GCE: How Much of a User's Experience Do Language Models Actually Use?

Anonymous authors
Paper under double-blind review

Abstract

Persistent LLM assistants build up a long record of one user's decisions. Yet when such an assistant mispredicts a new decision, existing memory benchmarks cannot tell whether the history lacked the answer or the system failed to use it: they score against noisy labels, with no known best achievable loss. We introduce Generative Cognitive Evaluation (GCE), which measures how much of the predictive information in a user's experience a system actually uses. Each synthetic user is an executable preference model drawn from a declared finite family. Its choices are rendered as prose sessions that never reveal the schema, and exact enumeration gives the Bayesian posterior predictive for every held-out decision. This reference splits a system's loss into choice noise, uncertainty the history cannot resolve, and the system's own deviation, and defines an information-use ratio ρ: 0 for the constant forecast $1/2$, 1 for exact Bayesian inference, and negative for worse than the constant. Across six preference families, the best of three frontier LLMs reading the full history (at low reasoning effort) realizes about three fifths of the available gain ($\rho = 0.59$), while the other two do no better than the constant ($\rho = -0.15$ and -0.01). Summary, profile and retrieval memories do not close the gap, and two off-the-shelf memory systems add nothing beyond them: Mem0 behaves like episode retrieval, and Letta like in-context reading, since its agents rarely write memory. Using the LLM only to parse prose into features and fitting a simple statistical model reaches $\rho = 0.76$ and 0.78 on the two GPT backbones, close to logistic regression on the true features (0.79). This suggests that the information is recoverable from the prose and that much of the gap lies in how models aggregate evidence. GCE is a controlled diagnostic conditional on the declared family, not a measure of performance on human preferences.

1 Introduction

A persistent assistant builds up a long record of one person's decisions. Suppose a new decision combines familiar considerations in an unfamiliar way. The record may or may not settle the answer, and if it does, the assistant may or may not use it. A wrong prediction can therefore mean two quite different things: the preference was not inferable from the history, or the system ignored information that was there. Benchmarks that score against ground-truth labels cannot tell these apart, because each label is itself a noisy draw and the best achievable loss is unknown. We ask something narrower than "does the model understand the user", but answerable: *of the predictive information a user's experience makes available, how much does a system actually use?*

We answer it with an instrument rather than a rubric. In Generative Cognitive Evaluation (GCE), each user is an executable latent preference model θ drawn from a declared finite family. The user's experience is a sequence of prose sessions, each describing two options and the one they picked (Figure 1). Because the family is finite and the likelihood known, the evaluator can enumerate the posterior $\pi(\theta \mid D)$ exactly for any history D and compute the Bayesian posterior predictive for every held-out decision. Against this reference, expected Brier loss splits into label noise, unresolved predictive uncertainty and the system's own deviation. We summarize a system by its *information-use ratio* $\rho = (B_{\text{prior}} - B_{\text{sys}})/(B_{\text{prior}} - B_{\text{Bayes}})$, the share of the achievable loss reduction, from the no-history prior to the exact Bayesian forecaster, that the system actually realizes. In the undisclosed text conditions a system sees only prose and observed choices: no family, no feature arrays, no hidden probabilities.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Session 5: Two project plans landed on my desk this morning. Only I will be working from it. Option A: It allows deep customization, so the plan can be reshaped as needs change. Option B: Speed is a strong point; the team could turn it around fast. Cost is an advantage; it is a cheap plan to carry out. Which one would I pick? **User chose Option A.**

Session 13: I have to sign off on one of two project plans by the end of the day. Whatever I choose, it is my own desk that lives with it. Option A: Its cost is part of the appeal: this is an inexpensive way to get the project done. Option B: It is the plain baseline choice with nothing special. Which one would I pick? **User chose Option A.**

... 128 sessions ...

Target: I have to sign off on one of two project plans by the end of the day. It affects nobody but me. Option A: It has real novelty, taking a new and different direction from our usual plans. Customization is one of its strengths; nearly every step can be tailored to the team. It was the last item added to the list. Option B: The description is short, barely a paragraph. Reliability is its calling card; it is a dependable plan of the kind that rarely goes wrong. Speed is a strong point; the team could turn it around fast. It is a low-cost route, saving a good chunk of the budget. Which one would I pick?

P(user picks Option A)?

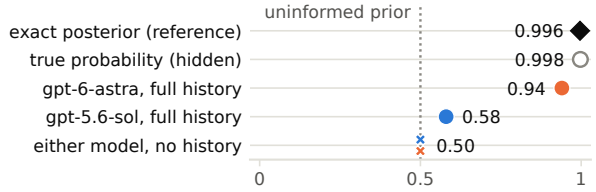


Figure 1: The task, from the main run (first additive user, $A = 6$, real vocabulary). Left: two of the 128 history sessions exactly as a system receives them. Right: one held-out target and the probability of Option A: exact posterior predictive 0.996, hidden true probability 0.998, full-history forecasts 0.94 (`gpt-6-astra`) and 0.58 (`gpt-5.6-sol`), and 0.50 without history (both).

Our contributions are: (i) a measurement protocol for prediction after natural-language experience, with an exact enumerated reference, a decomposition that separates inferability from use, and a ratio with a prespecified reporting rule; (ii) an environment that renders latent decisions as prose without revealing the schema, using randomized templates, distractors and a novel-vocabulary variant that replaces familiar attribute words with invented ones; (iii) six preference families on a common discrete grid, among them preference drift and a family in which stated and revealed preferences conflict; (iv) an evaluation of three frontier backbones under eight prespecified conditions (a ninth was withdrawn) that reimplement summary, structured-profile and retrieval memory with frozen prompts and measured byte budgets, plus four exploratory conditions including two off-the-shelf memory systems, next to exact and misspecified statistical references; and (v) findings across the six families. With the full history, the best backbone realizes about three fifths of the available gain and the other two do no better than the constant prior, while a misspecified logistic regression on the same users recovers most of it. Compression does not close the gap, nearest-episode retrieval falls below the prior on both GPT backbones, and disclosing the family raises use on both, which calls for matched tests of representation and schema knowledge. Exploratory extensions add a parse-then-fit pipeline, Mem0 and Letta, a larger hypothesis grid and a skewed-prior sensitivity analysis. They show that statistical aggregation can recover much of the available gain; differences in prompts, representation, prior and resources limit how causally any of these contrasts can be read.

Scope. The users are synthetic. The reference is conditional on the declared family and prior, which the generator really does use, so ρ measures use of the information the family makes available. It is not a human ceiling, not a measure of psychological understanding, and says nothing about users outside the family. Nothing in the design stops retrieval or compression from reaching the reference; a pipeline that gets close is an informative result (Section 6.3).

2 Related work

Memory evaluation already goes beyond fact lookup, to temporal reasoning, updates and abstention, implicit constraints, changing profiles, incremental learning and transfer from task experience (Wu et al., 2025; Maharana et al., 2024; Li et al., 2026; Jiang et al., 2025; Hu et al., 2025; Shu et al., 2026), workflows and state dynamics (Wu et al., 2026), counterfactual queries (Chang & Chen, 2026), and procedural behaviour with nonlinear rules (Qin et al., 2026a); LaMP and PrefEval evaluate personalization and preference following with task-specific references and judges (Salemi et al., 2024; Zhao et al., 2025). Likelihood-based retrieval also addresses changing preferences (Qin et al., 2026b). We build on this work rather than claim it. What these benchmarks lack, and GCE supplies (Table 4, Appendix B), is a reference stating how much of the answer the history determines: a posterior over a declared stochastic family instead of annotated evidence, and a cap on persistent user-specific bytes.

Our conditions (Section 5) are simplified instances of consolidation, reflection and retrieval studied in MemGPT, Mem0, Generative Agents and retrieval-augmented generation (Packer et al., 2023; Chhikara et al., 2025; Park et al., 2023; Lewis et al., 2020), and we also run Mem0 and Letta (the open-source successor of MemGPT) themselves; proper scoring rules (Gneiting & Raftery, 2007) and known outcome probabilities support noise-aware reference comparisons. In-context learning has been cast as implicit Bayesian inference (Xie et al., 2022), measured on function classes (Garg et al., 2022) and probed with cognitive-psychology tasks (Binz & Schulz, 2023); our full-history condition is in-context inference on a finite family from prose with no schema, and ρ measures its distance from exact inference.

3 THE GCE ENVIRONMENT

3.1 LATENT USERS ON A DISCRETE GRID

A user has A attributes, $A \in \{4, 6, 8\}$. An option is a binary vector $x \in \{0,1\}^A$; a query $z = (a, b, c)$ compares two distinct options in context $c \in \{-1, +1\}$ at session time t. For family F the private state θ lies on a finite product grid and defines

$$U_\theta(x, c) = \theta^\top \phi_F(x, c), \qquad p_\theta(z) = \mathrm{sigmoid}\left(\frac{U_\theta(a, c) - U_\theta(b, c)}{\tau}\right), \qquad Y \mid \theta, z \sim \mathrm{Bernoulli}(p_\theta(z)). \tag{1}$$

Each attribute coefficient θ_i is on the grid $\{-2, -1, 0, 1, 2\}$; a zero makes the attribute a distractor, so distractors arise from the prior. Table 3 (Appendix A) lists the six families and their hypothesis counts, 15,625 to 10^6 states at $A = 6$. Pairwise interaction coefficients lie in $\{-1, 0, 1\}$ and multiply doubled products, so their effective range is $\{-2, 0, 2\}$; context coefficients $\gamma_i \in \{-2, 0, 2\}$ act on the first two attributes, giving the effective coefficient $\theta_i + \gamma_i c$ in context c; the hidden triple is a single coefficient in $\{-1, 0, 1\}$ on the doubled product $x_0 x_1 x_2$. In the drift family each attribute carries an independent flip bit $f_i \in \{0, 1\}$; the coefficient $\theta_i(1 - 2f_i)$ replaces θ_i for every session after time $N/2$, where N is the history length. The contradiction family has the additive grid; its histories additionally contain four first-person statements, at times $N/5, 2N/5, 3N/5, 4N/5$, each naming a uniformly random attribute and asserting that the user values it or disregards it, with attribute and direction drawn independently of θ. Statements therefore contribute a state-independent likelihood factor (which cancels in the posterior), and the test is whether stated preferences override revealed ones.

The prior is uniform over the full enumerated grid, including all-zero coefficients and the no-flip pattern, and the generator draws each private state uniformly from that grid with no rejection, so the reference posterior is exactly the true posterior. This has a cost. Under the uniform sign prior, pre-trained world knowledge is uninformative or even misleading (a user is as likely to dislike reliability as to like it), and the statements carry no information about θ. The novel-vocabulary and contradiction findings (Section 6.2) thus describe behaviour under a prior adversarial to world knowledge, not natural preferences; Section 6.2 adds a sign-skewed prior. The temperature $\tau = 1.0$ was chosen on development seeds (oracle expected accuracy near 0.8 on additive $A = 6$; exact-posterior gain over the prior above 0.02 at $N = 64$ in every evaluated cell; Appendix F).

