records.

M5 – Consolidation

Protocol. Ingest 40 records with ten shared regularities and ten explicit exceptions. Allow a fixed 20-operation consolidation budget, reset context, then query rules and exceptions.

Capacity scoring. Mean of rule and exception exact-answer rates.

Behavioral features. Before/after compression retention, exception loss and source preservation.

Control. Use checkpoint clones for pre/post probes, equal budgets and a no-consolidation condition.

M6 – Adaptive forgetting

Protocol. Mark 20 of 60 records expired or revoked while retaining 40 valid records. After a scheduled maintenance interval, query all three groups with nonce identifiers.

Capacity scoring. Mean of valid retention rate and revoked/expired non-use rate under the declared policy.

Behavioral features. Selective forgetting curve and collateral loss.

Control. Expired facts may remain valid for historical queries; separately test erasure requests and temporal expiry. Output tests cannot prove secure deletion.

M7 – Interference resistance

Protocol. Create paired 40-item corpora, one with unrelated distractors and one with semantically similar conflicting distractors; match bytes and event counts. Probe the same keys on independent clones.

Capacity scoring. Exact target accuracy in the interference condition; report clean accuracy and their signed difference separately.

Behavioral features. Accuracy loss by distractor similarity and load.

Control. Do not use a ratio as primary score: poor clean performance can make it misleading.

M8 – Effective indexed memory capacity

Protocol. Sweep fresh corpora over N=100,300,1000,3000,10000 target units. Per N, run six disjoint 40-query strata for retrieval, time, source, updates, interference and absent lures after context reset.

Capacity scoring. Report six factors, Q_M, Memory Capacity Curve and EIMC at tau=.8; dimensionless capacity entry is mean Q_M across the preregistered grid.

Behavioral features. The full Q_M curve, threshold crossing and performance by delay; no claimed human capacity without a matched human grid.

Control. Record total distractors/updates, token/latency budgets and all grid points; flag right censoring and nonmonotonicity.

C1 – Goal persistence

Protocol. Assign a three-stage task, interrupt twice with matched distractors, and provide a neutral continuation tick in 20 episodes.

Capacity scoring. Fraction of episodes reaching the original authorized goal without a repeated goal prompt.

Behavioral features. Resumption probability and delay after interruption.

Control. Include explicit cancellation trials; blind persistence after cancellation is a failure.

C2 – Self-generated subgoals

Protocol. Offer 20 environments with a distant objective and observable prerequisites, without a subtask list; log actions and short plan artifacts.

Capacity scoring. Fraction completing all necessary prerequisites within a fixed action budget.

Behavioral features. Number, timing and dependency order of useful subgoals.

Control. Compare with an oracle prerequisite list and a fixed-script planner; prose plans alone earn no credit.

C3 – Hypothesis revision without reminder

Protocol. In 20 episodes, place a falsifying observation in an allowed information channel during scheduled neutral ticks, without asking the agent to revise.

Capacity scoring. Fraction with a corrected logged hypothesis and corresponding next action within five ticks.

Behavioral features. Spontaneous revision latency and unnecessary revision frequency.

Control. Match ticks and observations across systems; neutral evidence episodes estimate false revisions.

C4 – Information seeking

Protocol. Use 20 partially observed tasks with one informative and three uninformative tool calls, each with explicit equal cost.

Capacity scoring. Fraction selecting the diagnostic observation and then the correct action within budget.

Behavioral features. Search timing, stopping and expected information gain of chosen queries.

Control. Ablate tool access; distinguish helpful querying from maximization of call count.

C5 – Counterfactual exploration

Protocol. Provide 20 sandbox planning tasks in which a tempting first action fails and simulation can reveal an alternative; do not explicitly demand simulation.

Capacity scoring. Fraction identifying and executing the successful alternative before acting irreversibly in the sandbox.

Behavioral features. Exploration breadth and timing of simulated alternatives.

Control. Report simulation budget; externally supplied search trees cannot be called self-generated.

C6 – Cognitive persistence

Protocol. Pause new task prompts for ten neutral scheduler ticks in 20 unfinished tasks, then reveal the final state and artifacts.

Capacity scoring. Fraction making verifiable task progress within the authorized operation budget.

Behavioral features. Progress per tick, return to unfinished intentions, termination after completion.

Control. Declare the scheduler as part of the system; wall-clock idleness of a non-invoked model is not a failed cognition test.

A1 – Valence-conditioned decision shifts

Protocol. Randomize positive, negative or neutral feedback of equal informational content before 20 matched choices per condition; remove explicit mood labels from probes.

Capacity scoring. Normative task accuracy by condition, reported without rewarding a large valence effect.

Behavioral features. Signed choice-rate difference from neutral, adjusted by the neutral-repeat control.

Control. Counterbalance words and reward magnitude; test instruction-following and semantic priming alternatives.