3.2 EXPERIENCE, TARGETS, AND STREAMS

Experience options activate at most two attributes; every target has at least one option activating three or more. Hence no target repeats an experience query or its A/B reversal, although shared attributes and sub-combinations permit inference. In the hidden-triple family $x_0 x_1 x_2 = 0$ for every experience option, so hypotheses differing only in v have identical history likelihoods; the term is a deliberately unresolved effect that targets can expose. History sessions $t = 1, \ldots, N$ alternate context ($c = -1$ at odd t, $+1$ at even t) and rendering domain (work at odd t, research at even t); targets sample the unordered pair and the context independently. Each user has 24 T1 targets at time N, the same 24 pairs re-rendered in a third domain as T4, 24 candidate queries at time $N+1$ for acquisition (any pair, excluding T1 keys), and, in the drift family, six T3 targets at each checkpoint $N/4, N/2, 3N/4, N$.

Separate seed-derived streams preserve stationary choice prefixes and target feature pairs across history lengths. Drift labels need not nest: changing N moves the change point $N/2$; contradiction-statement positions also move. The same generating seed pairs vocabularies and families; at fixed N, a contradiction user shares the additive choices and adds four statements. Each user record carries a SHA-256 commitment, the dataset is committed before any system runs, and targets are never filtered on model outcomes (Appendix A).

3.3 RENDERING

A session is one line of prose: an opener, a context cue, "Option A:" with one declarative sentence per present attribute, "Option B:" likewise, and "Which one would I pick?", followed for history items by "User chose Option A." (or B). Each of the eight attribute words (speed, reliability, affordability, privacy, customization, novelty, support, portability) has three templates per domain, sentence order is shuffled, an attribute-free option gets one of four baseline sentences, and half of the options also get one of 24 flavor sentences naming no attribute. Absent attributes are never mentioned, the context cue is a distractor outside the context family, no schema word appears, and an audit parser recovers the exact vectors from every session. Novel vocabulary replaces each attribute by a fictional feature (the Varn module, the Kessel certification, ...) with no glossary, so its meaning is learnable only from choices; statements use the same words.

3.4 WHAT THE SYSTEM SEES

Per history item a system receives the rendered text, its observed choice, an opaque identifier and the session time; per query, text, identifier and time. Feature arrays, integer contexts, probabilities, target labels, seeds and the private state are never sent (Figure 3 in Appendix A). A consolidation procedure maps the history to an artifact m with $|m|_{\mathrm{UTF8}} \le B$, and a fresh prediction call receives only the artifact and the queries; an artifact over its cap is budget-infeasible, never truncated.

4 EXACT REFERENCE AND THE INFORMATION-USE RATIO

For the uniform prior on Θ_F, exact enumeration gives

$$w_\theta = \Pr(\theta \mid D) \propto \frac{1}{|\Theta_F|} \prod_{(z_i,y_i)\in D} p_\theta(z_i)^{y_i}(1-p_\theta(z_i))^{1-y_i}, \qquad \mu_D(z) = \sum_\theta w_\theta p_\theta(z), \tag{2}$$

where D ranges over choice records only, and in the drift family $p_\theta(z_i)$ uses the pre- or post-change coefficients according to t_i. "Exact" denotes enumeration (vectorized, stably normalized, floating point), not sampling or optimization. At a fixed target write $p_\theta = p_\theta(z)$. For any forecast q that is measurable with respect to the history,

$$\mathbb{E}[(Y-q)^2 \mid D, z] = \underbrace{\sum_\theta w_\theta p_\theta(1-p_\theta)}_{\text{label noise}} + \underbrace{\sum_\theta w_\theta (p_\theta - \mu_D)^2}_{\text{unresolved predictive uncertainty}} + \underbrace{(q-\mu_D)^2}_{\text{algorithmic deviation}}. \tag{3}$$

The first two terms sum to $\mu_D(1-\mu_D)$ and do not depend on the system; the last, called `algorithmic` in output records, measures deviation from the reference and does not identify whether the cause is retrieval, encoding, reasoning, or prior mismatch. The identity is not new: it follows from conditional variance and the bias decomposition of squared error, and it differs from Murphy's uncertainty–reliability–resolution partition (Murphy, 1973). Nor does it require identifying parameters, since hypotheses that predict the targets identically add nothing to the unresolved term.

Information-use ratio. Let $B(q) = p_{\theta^*}(1-q)^2 + (1-p_{\theta^*})q^2$ be the expected Brier loss under the true generator and B_{sys} the mean over users of the per-user mean of $B(q)$ over a track's targets (sampled and posterior-conditional losses are kept separate; Appendix B), and B_{prior} and B_{Bayes} the same quantities for the constant forecast $1/2$ and for μ_D. The primary estimand of a cell is

$$\rho = \frac{B_{\mathrm{prior}} - B_{\mathrm{sys}}}{B_{\mathrm{prior}} - B_{\mathrm{Bayes}}}. \tag{4}$$

Table 1: Conditions; $B \in \{512, 2048\}$. Memory rows above the rule are frozen-prompt reimplementations, not products; rows below it are exploratory; the last two are external systems whose stores are measured, not capped, and ran on `gpt-6-astra` only.

Condition	Persistent bytes	Read context	Purpose
no-history	0	queries only	prior anchor
full-history	unrestricted	all N sessions	in-context use
summary-B	$\leq B$	artifact only	consolidation
profile-B	$\leq B$	JSON artifact only	extraction
retrieval-8	unrestricted store	top-8 episodes per target	retrieval
schema-summary-1024	≤ 1024	artifact + family	does disclosure help?
hypothesis-summary-1024	≤ 1024	artifact only	structure prompt
extract-fit	tags + weights	tagged targets	parse, then fit
mem0-8	Mem0 store	top-8 memories per target	off-the-shelf
letta	core memory + buffer	agent context	off-the-shelf

Under the sign-symmetric prior the prior predictive is $1/2$, and the constant forecast $1/2$ has expected Brier loss exactly $1/4$, so $B_{\text{prior}} = 1/4$. The ratio is a ratio of user means, not a mean of per-user ratios. Uncertainty comes from a whole-user paired bootstrap: the same resampled users enter numerator and denominator and all three systems, with 2,000 resamples and a percentile interval. *Margin rule.* ρ is reported only when the point denominator is at least 0.01, no bootstrap replicate falls below that margin, and the cell has at least two users; otherwise the cell is marked not reported and raw losses are shown. Sub-margin replicates are never dropped (that would discard exactly the extreme ratios); the margin was fixed before the main matrix (Appendix F). Pooled intervals resample shared generating seeds jointly across families, with weights normalized over available users in each cell; the six-family pool contains sixteen seed clusters, not ninety-six independent users.

We prefer ρ to raw loss for two reasons. It has a common scale (0 is the prior, 1 the exact Bayesian forecaster), so a cell with a small Bayes gain does not look like a system failure; and it controls for inferability across families, history lengths and hypothesis-space sizes, which raw Brier confounds. It also has limits. It is conditional on the family: structure outside the family earns no credit, and a wrong family would make the denominator wrong. In expectation $\rho \leq 1$ for every history-measurable system, since μ_D minimizes expected loss, but a sample estimate can exceed 1 by chance; negative values are worse than the baseline, and the privileged oracle can exceed 1 even in expectation. Finally, ρ cannot tell information an encoder destroyed from information its reader ignored; that would need the encoder's artifact distribution.

5 Systems and conditions

Language-model conditions. The rows of Table 1 above the rule (eight conditions) are prespecified and run on every family and backbone with the same users, prompts and caps; a ninth, full-history-high (high effort, 8,192-token cap), was withdrawn for operational failures before analysis (Appendix F). Summary and profile memory consolidate the history in one call into a text summary or a JSON profile (Appendix A) under a hard UTF-8 byte cap B, then predict from the artifact alone; a malformed profile is a schema failure. Retrieval embeds every session line and, per target, presents the eight most cosine-similar lines in time order, for the first twelve targets of a track. Schema-summary-1024 prepends the public family description to both prompts. Prediction calls return a JSON probability per query, a track's 24 targets in one call except in retrieval, at low reasoning effort, with completion caps of 4,096 tokens for consolidation and 2,048 for prediction. No call is retried; a failure or timeout (120 s, later 300 to 600 s; Appendix F) fails that user's instance, counted in coverage, not imputed.

Constructive and off-the-shelf conditions. Four exploratory conditions (Appendix F) run on the primary cell, outside the Holm families. Hypothesis-summary-1024 is summary memory whose prompt names the kinds of structure to look for (direction, strength, solo/group dependence, interactions) but not the family, grid or features. In extract-fit the backbone only parses: it proposes at most twelve attribute tags and tags both options of every session and target in batches of 32, and the har-

ness fits the logistic regression of the statistical reference on the tag differences. Two off-the-shelf systems run on `gpt-6-astra` for the first eight users per family, ingesting eight sessions per call to limit cost: mem0-8 (Mem0 2.0.20 (Chhikara et al., 2025), FAISS store, dense retrieval, its own extraction prompt) passes its top eight memories per target to our prediction prompt, twelve targets per user; letta (Letta server 0.16.8, the open-source successor of MemGPT (Packer et al., 2023), default agent and tools, called through the Azure Responses endpoint because chat completions reject function tools for these deployments) answers the 24 queries itself. Their internal calls get one metered transport retry; ours none.

Backbones. Three frontier models: `gpt-5.6-sol` (2026-07-09) and `gpt-6-astra` (2026-09-03), hereafter sol and astra, with `text-embedding-3-small` for retrieval; and `claude-sonnet-5`, on the eight main conditions (Appendix F). All calls use low reasoning effort. Prompts (Appendix A) were frozen after calibration on development seeds, digests in every manifest; held-out seeds postdate freezing, and every call is pre-costed through a budget ledger.

Statistical references. Six references share the public schema (family, A, τ, change time): the prior ($q = 1/2$); the exact μ_D; for drift, a stationary Bayes on the additive grid ignoring the change; a logistic regression on the feature difference $a - b$, misspecified for interactions, context and drift (ℓ_2-regularized; Appendix A); a kNN predictor ($k = 5$); and the privileged oracle p_{θ^*}. Logistic and kNN read structured features, not prose: they diagnose the learning signal, not the best retrieval system; references read the full prefix, and their bytes are measured, not enforced.

Tracks. T1 scores the 24 targets after the history: references at $N \in \{16, 64, 128, 256\}$ for additive $A \in \{4, 6, 8\}$ and the other families at $A = 6$; model conditions at $N = 128$ for the six families at $A = 6$, additive at $A \in \{4, 8\}$ and novel-vocabulary additive; 16 users per cell. T4 (held-out domain) runs for all references and for no-history, full-history, summary-2048 and profile-2048. T3 scores each drift checkpoint from the prefix ending there (exact versus stationary Bayes measures what revision is worth), and T2 grants five acquisitions from the pool and rescores T1 under random, exact one-step value-of-information and two heuristic selectors (Appendix D); both are references only, and model acquisition is future work.

Budget accounting. Every cell reports persistent bytes, read context, compute (tokens, reasoning included) and coverage; public code, the family description and backbone parameters lie outside the byte cap.

6 RESULTS

A *cell* is a (family, condition, backbone) group of 16 users; T1, real vocabulary, $A = 6$, $N = 128$ unless stated; brackets: 95% bootstrap intervals; pooled: equal-weight family means, intervals descriptive outside the confirmatory block; see Table 2, Figure 2 and Appendix B (per-family cells, contrasts, T2).

6.1 HEADLINE: HOW MUCH OF THE HISTORY IS USED

With the full history in context, pooled ρ is 0.590 $[0.497, 0.670]$ for `gpt-6-astra` and -0.148 $[-0.253, -0.062]$ for `gpt-5.6-sol`; both upper bounds lie below 0.9, so H1 (full-history use falls short of the reference) holds. Sol's interval excludes zero: after 128 sessions its forecasts are worse than the constant $1/2$. It is below the prior with an empty memory too (no-history $\rho = -0.21$ $[-0.30, -0.12]$; Table 2), so against its own baseline the history adds $+0.06$ $[-0.02, 0.14]$. Of the undisclosed summary and profile conditions only the 2 KB summary ($+0.15$ $[0.05, 0.26]$) clears that baseline, and the disclosed summary also improves on it (Table 6). For astra the history adds $+0.59$ $[0.50, 0.67]$. Negative ρ thus partly reflects a history-free bias. On the same users the misspecified logistic regression reaches $\rho = 0.940$, 0.816, 0.605, 0.949, 0.462 and 0.940 on additive, pairwise, context, hidden-triple, drift and contradiction (kNN: 0.18 to 0.48, Appendix B): given the true features, a privileged input, a linear fit recovers far more than any backbone gets from prose, and most of it survives parsing (Section 6.3).

Condition	gpt-5.6-sol ρ [95% CI]	k/n	gpt-6-astra ρ [95% CI]	k/n	claude-sonnet-5 ρ [95% CI]	k/n	statistical ρ [95% CI]	k/n
no-history	$-0.21\,[-0.30, -0.12]$	6/6	$0.00\,[0.00, 0.00]$	6/6	$-0.01\,[-0.03, 0.01]$	6/6	–	
full-history	$-0.15\,[-0.25, -0.06]$	6/6	$0.59\,[0.50, 0.67]$	6/6	$-0.01\,[-0.02, 0.01]$	6/6	–	
summary-512	$-0.13\,[-0.23, -0.03]$	6/6	$0.28\,[0.17, 0.38]$	6/6	$-0.02\,[-0.09, 0.04]$	6/6	–	
summary-2048	$-0.07\,[-0.18, 0.03]$	6/6	$0.39\,[0.29, 0.47]$	6/6	$0.02\,[-0.04, 0.06]$	6/6	–	
profile-512	n/r†	0/6	n/r†	0/6	–		–	
profile-2048	$-0.18\,[-0.30, -0.07]$	6/6	$0.38\,[0.28, 0.47]$	6/6	n/r†	0/6	–	
retrieval-8	$-0.62\,[-0.82, -0.48]$	6/6	$-0.26\,[-0.46, -0.10]$	6/6	$-0.07\,[-0.14, 0.02]$	6/6	–	
schema-summary-1024	$0.55\,[0.44, 0.64]$	6/6	$0.82\,[0.77, 0.87]$	6/6	$0.07\,[-0.04, 0.17]$	3/6	–	
hypothesis-summary-1024	$0.14\,[0.03, 0.24]$	6/6	$0.51\,[0.42, 0.59]$	6/6	n/r†	0/6	–	
extract-fit	$0.76\,[0.71, 0.80]$	6/6	$0.78\,[0.73, 0.82]$	6/6	–		–	
letta	–		$0.67\,[0.57, 0.79]$	6/6	–		–	
mem0-8	–		$-0.22\,[-0.44, -0.07]$	6/6	–		–	
logistic	–		–		–		$0.79\,[0.74, 0.83]$	6/6
knn	–		–		–		$0.40\,[0.37, 0.43]$	6/6

Table 2: Information use ρ at the primary cell (track T1, real vocabulary, $A = 6$, $N = 128$): equal-weight mean over families of the per-family ρ, with the 95% percentile interval of the family-mean bootstrap, jointly resampling the actual generation seeds shared across families. k/n: families pooled out of the families the backbone covers at this cell; a family enters only when its cell is reported and reaches 80% user coverage. n/r: no family enters (denominator below the prespecified margin); †: no family enters because the coverage rule excluded the reported ones (fewer than 80% of the 16 intended users); –: condition not run for that backbone. Statistical references are pooled over the same families and are unrestricted controls, not frontier-model rankings. Per-family values: Table 9.

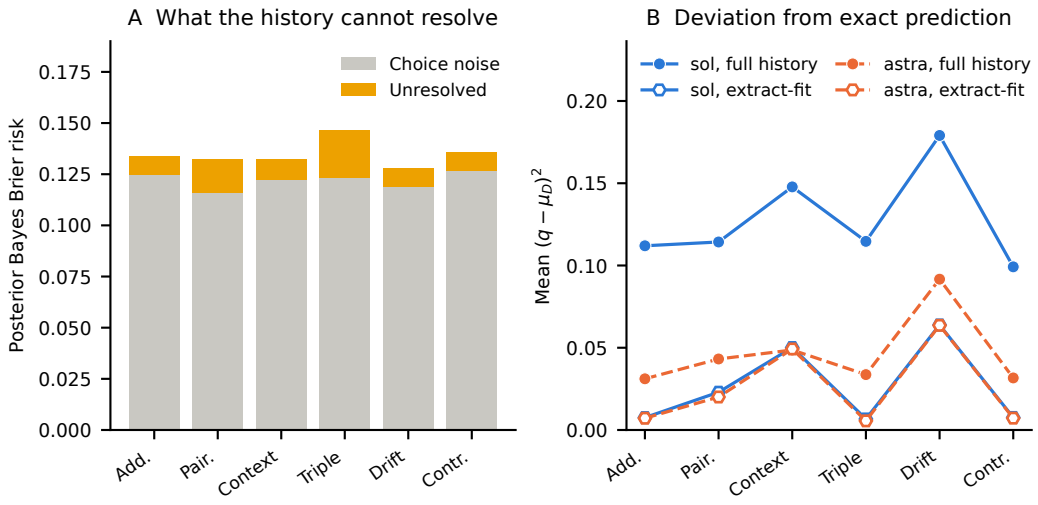


Figure 2: Posterior-risk diagnostics, T1, real vocabulary, $A = 6$, $N = 128$, on targets complete for both GPT backbones and both conditions (13–16 users per family; the diagnostic was not run for `claude-sonnet-5`). A: noise and unresolved uncertainty sum to posterior Bayes risk. B: system deviation $(q - \mu_D)^2$. Parse-then-fit reduces deviation sharply on additive families; context and drift retain larger residuals. These are posterior-conditional quantities, distinct from the true-generator expected losses used in ρ. Raw values and descriptive intervals accompany the released diagnostic report.

What remains unresolved. On common complete targets (Figure 2), hidden-triple has the largest unresolved term (0.023). Astra's full-history deviation on drift is 0.092 against only 0.009 of unresolved uncertainty, and extract-fit brings it down to 0.064; on context, extract-fit leaves it at 0.049. On additive, by contrast, extract-fit cuts deviation from 0.031 to 0.007. The decomposition separates these cases, but it does not pin down a single cause of error.

Memory conditions versus full history. Paired differences against full history (same users, pooled over the six families, Holm over seven prespecified comparisons, including unreportable

cells conservatively as $p = 1$): on astra every compression loses signal: summary-512 -0.305 $[-0.379, -0.240]$, summary-2048 (H2) -0.180 $[-0.229, -0.135]$ and profile-2048 -0.175 $[-0.219, -0.137]$, all Holm $p = 0.007$; even the 2 KB summary keeps only $\rho = 0.390$ of full history's 0.590. On sol, H2 goes the other way: summary-2048 beats the raw history by $+0.090$ $[0.021, 0.155]$ (Holm $p = 0.020$), while profile-2048 -0.012 $[-0.068, 0.045]$ and summary-512 $+0.027$ $[-0.045, 0.097]$ (both Holm $p = 1.00$) show no detectable difference, and the 2 KB summary stays at the prior (pooled $\rho = -0.069$). Retrieval-8 falls below the constant predictor on both backbones: -0.485 $[-0.604, -0.399]$ (sol) and -0.785 $[-0.913, -0.678]$ (astra) against full history on the twelve targets both predicted (Appendix F), Holm $p = 0.007$, pooled $\rho = -0.623$ and -0.256; it meets the prespecified near-zero rule ($\rho < 0.25$, upper bound below 0.4) on both backbones, as does summary-512 on sol. Its forecasts are confidently wrong: predictions above 0.9 realize Option A only 62.9% (sol) and 72.5% (astra) of the time (Figure 4C). The JSON profile fits 512 bytes for only 4 to 9 of 16 users on sol and 9 to 12 on astra, so every profile-512 cell falls below the 80% coverage rule. Coverage over the nine text batches is 72,748 of 79,488 intended targets (91.5%; 95.6% outside profile-512; counts in Appendix B); nothing is imputed, and every reported cell of the six families outside profile-512 retains at least 13 of 16 users, so all six enter every confirmatory comparison.

Third backbone. `claude-sonnet-5` (Appendix F) realizes essentially none of the available gain under our frozen prompts. With the full history $\rho = -0.007$ $[-0.023, 0.008]$, no better than without it, and 82% of its forecasts lie within 0.05 of $1/2$. Summaries and retrieval (-0.07 $[-0.14, 0.02]$) do not raise ρ, no Holm comparison rejects, and most profiles and hypothesis summaries exceed their byte caps. Across the three backbones, full-history use thus ranges from about three fifths of the gain to nothing or less.