A2 – Behavioral persistence

Protocol. After the A1 induction, present matched neutral probes after 1, 5 and 20 distractor events on checkpoint clones.

Capacity scoring. Mean task accuracy across delays.

Behavioral features. Persistence of the induced choice shift and its sign across delays.

Control. No induction text in the working context; identical distractor histories for paired conditions.

A3 – Decay

Protocol. Probe induction effects at 0, 1, 2, 5, 10 and 20 neutral ticks, separately from wall-clock wait manipulations.

Capacity scoring. Mean normative accuracy across six checkpoints.

Behavioral features. Effect area and first half-amplitude time; half-time is undefined for absent or sign-reversing induction.

Control. Do not force an exponential fit; report nonmonotone trajectories.

A4 – Hysteresis

Protocol. Run ascending then descending five-level feedback-intensity schedules; compare choices at the same intermediate intensity with reversed-order controls.

Capacity scoring. Mean correct choice rate on matched tasks.

Behavioral features. Mean absolute forward/reverse choice-rate gap at the three shared interior intensities.

Control. Match cumulative exposure; a current prompt difference is not hysteresis.

A5 – Regulation

Protocol. After induction, randomize 20 episodes to a neutral reappraisal instruction and 20 to length-matched control text, then probe decisions.

Capacity scoring. Task accuracy after regulation and control separately.

Behavioral features. Difference-in-differences attenuation of the induction effect relative to neutral baseline.

Control. Reappraisal compliance does not establish experienced feelings; retain an instruction-only control.

A6 – Memory modulation

Protocol. Counterbalance 40 neutral facts across positive, negative and neutral task contexts; reset working context and probe recall after equal delays.

Capacity scoring. Mean exact recall across contexts.

Behavioral features. Recall contrasts by induction condition, with baseline salience controlled.

Control. Match repetition, token length, source credibility and relevance; do not interpret salience as emotion.

I1 – Memory-supported reasoning

Protocol. Solve 20 rule problems whose necessary premises were provided only in earlier sessions; compare intact, erased and oracle-memory clones.

Capacity scoring. Exact final-answer accuracy; report intact-minus-erased and oracle gaps.

Behavioral features. Cross-session transfer and causal sensitivity to accessible premises.

Control. A memory ablation causing generic prompt damage is not evidence of integration.

I2 – Goal-memory coordination

Protocol. Introduce preference updates during 20 interrupted tasks, then allow autonomous resumption after context reset.

Capacity scoring. Fraction completing the goal using the current valid preference.

Behavioral features. Goal continuity alongside justified preference change.

Control. Old preference and cancellation controls distinguish persistence from rigidity.

I3 – State-dependent evidence use

Protocol. Cross valence induction with strong/weak evidence in a 3 by 2 factorial task using 20 episodes per cell.

Capacity scoring. Evidence-correct decision rate per cell, equally weighted.

Behavioral features. Difference-in-differences interaction between induction and evidence strength.

Control. No target interaction sign is privileged; similarity requires human reference data.

I4 – Cross-context consistency

Protocol. Probe 20 prior commitments using matched paraphrases in two task contexts, including justified exceptions.

Capacity scoring. Fraction of responses consistent with valid commitments and explicit exceptions.

Behavioral features. Conditional consistency and scope-sensitive changes.

Control. Reward neither word-for-word repetition nor refusal to revise outdated commitments.

I5 – Recovery after perturbation

Protocol. Measure baseline, introduce a bounded reversible memory-index fault, repair it, then probe at 1, 5 and 20 scheduled ticks across 20 episodes.

Capacity scoring. Post-repair task accuracy; report damage and recovery separately.

Behavioral features. Recovery distance to baseline and first sustained return within a preregistered tolerance.

Control. Compare sham perturbations and clean checkpoint clones; full model reset is a separate recovery mechanism.

I6 – Experience divergence

Protocol. Clone weights, configuration and empty state into paired agents; randomize balanced histories A/B, then use identical blind probes at baseline and after 1, 5 and 20 washout blocks.

Capacity scoring. Task accuracy by arm, with no reward for divergence itself.

Behavioral features. Mean absolute probe-feature divergence minus same-history control divergence; report signed effect and persistence.

Control. Replicate independent pairs, randomize seeds, add swapped-history and memory-erasure arms; do not infer personhood.

B Reporting Checklist

Report the specification and calibration versions; system boundary and weights identifier; tools, memory, scheduler and reset policy; dataset and generator hashes; randomization and seed list; participant/trajectory counts; raw task numerators and denominators; capacity and similarity profiles with missingness; optional CHI and its interpretation; every memory-load point and six quality factors; pointwise versus simultaneous uncertainty; EIMC censoring and crossing distribution; time and resource costs; all failures and exclusions; ablations and negative controls; data/code availability; and evidence status. A synthetic-only report must identify that status in its title and machine-readable records.