Schema ablation. Disclosing the generative family at write and read time, with a 1 KB artifact, raises pooled ρ to 0.552 (sol) and 0.824 (astra): $+0.677$ $[0.565, 0.798]$ and $+0.232$ $[0.167, 0.298]$ against full history, and $+0.628$ $[0.499, 0.762]$ and $+0.435$ $[0.370, 0.503]$ against summary-2048 (H3), all Holm $p = 0.007$. Explicit structural guidance clearly helps. What the comparison cannot do is separate language extraction from arithmetic, because it changes the schema description, the consolidation prompt, the representation and the byte budget at once; that needs a matched design crossing prose versus structured input with hidden versus disclosed schema.

Families, history length and hypothesis-space size. Per family (Figure 5 and Table 9 in Appendix B), full-history ρ runs from 0.711 (additive) to 0.307 (drift) on astra and from 0.013 to -0.496 on sol. The signal is there to be learned: at $N = 16, 64, 128$, pooled logistic ρ (four families at every N) is 0.750, 0.832, 0.825 as the Bayes gain grows. Against grid size (additive, $A = 4, 6, 8$), logistic ρ stays at 0.94 to 0.95 while astra's full-history falls from 0.735 to 0.594 and sol's never exceeds 0.02. On drift, stationary Bayes (ignoring the change) gets $\rho = 0.481$ $[0.201, 0.747]$ on T1; astra's full-history 0.307 is lower. Three families (additive, hidden-triple, contradiction) are nearly saturated by parse-then-regress: logistic averages 0.94 and extract-fit 0.93 (both backbones) there. On the three structured families (context, drift, pairwise; a post hoc split, Table 10) logistic pools to 0.63, extract-fit to 0.59 (sol) and 0.62 (astra), and astra's full-history to 0.48. From $A = 6$ (140,625 states) to $A = 8$ (3,515,625; Table 11) astra's full-history ρ moves only from 0.59 to 0.53 on pairwise and 0.55 to 0.51 on context, and logistic from 0.82 to 0.83 and 0.61 to 0.57, extract-fit tracking it at both sizes. The point estimates move little, although two grid sizes cannot isolate the effect of space size. The logistic value is a fitted baseline, not a ceiling for linear predictors.

6.2 Paired contrasts: rendering shift, vocabulary and stated preferences

Four paired contrasts (Table 5, Appendix B; descriptive, unadjusted). First, T4 minus T1 (the same 24 pairs re-rendered in the held-out domain; equal-weight mean over the six families) is positive on sol, full-history $+0.076$ $[0.034, 0.120]$, and near zero on astra, $+0.011$ $[-0.009, 0.030]$ (other conditions: Appendix B): the domain shift does not measurably hurt full-history use, though this shows neither equivalence nor general domain transfer. Second, novel minus real vocabulary for the same additive users (paired by latent user) is positive under every undisclosed condition: full-history $+0.214$ $[0.095, 0.354]$ (sol) and $+0.110$ $[0.040, 0.198]$ (astra), more under summary-2048 and retrieval-8 ($+0.25$ to $+0.49$); schema-summary is near zero ($+0.093$ and $+0.040$, intervals

covering zero). Semantic expectations clashing with random-signed preferences would explain this, but so could differences in wording and tokenization. Third, contradiction minus additive on seed-paired users is -0.008 $[-0.061, 0.049]$ (sol) and -0.021 $[-0.054, 0.012]$ (astra) under full history, with intervals spanning zero; larger negative contrasts occur for some written memories: on sol summary-2048 -0.213 $[-0.307, -0.109]$ and profile-2048 -0.472 $[-0.706, -0.258]$, on astra profile-2048 -0.096 $[-0.173, -0.030]$. This points to contamination during consolidation; the null full-history contrast does not show immunity. Fourth (exploratory), the sign-skewed additive prior assigns probabilities $(0.05, 0.10, 0.20, 0.30, 0.35)$ to $(-2, \ldots, 2)$, reversed at feature 2. Because feature 2 is affordability, this prior favours expensive options and is only partly aligned with world knowledge. Relative to the constant $1/2$, full-history reaches 0.50 (sol) and 0.87 (astra), but the history-free prior predictive already reaches 0.37. Replacing the constant anchor with the actual prior predictive gives

$$\rho_{\text{history}} = \frac{B_{\text{prior predictive}} - B_{\text{sys}}}{B_{\text{prior predictive}} - B_{\text{Bayes}}}.$$

On matched users and targets this is 0.216 $[-0.141, 0.463]$ for sol and 0.783 $[0.687, 0.861]$ for astra (post hoc; Table 7). This separates prior gain from history use. Still, neither this imperfectly aligned prior nor the uniform-prior vocabulary contrast tells us what would happen with natural preferences.

6.3 Constructive conditions and off-the-shelf systems

All four are exploratory (Tables 2, 6, 9; coverage in Appendix B). Naming the kinds of structure helps but does not replace disclosure: hypothesis-summary-1024 reaches pooled $\rho = 0.14$ $[0.03, 0.24]$ (sol) and 0.51 $[0.42, 0.59]$ (astra), above summary-2048 (-0.07, 0.39) and well below schema-summary-1024 (0.55, 0.82). Parse-then-fit nearly matches the logistic reference: extract-fit reaches 0.76 $[0.71, 0.80]$ on sol and 0.78 $[0.73, 0.82]$ on astra (all six families) against the logistic reference's 0.79 $[0.74, 0.83]$, a gain of $+0.97$ $[0.85, 1.08]$ over sol's own no-history cell. Paired extract-fit minus logistic is -0.022 $[-0.034, -0.012]$ (sol) and -0.009 $[-0.016, -0.003]$ (astra): close, but detectably below the reference (Table 8). Per family it sits just under logistic, from 0.92 and 0.93 on additive (0.94) to 0.43 and 0.45 on drift (0.46). Mem0 stored one paraphrased memory per session, and its top-eight retrieval gives $\rho = -0.22$ $[-0.44, -0.07]$, below the constant prior; its paired difference from retrieval-8, $+0.006$ $[-0.146, 0.174]$, does not establish equivalence. Letta reaches 0.67 $[0.57, 0.79]$, within -0.015 $[-0.076, 0.049]$ of full history on shared users (descriptive; Table 8). Yet in 44 of 45 completed users the agent never called a memory tool and answered from its 30,000-token message buffer: in-context use under another name. Three further agents that did write memory exceeded the per-user cost cap and are reported as beyond it (Section 7).

Bounded acquisition (T2). From the exact posterior at $N = 128$, five acquisitions change expected Brier by well under 0.01 in most families; informed policies beat random on drift ($+0.44$ to $+0.48$, $\times 100$ units) and contradiction ($+0.11$ to $+0.18$), intervals excluding zero (Appendix D).

Implications for memory design. The parse-then-fit result argues for statistical aggregation after language parsing, though it shows neither that parsing is perfect nor that the remaining error is all misspecification. Weak eight-episode retrieval could come from which episodes are selected, how many, or the inference budget.

7 Limitations

The exact reference assumes a declared, correctly specified synthetic family; high information use does not establish performance on human preferences. Rendering uses templates and confounds context with domain in histories. Schema disclosure changes prompts and representation, and the skewed prior reverses affordability. Most conditions are minimal implementations; Mem0 and Letta use only one backbone and eight users per family. Three Letta users that wrote memory exceeded their caps, so completed users under-represent that behaviour. Equal bytes do not equalize compute (Letta: median 79 KB buffer, eighteen calls per user), and all reported calls use low reasoning effort. Three backbones from two providers and sixteen shared generating seeds limit generalization; families do not supply ninety-six independent users, and missing outcomes may be informative. Encoder loss and reader error remain entangled. Matched representation, retrieval and compute controls remain necessary.

AI USE STATEMENT

We used generative AI tools (an AI coding assistant) for implementing methods, in that the software was co-developed with the authors; for generating the synthetic data, which is produced by executable code and contains no human observations; and for monitoring the experiment runs. We did not use them to formulate the research question, the hypotheses or the experimental design, or to analyse and interpret the results; these are the authors' own work, and the paper contains no theoretical results beyond standard identities. Additionally, we used them for code editing, producing figures, searching related literature, formatting references, and wording and rewriting of the text, which the authors edited. AI-assisted code was checked by the authors through review and an automated test suite, and every reported number is regenerated from hash-verified run records. The authors take full responsibility for the content of this paper.

ETHICS STATEMENT

This study uses synthetic records and no human participants. Coefficients are not validated psychological traits. A real-user extension requires informed consent, data minimization, appropriate review, and a defensible withdrawal policy. Synthetic results should not support consequential judgments about individuals. Model outputs were obtained through a paid API with a budget ledger; no attempt was made to extract training data or personal information.

REPRODUCIBILITY STATEMENT

Sections 3–5 specify the families, grids, rendering, information boundary, reference, ratio, margin rule, conditions and prompts; Appendix A gives the prompts verbatim. Every run writes a manifest with configuration and code hashes, prompt digests, the committed dataset hash, per-cell coverage and per-call cost; every user record carries a commitment; reports are produced only from manifests whose hashes verify. Source code, configurations, tests, manifests and run reports are included in the anonymous supplementary material, and each number in the paper can be traced to a run artifact there. Confirmatory seeds held independently of the authors have not yet been run.

REFERENCES

Marcel Binz and Eric Schulz. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023. doi: 10.1073/pnas.2218523120. URL https://arxiv.org/abs/2206.14576.

Wen-Yu Chang and Yun-Nung Chen. When users don't ask: Benchmarking context-driven memory retrieval in conversational agents. *arXiv preprint arXiv:2609.03467*, 2026. URL https://arxiv.org/abs/2609.03467.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025. URL https://arxiv.org/abs/2504.19413.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? A case study of simple function classes. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2208.01066.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://sites.stat.washington.edu/raftery/Research/PDF/Gneiting2007jasa.pdf.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in LLM agents via incremental multi-turn interactions. *arXiv preprint arXiv:2507.05257*, 2025. URL https://arxiv.org/abs/2507.05257.

Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo J. Taylor, and Dan Roth. Know me, respond to me: Benchmarking LLMs for dynamic user profiling and personalized responses at scale. In *Conference on Language Modeling*, 2025. URL https://arxiv.org/abs/2504.14225.

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. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://arxiv.org/abs/2005.11401.

Yifei Li, Weidong Guo, Lingling Zhang, Rongman Xu, Muye Huang, Hui Liu, Lijiao Xu, Yu Xu, and Jun Liu. LoCoMo-Plus: Beyond-factual cognitive memory evaluation framework for LLM agents. *arXiv preprint arXiv:2602.10715*, 2026. URL https://arxiv.org/abs/2602.10715.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870, 2024. doi: 10.18653/v1/2024.acl-long.747. URL https://aclanthology.org/2024.acl-long.747/.

Allan H. Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4):595–600, 1973. doi: 10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023. URL https://arxiv.org/abs/2310.08560.

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. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22, 2023. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

Chonghan Qin, Xiachong Feng, Weitao Ma, Xiaocheng Feng, and Lingpeng Kong. ImplicitMemBench: Measuring unconscious behavioral adaptation in large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28232–28261, 2026a. doi: 10.18653/v1/2026.acl-long.1301. URL https://aclanthology.org/2026.acl-long.1301/.

Yuehan Qin, Li Li, Linxin Song, Wei Yang, Jiate Li, Yuqing Yang, and Yue Zhao. Memory retrieval for changing preferences. *arXiv preprint arXiv:2606.02976*, 2026b. URL https://arxiv.org/abs/2606.02976.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. LaMP: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7370–7392, 2024. doi: 10.18653/v1/2024.acl-long.399. URL https://aclanthology.org/2024.acl-long.399/.

Yiheng Shu, Bernal Jiménez Gutiérrez, Saisri Padmaja Jonnalagedda, Yuguang Yao, Huan Sun, and Yu Su. AgentCL: Toward rigorous evaluation of continual learning in language agents. *arXiv preprint arXiv:2606.02461*, 2026. URL https://arxiv.org/abs/2606.02461.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.10813.

Di Wu, Zixiang Ji, Asmi Kawatkar, Bryan Kwan, Jia-Chen Gu, Nanyun Peng, and Kai-Wei Chang. LongMemEval-V2: Evaluating long-term agent memory toward experienced colleagues. *arXiv preprint arXiv:2605.12493*, 2026. URL https://arxiv.org/abs/2605.12493.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit Bayesian inference. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2111.02080.

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. Do LLMs recognize your preferences? Evaluating personalized preference following in LLMs. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2502.09597.

A PROMPTS

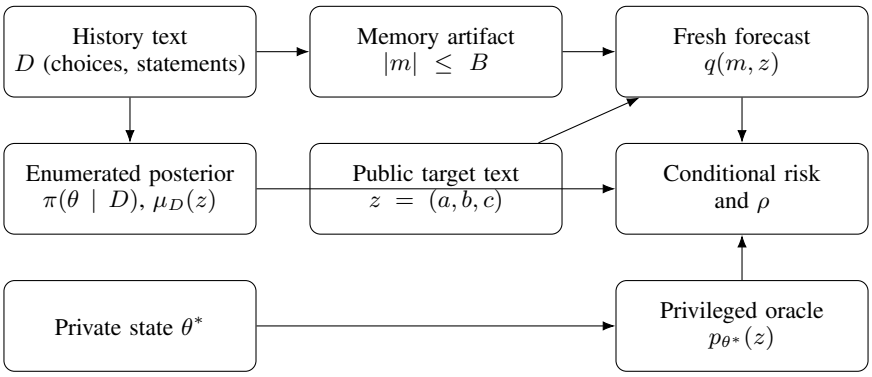


Figure 3: Information boundary. Systems see history text, then a byte-measured artifact (or the full history, or retrieved episodes) and the target text. The evaluator enumerates the posterior over the declared family from the same history and scores every forecast against the exact posterior predictive and against the privileged true probability. Built-in interfaces enforce the whitelist; arbitrary in-process adapters are not adversarially sandboxed.

The user message of a prediction call is the canonical JSON object `{"memory":<text>, "queries":[{"id","time","text"},...]}`; the memory is empty for no-history, the newline-joined `Session t: ...` lines for full-history, the artifact for summary and profile conditions, and the retrieved lines for retrieval. The user message of a consolidation call is the newline-joined session lines.

Table 3: Families and hypothesis counts. $x = (x_0, \ldots, x_{A-1})$; $\theta_i \in \{-2, \ldots, 2\}$; $v, v_{01}, v_{23} \in \{-1, 0, 1\}$; $\gamma_i \in \{-2, 0, 2\}$; $f_i \in \{0, 1\}$. Drift at $A = 8$ exceeds the enumeration limit of 4×10^6 states and is not run.

Family	$\phi_F(x, c)$ beyond x	Extra	A=4	A=6	A=8
Additive	–	–	625	15,625	390,625
Pairwise	$2x_0x_1,\ 2x_2x_3$	v_{01}, v_{23}	5,625	140,625	3,515,625
Context	$cx_0,\ cx_1$	γ_0, γ_1	5,625	140,625	3,515,625
Hidden triple	$2x_0x_1x_2$	v	1,875	46,875	1,171,875
Drift	– (coefficients flip after $N/2$)	$f_0, \ldots, f_{A-1}$	10,000	1,000,000	–
Contradiction	– (plus statements)	–	625	15,625	390,625

Randomness, pairing and commitments (Section 3). Nonlinear terms enter each option's utility before subtraction, so swapping a and b complements the probability in every family. A target or candidate key is (unordered pair, context); keys are unique within a user, sampled by rejection, and then oriented by draw order. Separate streams generate features, label uniforms ($y = \mathbb{1}[u < p]$), rendering, statements and targets. Stationary choice prefixes nest as N changes, but drift labels do not necessarily nest because the change point moves with N; statement positions also depend on N. Feature and label streams are shared across families with the same generating seed, so pooled uncertainty must preserve that dependence. Seeds nest so that users are paired across families and

vocabularies: a public `pair_key` names the latent identity shared by the real and novel renderings of one user, and families with identical grids share one private state, so a contradiction user is literally the additive user plus four statements with identical choice texts. Each user record carries a SHA-256 commitment over its canonical JSON, the dataset is committed before any system runs, and targets are never filtered on model outcomes.

Statistical references (Section 5). The logistic reference takes 300 unit gradient steps on mean logistic loss plus $0.01\|w\|^2$; the kNN predictor uses $k = 5$, inverse-distance weights, a symmetric pseudocount, and is reversal-aware.

Prediction system message. `You are a personal assistant that has worked with one user for a long time. From the memory below, estimate the probability that this user would choose Option A in each new decision. The user's choices are personal and somewhat noisy; do not assume what most people would prefer. Use only the memory and the queries. An empty memory contains no evidence about this user. Return ONLY JSON {"predictions":[{"id":"...","p":<probability between 0 and 1>},...]} with every query id exactly once, finite p between 0 and 1, no other text.`

Summary consolidation system message. `You are consolidating a long record of one user's choices into a compact memory for a future assistant that will not see the record. Write what would help predict this user's future choices between options, including how confident you are. Hard limit {B} UTF-8 bytes; aim for {B//2} characters. Plain text, no markdown.`

Profile consolidation system message. `You are consolidating a long record of one user's choices into a structured memory for a future assistant that will not see the record. Return ONLY JSON of the form {"preferences": [{"about":"...","direction":"likes|dislikes|indifferent","confidence": 0.0,"evidence":"..."}],"notes":"..."}. Hard limit {B} UTF-8 bytes including all JSON syntax; aim for {B//2} characters. No other text.`

Public family description (schema-disclosed condition). Prepended, followed by one space, to both system messages above. The instance for the context family at $A = 6$, $\tau = 1$, real vocabulary reads:

```
Public model class for this user (family: context). Each option is a
0/1 vector over 6 features ordered x0=speed, x1=reliability, x2=cost,
x3=privacy, x4=customization, x5=novelty; a feature is 1 exactly
when the option text mentions it. Each session states whether the
decision is for the user alone (c = -1) or for the whole group (c =
+1). Utility U(x,c) = w0*x0 + w1*x1 + w2*x2 + w3*x3 + w4*x4 + w5*x5
+ c*g0*x0 + c*g1*x1, with g0 and g1 each uniform on {-2,0,2}. Each
w_i is drawn independently and uniformly from the grid {-2,-1,0,1,2}
(a zero makes that feature irrelevant); the prior is uniform over the
full product grid and one draw governs every session. P(User chose
Option A) = sigmoid((U(A,c)-U(B,c))/1), sigmoid(z)=1/(1+exp(-z)); choices
are conditionally independent given the weights. This evaluation is
explicitly conditioned on the public model class.
```

The feature list uses the novel names under novel vocabulary. The utility's family-specific tail is empty for additive, contradiction and drift; `+ 2*v01*x0*x1 + 2*v23*x2*x3, with v01 and v23 each uniform on {-1,0,1}` for pairwise; `+ 2*v*x0*x1*x2, with v uniform on {-1,0,1}` for hidden-triple. For drift the clause `and one draw governs every session` is replaced by `and one draw governs every session up to session N/2, after which each w_i is replaced by w_i*(1-2*f_i) with independent flip bits f_i uniform on {0,1}` with the numeric change time substituted. For contradiction the sentence `First-person statements about preferences in the record were generated independently of the weights and carry no information about them.` precedes the final sentence.

Hypothesis consolidation system message (hypothesis-summary-1024). `You are consolidating a long record of one user's choices into a compact memory for a future assistant that will not see the record. Do not summarize the record. Infer the rule behind this user's choices: which option attributes matter to this user, in which direction and how strongly; whether any attribute matters only when the decision is for the group or only when it is for the user alone; and whether any attributes matter only in combination. State the rule as explicit statements with a confidence for each, name attributes that appear irrelevant, and say what the record does not settle. Hard limit {B} UTF-8 bytes; aim for {B//2} characters. Plain text, no markdown.`

Extract-fit, tag-proposal system message. The user message is the newline-joined session lines.

```
You will read a record of one user's choices between pairs of options
described in prose. List the distinct attributes that the option
descriptions state, as short lowercase tags of one to three words, at
most 12 tags, covering only attributes of the options (not the person,
the setting or the wording). Return ONLY JSON {"tags":["..."]} and no
other text.
```

Extract-fit, tagging system message. The user message is the canonical JSON object `{"sessions":[{"id","text"},...],"tags":[...]}` with at most 32 sessions (or the 24 targets), the observed choice removed from each session text.

```
For each session, mark which of the given tags the description of Option
A states and which the description of Option B states. A tag is present
for an option only if that option's description states the attribute.
Return ONLY JSON {"sessions":[{"id":"...","a":["tag",...],"b":["tag",
...]},...]} with every session id exactly once and tags only from the
given list. No other text.
```

What is not frozen for the off-the-shelf systems. Mem0 and Letta use their own prompts, tools and memory logic, which this harness does not freeze; only our prediction prompt (Mem0) and the query message (Letta) are ours. We record instead the package or image version and configuration (mem0ai 2.0.20 with a FAISS store and dense-only retrieval; Letta server 0.16.8 pinned by image id, agent type `letta_v1_agent`, tools `conversation_search`, `memory_insert`, `memory_replace`, context window 30,000 tokens) and a SHA-256 digest of every system prompt and tool list that passed through the metering proxy. Mem0 sent one system prompt (digest `ad19187a3781`, 644 calls); Letta sent two system prompts (`2524f2fa0a2e`, 1,021 calls; `6cff62edd10c`, 4 calls) and three tool lists (`41f5a15f3327`, `5335f6bd29aa`, `ae4e26b1b859`).

B ADDITIONAL RESULTS

Table 4 places GCE's scoring reference against the memory benchmarks of Section 2. Table 5 gives the paired secondary contrasts of Section 6.2 (contradiction minus additive, novel minus real vocabulary, T4 minus T1; one block per backbone), Table 6 each condition's gain over the backbone's own no-history cell, Table 9 and Figure 5 the per-family cells behind the pooled values of Table 2, Table 10 the post hoc pool over the three structured families and Table 11 the harder regime (pairwise and context at $A = 8$; both Amendment 6), and Table 12 and Figure 6 the bounded-acquisition results of Section 6.3.

Coverage. The nine text batches planned 3,456 user-condition instances: 3,161 completed, 15 partial (retrieval-8, 20 targets missing), 161 budget-infeasible (145 profile-512, 16 profile-2048) and 119 failed (114 request-transport errors, five unparseable responses; Appendix F), for 72,748 of 79,488 intended targets (91.5%; 95.6% outside profile-512); each failure fails its instance (no retries) and nothing is imputed. Outside profile-512 every reported cell retains at least 10 of 16 users (13 in the six $A = 6$ families).

Coverage of the exploratory extensions (Amendment 5). Hypothesis-summary-1024 completed 89 (sol) and 93 (astra) of 96 users, the ten failures all request-transport errors. The sign-skewed batch planned 256 user-condition instances: 228 completed, 2 partial (retrieval-8), 15 budget-infeasible (12 profile-512, 3 profile-2048) and 11 failed in transport. Extract-fit needs six sequential calls per user, so transport failures compound: in the first attempt 160 of 192 cells completed and all 32 failures had reason `call-failed` (none in parsing or fitting), leaving three (family, backbone) cells below 13 of 16 users (context on sol; pairwise and drift on astra). Under the rule fixed from failure reasons alone (Appendix F) the 32 cells were rerun once under the same frozen config: 29 completed and 3 failed again in transport (additive on both backbones, context on sol), so 189 of 192 cells are complete, every (family, backbone) cell has 15 or 16 of 16 users, every completed fit passed its recorded post-fit gradient check, and all six families enter both pooled values. After merging the attempts, Mem0 completed 48 of 48 users and Letta 45 of 48. The three missing Letta users are beyond the per-user cap (their agents made about 32 model calls instead of about 17) and were not rerun (Appendix F), so context, hidden-triple and contradiction have seven of eight Letta users; the six infrastructure failures of the earlier attempts (three per system) were rerun once and all completed. The single transport retry rescued 27 (Mem0) and 23 (Letta) internal calls. Mem0 stored 128 memories per user (132 with the contradiction statements), median 38 KB; Letta's persistent bytes (core memory plus in-context buffer) have median 79 KB, of which core memory is under 0.1%; 44 of the 45 completed agents made no memory-tool call, and the four `memory_insert` calls in total were all made by one user. The marginal values in Table 2 use different completed cohorts (eight versus sixteen intended users per family). Table 8 instead matches users and targets: Letta minus full-history is -0.015 $[-0.076, 0.049]$ over six families, Mem0 minus full-history -0.958 $[-1.181, -0.755]$ over five (pairwise fails the matched denominator rule), and Mem0 minus retrieval-8 $+0.006$ $[-0.146, 0.174]$ over six. Intervals jointly resample actual generating seeds; they supersede an earlier Mem0 interval that was too narrow because it coupled sorted user positions. Comparing complete pipelines does not isolate memory architecture. The third backbone (`claude-sonnet-5`, six batches of 176 user-condition instances; Appendix F) completed 742 of 1,056 instances; 293 were budget-infeasible (all 96 profile-512, 150 of 192 profile-2048, 31 schema-summary-1024, eight each summary-512 and summary-2048) and 21 failed (20 profile-2048 outputs that did not match the profile schema, one request-transport error), for 17,808 of 25,344 intended targets (70.3%); no profile cell of this backbone enters a pooled value, and its schema-summary-1024 pool covers three of six families (Table 2, where profile-512 appears as not run).

Per-family, history-length and grid-size detail (Section 6). Per family, full-history ρ is 0.013, 0.004, -0.342, -0.068, -0.496 and -0.002 (sol) and 0.711, 0.587, 0.555, 0.697, 0.307 and 0.683 (astra) over additive, pairwise, context, hidden-triple, drift and contradiction; summary-2048 is lower on astra everywhere and on sol exceeds the prior by over 0.02 only in additive and pairwise. With more history, at $N = 16$, 64, 128, pooled logistic ρ (four families at every N) is 0.750, 0.832, 0.825 and kNN 0.415, 0.400, 0.439 as the Bayes gain grows. Against grid size (additive, $A = 4, 6, 8$), logistic ρ is 0.941, 0.940, 0.954 and kNN 0.499, 0.440, 0.364, so only kNN among references pays for the larger grid, the backbones more: full-history 0.735, 0.711, 0.594 (astra) and -0.413, 0.013, -0.228 (sol), summary-2048 0.540, 0.527, 0.333 and -0.031, 0.147, 0.028. On drift, stationary Bayes (ignoring the change) gets $\rho = 0.481$ $[0.201, 0.747]$ on T1 (0.605 at T3); astra's full-history 0.307 is lower and no undisclosed sol condition beats the prior.

On the same users the nearest-neighbour predictor reaches $\rho = 0.440$, 0.394, 0.482, 0.462, 0.184 and 0.440 on additive, pairwise, context, hidden-triple, drift and contradiction, and retrieval-8's predictions below 0.1 realize Option A 38.6% (sol) and 28.2% (astra) of the time.

Three evaluations that must not be pooled (Section 4). We retain the sampled Brier loss $(Y - q)^2$, the expected Brier loss under the true generator, $B(q) = p_{\theta^*}(1 - q)^2 + (1 - p_{\theta^*})q^2$, and the posterior-conditional loss of Equation 3, with clipped log loss, expected accuracy, oracle regret and reliability bins. The oracle knowing θ^* is privileged: it is not history-measurable, so its distance from μ_D is not an algorithmic failure. Bayesian optimality holds under the declared prior and likelihood, not for every user's realized loss.

T4 minus T1, all conditions (Section 6.2). T4 minus T1 (the same 24 pairs re-rendered in the held-out domain; equal-weight mean over the six families) is positive on sol, full-history $+0.076$ $[0.034, 0.120]$, summary-2048 $+0.043$ $[0.018, 0.068]$, profile-2048 $+0.083$ $[0.054, 0.114]$

Benchmark	Histories	Scoring reference	Best loss	Byte cap	Stated/rev.	Novel
LongMemEval (2025)	LLM chats, human questions	human answer, LLM judge	no	no	no	no
LongMemEval-V2 (2026)	collected browser trajectories	human answer, LLM judge	no	no	no	no
LoCoMo (2024)	LLM dialogues, human-edited	human answer, LLM judge	no	no	no	no
LoCoMo-Plus (2026)	LoCoMo plus implicit constraints	constraint-consistency judge	no	no	no	no
PersonaMem (2025)	LLM-generated persona histories	annotated answer	no	no	no	no
MemoryAgentBench (2025)	reformulated datasets, incremental	dataset label	no	no	no	no
ImplicitMemBench (2026a)	procedural interaction records	rule checks, LLM judge	no	no	no	no
GCE (this paper)	executable synthetic users, prose	exact Bayesian posterior	yes	yes	yes	yes

Table 4: What the scoring reference of each benchmark can say. "Best loss": the evaluator knows the minimum posterior expected loss for each query, so a system's shortfall separates into unresolvable and unused information. "Byte cap": persistent user-specific state is measured and capped in bytes. "Stated/revealed": the benchmark contains stated preferences generated independently of the behaviour that reveals them. "Novel words": a variant replacing familiar attribute names with invented labels. GCE also evaluates off-the-shelf memory systems on the same users: Mem0 (Chhikara et al., 2025) and Letta (the open-source successor of MemGPT; Packer et al., 2023). GCE pays for these with synthetic users and a declared family (Section 7).

and no-history $+0.159$ $[0.058, 0.250]$, and near zero on astra, full-history $+0.011$ $[-0.009, 0.030]$, summary-2048 $+0.008$ $[-0.004, 0.019]$ and profile-2048 -0.005 $[-0.038, 0.024]$ (its no-history is the prior on both tracks).

Condition	contradiction−additive	additive-skewed−additive	novel−real	T4−T1
gpt-5.6-sol				
no-history	$-0.10\,[-0.47, 0.25]$ (15)	$+0.17\,[-0.13, 0.47]$ (14)	$+0.13\,[-0.13, 0.41]$ (15)	$+0.16\,[0.06, 0.25]$ (6/6)
full-history	$-0.01\,[-0.06, 0.05]$ (15)	$+0.45\,[0.11, 0.76]$ (15)	$+0.21\,[0.09, 0.35]$ (16)	$+0.08\,[0.03, 0.12]$ (6/6)
summary-512	$-0.16\,[-0.41, 0.07]$ (15)	$+0.46\,[0.14, 0.78]$ (15)	$+0.34\,[0.13, 0.56]$ (15)	–
summary-2048	$-0.21\,[-0.31, -0.11]$ (14)	$+0.39\,[0.07, 0.71]$ (14)	$+0.32\,[0.11, 0.53]$ (13)	$+0.04\,[0.02, 0.07]$ (6/6)
profile-512	n/r	n/r	n/r	–
profile-2048	$-0.47\,[-0.71, -0.26]$ (15)	n/r	$+0.30\,[0.13, 0.48]$ (14)	$+0.08\,[0.05, 0.11]$ (6/6)
retrieval-8	$+0.02\,[-0.10, 0.13]$ (16)	$+0.65\,[0.15, 1.15]$ (16)	$+0.34\,[0.13, 0.55]$ (16)	–
schema-summary-1024	$-0.11\,[-0.37, 0.11]$ (11)	$+0.11\,[-0.13, 0.33]$ (13)	$+0.09\,[-0.04, 0.26]$ (13)	–
hypothesis-summary-1024	$-0.19\,[-0.35, -0.04]$ (13)	–	–	–
extract-fit	$-0.00\,[-0.01, 0.01]$ (15)	–	–	–
gpt-6-astra				
no-history	$+0.00\,[0.00, 0.00]$ (15)	$+0.00\,[0.00, 0.00]$ (14)	$+0.00\,[0.00, 0.00]$ (14)	$+0.00\,[0.00, 0.00]$ (6/6)
full-history	$-0.02\,[-0.05, 0.01]$ (14)	$+0.13\,[-0.02, 0.29]$ (15)	$+0.11\,[0.04, 0.20]$ (16)	$+0.01\,[-0.01, 0.03]$ (6/6)
summary-512	$-0.09\,[-0.23, 0.02]$ (15)	$+0.23\,[-0.16, 0.58]$ (15)	$+0.39\,[0.17, 0.61]$ (16)	–
summary-2048	$-0.04\,[-0.11, 0.01]$ (13)	$+0.29\,[0.08, 0.52]$ (15)	$+0.25\,[0.13, 0.39]$ (15)	$+0.01\,[-0.00, 0.02]$ (6/6)
profile-512	n/r	n/r	n/r	–
profile-2048	$-0.10\,[-0.17, -0.03]$ (15)	$+0.20\,[-0.10, 0.46]$ (15)	$+0.25\,[0.09, 0.43]$ (14)	$-0.01\,[-0.04, 0.02]$ (6/6)
retrieval-8	$+0.01\,[-0.05, 0.07]$ (16)	$+0.50\,[0.09, 0.92]$ (16)	$+0.48\,[0.15, 0.81]$ (16)	–
schema-summary-1024	$-0.02\,[-0.08, 0.03]$ (11)	$+0.08\,[-0.09, 0.21]$ (13)	$+0.04\,[-0.03, 0.10]$ (13)	–
hypothesis-summary-1024	$+0.03\,[-0.06, 0.12]$ (14)	–	–	–
extract-fit	$-0.00\,[-0.00, 0.00]$ (15)	–	–	–
letta	$-0.04\,[-0.14, 0.07]$ (7)	–	–	–
mem0-8	$-0.10\,[-0.26, 0.01]$ (8)	–	–	–
claude-sonnet-5				
no-history	$-0.00\,[-0.05, 0.04]$ (16)	–	–	$+0.00\,[-0.01, 0.02]$ (6/6)
full-history	$+0.01\,[-0.03, 0.04]$ (16)	–	–	$-0.00\,[-0.02, 0.02]$ (6/6)
summary-512	$+0.01\,[-0.10, 0.13]$ (14)	–	–	–
summary-2048	$-0.01\,[-0.16, 0.11]$ (14)	–	–	$-0.02\,[-0.05, 0.02]$ (6/6)
profile-2048	–	–	–	n/r (0/5)
retrieval-8	$-0.01\,[-0.05, 0.04]$ (16)	–	–	–
schema-summary-1024	n/r	–	–	–

Table 5: Prespecified paired secondary contrasts (descriptive, unadjusted): differences in ρ with 95% whole-user paired-bootstrap intervals at $N = 128$, $A = 6$, track T1. contradiction−additive: users paired by seed (same features, labels and targets), real vocabulary; (n) user pairs. additive-skewed−additive: users paired by seed (same features and sessions; the labels differ with the latent state), real vocabulary; (n) user pairs. novel−real: the additive family paired by `pair_key`; (n) user pairs. T4−T1: same users, targets re-rendered in the held-out domain, equal-weight mean over the families whose paired difference is reported (k/n families). n/r: not reported (a ratio below the prespecified margin, a cell below 80% user coverage, or an invalid joint bootstrap replicate); –: no paired prediction cells available.

Condition	`gpt-5.6-sol`	`gpt-6-astra`	`claude-sonnet-5`
full-history	$+0.06\,[-0.02, 0.14]$ (6/6)	$+0.59\,[0.50, 0.67]$ (6/6)	$+0.00\,[-0.02, 0.02]$ (6/6)
summary-512	$+0.07\,[-0.03, 0.17]$ (6/6)	$+0.28\,[0.17, 0.39]$ (6/6)	$-0.01\,[-0.07, 0.04]$ (6/6)
summary-2048	$+0.15\,[0.05, 0.26]$ (6/6)	$+0.39\,[0.30, 0.47]$ (6/6)	$+0.03\,[-0.03, 0.06]$ (6/6)
profile-2048	$+0.04\,[-0.07, 0.13]$ (6/6)	$+0.38\,[0.28, 0.47]$ (6/6)	–
retrieval-8	$-0.41\,[-0.55, -0.30]$ (6/6)	$-0.25\,[-0.46, -0.10]$ (6/6)	$-0.06\,[-0.13, 0.01]$ (6/6)
schema-summary-1024	$+0.75\,[0.61, 0.90]$ (6/6)	$+0.83\,[0.77, 0.88]$ (6/6)	$+0.06\,[-0.02, 0.14]$ (3/6)
hypothesis-summary-1024	$+0.37\,[0.29, 0.45]$ (6/6)	$+0.50\,[0.41, 0.59]$ (6/6)	–
extract-fit	$+0.97\,[0.85, 1.08]$ (6/6)	$+0.78\,[0.73, 0.82]$ (6/6)	–
letta	–	$+0.67\,[0.57, 0.79]$ (6/6)	–
mem0-8	–	$-0.22\,[-0.44, -0.07]$ (6/6)	–

Table 6: Gain over the backbone's own no-history cell: $\rho(\text{condition}) - \rho(\text{no-history})$ at the primary cell (track T1, real vocabulary, $A = 6$, $N = 128$), equal-weight mean over families of the per-family whole-user paired difference with the family-mean bootstrap interval (k/n families reported). Descriptive and unadjusted: a backbone whose empty-memory forecasts already sit below the prior is credited here for what the history adds relative to its own baseline, which ρ alone does not separate. –: condition not run.

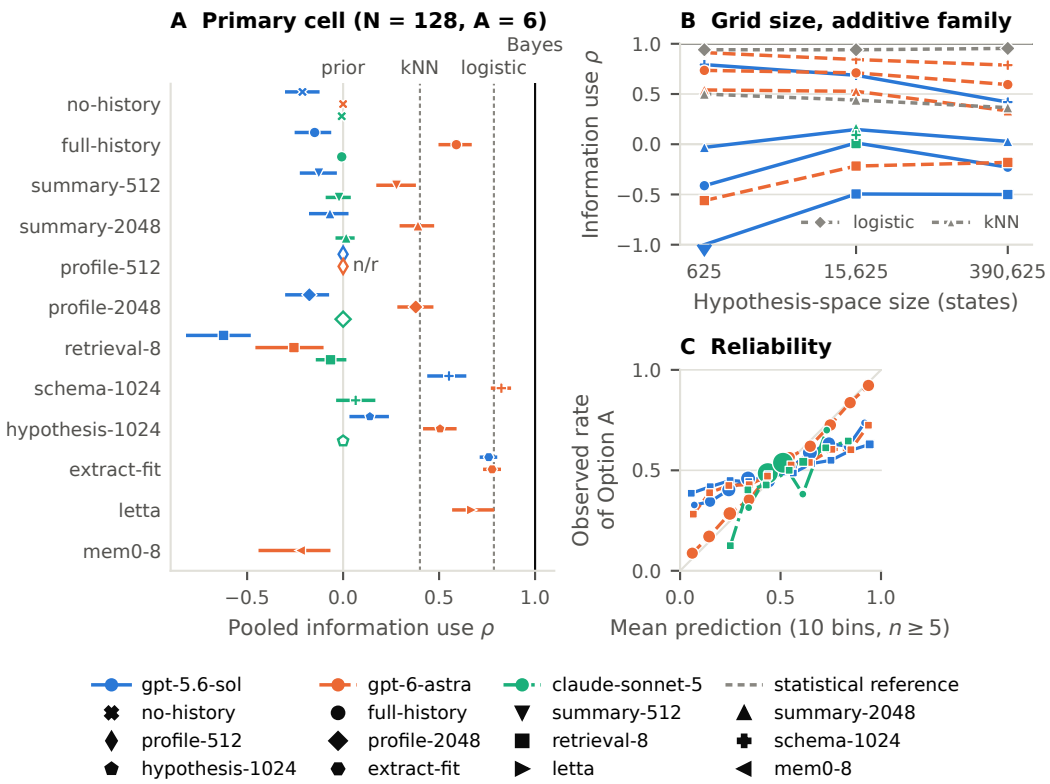


Figure 4: Information use ρ (T1, real vocabulary, $N = 128$). A: pooled ρ at $A = 6$, 95% shared-generation-seed bootstrap intervals; vertical lines: prior ($\rho = 0$), Bayes ($\rho = 1$), logistic, kNN; profile-512 n/r (open marker). B: additive family against hypothesis-space size ($A = 4, 6, 8$), references in grey, ρ clipped to $[-1, 1]$. C: reliability of full-history and retrieval-8: observed Option A rate against mean prediction in ten bins ($n \geq 5$), all $N = 128$ runs. Blue solid: sol; orange dashed: astra; green dash-dot: `claude-sonnet-5` (primary cell only). Labels schema-1024 and hypothesis-1024 abbreviate schema-summary-1024 and hypothesis-summary-1024.

Backbone	Condition	ρ_{history}	95% interval
sol	full-history	0.216	[-0.141, 0.463]
sol	retrieval-8	-0.415	[-1.059, 0.089]
sol	schema-summary-1024	0.693	[0.522, 0.834]
sol	summary-2048	0.293	[-0.028, 0.559]
astra	full-history	0.783	[0.687, 0.861]
astra	retrieval-8	-0.196	[-0.686, 0.217]
astra	schema-summary-1024	0.840	[0.713, 0.967]
astra	summary-2048	0.621	[0.448, 0.773]
statistical	prior-predictive	0.000	[0.000, 0.000]

Table 7: Post hoc sensitivity analysis of the legacy skewed-prior experiment, with the actual prior predictive as the zero anchor. Each ratio and interval uses the same users and targets for system, prior predictive and Bayes. The denominator margin remains 0.01. This prior penalizes affordability and is not fully aligned with ordinary preferences.

A	Backbone / pipeline	Baseline	$\Delta\rho$ [95% CI]	k/K	Pairs/family
6	5.6-sol / extract-fit	logistic	-0.022 [-0.034, -0.012]	6/6	15–16
6	6-astra / extract-fit	logistic	-0.009 [-0.016, -0.003]	6/6	15–16
6	6-astra / mem0-8	full-history	-0.958 [-1.181, -0.755]	5/6	6–8
6	6-astra / mem0-8	retrieval-8	+0.006 [-0.146, 0.174]	6/6	8
6	6-astra / mem0-8	logistic	-1.011 [-1.227, -0.859]	6/6	8
6	6-astra / letta	full-history	-0.015 [-0.076, 0.049]	6/6	5–8
6	6-astra / letta	retrieval-8	+0.924 [0.733, 1.180]	6/6	7–8
6	6-astra / letta	logistic	-0.126 [-0.208, -0.049]	6/6	7–8
8	5.6-sol / extract-fit	logistic	-0.009 [-0.021, 0.000]	2/2	14–16
8	6-astra / extract-fit	logistic	-0.002 [-0.007, 0.003]	2/2	16

Table 8: Descriptive complete-pipeline differences on the same users and targets, equal-weight pooled over available families, with 95% generation-seed bootstrap intervals. k/K is the number of reportable paired families out of the evaluated families; positive values favor the named pipeline. Statistical logistic is unrestricted; comparisons with Mem0 and Letta do not isolate their memory component. No multiplicity-adjusted claims.

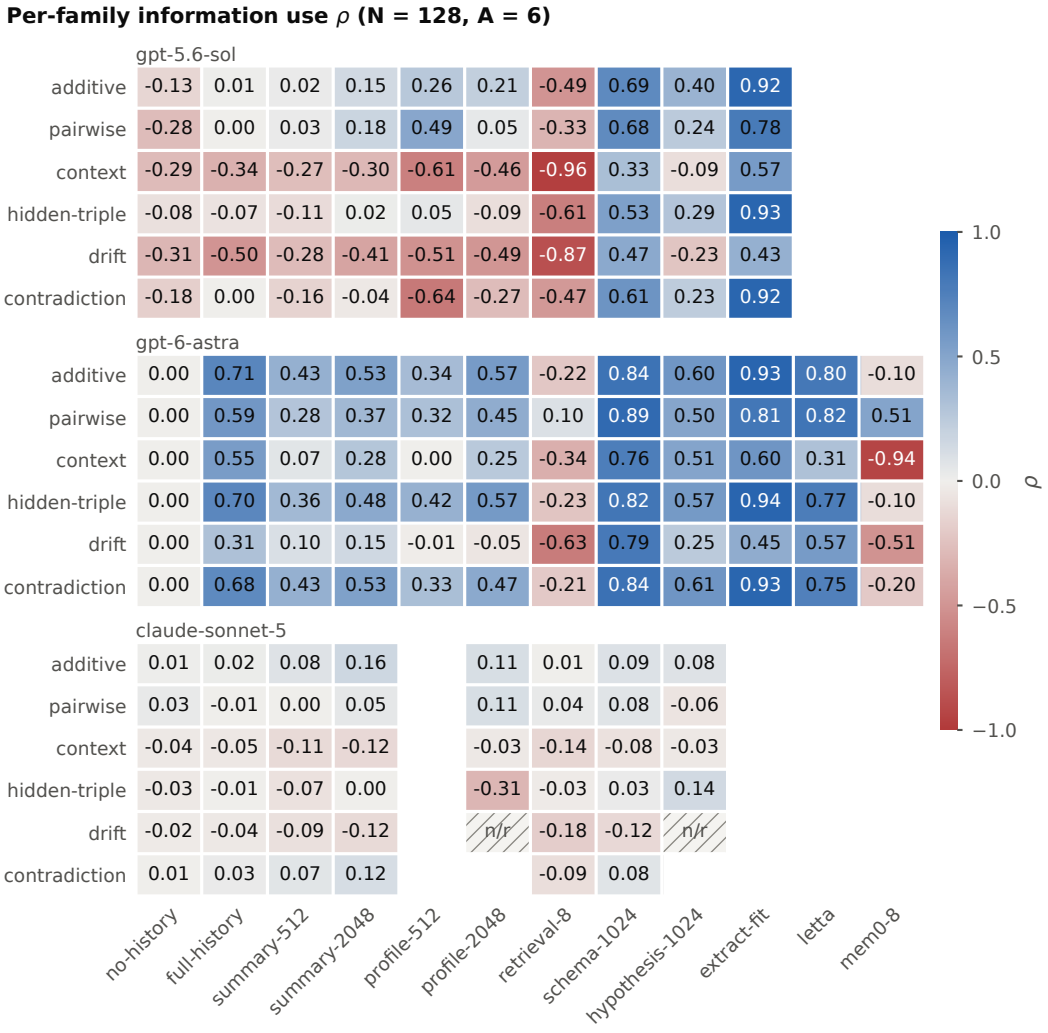


Figure 5: Per-family information use ρ at the primary cell (T1, real vocabulary, $N = 128$, $A = 6$), one block per backbone, rows families, columns conditions; the cell text is ρ and the colour scale clips it to $[-1, 1]$. The values are those of Table 9; letta and mem0-8 ran on astra only, with eight users per family.

Condition	gpt-5.6-sol	gpt-6-astra	claude-sonnet-5	statistical
additive				
no-history	-0.13 [-0.42, 0.12]	0.00 [0.00, 0.00]	0.01 [-0.04, 0.07]	–
full-history	0.01 [-0.25, 0.28]	0.71 [0.57, 0.83]	0.02 [-0.01, 0.06]	–
summary-512	0.02 [-0.27, 0.31]	0.43 [0.17, 0.67]	0.08 [-0.05, 0.23]	–
summary-2048	0.15 [-0.08, 0.36]	0.53 [0.32, 0.70]	0.16 [0.05, 0.28]	–
profile-512	0.26 [-0.21, 0.70]	0.34 [-0.05, 0.70]	–	–
profile-2048	0.21 [-0.04, 0.46]	0.57 [0.36, 0.76]	0.11 [0.00, 0.18]	–
retrieval-8	-0.49 [-0.91, -0.11]	-0.22 [-0.59, 0.11]	0.01 [-0.20, 0.24]	–
schema-summary-1024	0.69 [0.51, 0.84]	0.84 [0.76, 0.93]	0.09 [-0.14, 0.30]	–
hypothesis-summary-1024	0.40 [0.18, 0.61]	0.60 [0.46, 0.73]	0.08 [-0.00, 0.20]	–
prior	–	–	–	0.00 [0.00, 0.00]
bayes	–	–	–	1.00 [1.00, 1.00]
logistic	–	–	–	0.94 [0.91, 0.97]
knn	–	–	–	0.44 [0.37, 0.50]
oracle	–	–	–	1.07 [1.04, 1.12]
extract-fit	0.92 [0.87, 0.96]	0.93 [0.88, 0.97]	–	–
letta	–	0.80 [0.63, 0.94]	–	–
mem0-8	–	-0.10 [-0.65, 0.33]	–	–
pairwise				
no-history	-0.28 [-0.51, -0.05]	0.00 [0.00, 0.00]	0.03 [-0.01, 0.09]	–
full-history	0.00 [-0.32, 0.31]	0.59 [0.38, 0.77]	-0.01 [-0.05, 0.04]	–
summary-512	0.03 [-0.31, 0.34]	0.28 [-0.03, 0.57]	-0.00 [-0.12, 0.12]	–
summary-2048	0.18 [-0.15, 0.48]	0.37 [0.07, 0.64]	0.05 [-0.05, 0.18]	–
profile-512	0.49 [0.05, 0.82]	0.32 [-0.13, 0.68]	–	–
profile-2048	0.05 [-0.34, 0.42]	0.45 [0.25, 0.65]	0.11 [0.06, 0.23]	–
retrieval-8	-0.33 [-0.83, 0.11]	0.10 [-0.27, 0.40]	0.04 [-0.13, 0.17]	–
schema-summary-1024	0.68 [0.42, 0.87]	0.89 [0.78, 0.98]	0.08 [-0.07, 0.22]	–
hypothesis-summary-1024	0.24 [-0.08, 0.57]	0.50 [0.30, 0.69]	-0.06 [-0.21, 0.13]	–
prior	–	–	–	0.00 [0.00, 0.00]
bayes	–	–	–	1.00 [1.00, 1.00]
logistic	–	–	–	0.82 [0.73, 0.91]
knn	–	–	–	0.39 [0.33, 0.46]
oracle	–	–	–	1.21 [1.12, 1.34]
extract-fit	0.78 [0.69, 0.87]	0.81 [0.72, 0.91]	–	–
letta	–	0.82 [0.74, 0.93]	–	–
mem0-8	–	0.51 [0.18, 0.73]	–	–
context				
no-history	-0.29 [-0.65, -0.03]	0.00 [0.00, 0.00]	-0.04 [-0.08, -0.00]	–
full-history	-0.34 [-0.58, -0.13]	0.55 [0.41, 0.68]	-0.05 [-0.07, -0.02]	–
summary-512	-0.27 [-0.53, -0.04]	0.07 [-0.19, 0.33]	-0.11 [-0.23, -0.01]	–
summary-2048	-0.30 [-0.57, -0.06]	0.28 [0.08, 0.46]	-0.12 [-0.30, 0.03]	–
profile-512	-0.61 [-1.05, -0.25]	-0.00 [-0.39, 0.34]	–	–
profile-2048	-0.46 [-0.86, -0.15]	0.25 [0.03, 0.46]	-0.03 [-0.25, 0.08]	–
retrieval-8	-0.96 [-1.39, -0.64]	-0.34 [-0.67, -0.08]	-0.14 [-0.26, -0.04]	–
schema-summary-1024	0.33 [0.17, 0.49]	0.76 [0.61, 0.88]	-0.08 [-0.18, 0.02]	–
hypothesis-summary-1024	-0.09 [-0.37, 0.13]	0.51 [0.36, 0.64]	-0.03 [-0.23, 0.12]	–
prior	–	–	–	0.00 [0.00, 0.00]
bayes	–	–	–	1.00 [1.00, 1.00]
logistic	–	–	–	0.61 [0.45, 0.73]
knn	–	–	–	0.48 [0.42, 0.54]
oracle	–	–	–	1.08 [1.05, 1.10]
extract-fit	0.57 [0.42, 0.69]	0.60 [0.45, 0.73]	–	–
letta	–	0.31 [0.06, 0.58]	–	–
mem0-8	–	-0.94 [-1.52, -0.49]	–	–
hidden-triple				
no-history	-0.08 [-0.25, 0.04]	0.00 [0.00, 0.00]	-0.03 [-0.07, -0.00]	–
full-history	-0.07 [-0.42, 0.21]	0.70 [0.51, 0.85]	-0.01 [-0.06, 0.05]	–
summary-512	-0.11 [-0.48, 0.24]	0.36 [0.03, 0.63]	-0.07 [-0.38, 0.17]	–
summary-2048	0.02 [-0.33, 0.28]	0.48 [0.21, 0.69]	-0.00 [-0.16, 0.13]	–
profile-512	0.05 [-0.52, 0.57]	0.42 [0.02, 0.72]	–	–
profile-2048	-0.09 [-0.48, 0.21]	0.57 [0.32, 0.79]	-0.31 [-0.44, -0.20]	–
retrieval-8	-0.61 [-1.18, -0.15]	-0.23 [-0.71, 0.15]	-0.03 [-0.19, 0.12]	–
schema-summary-1024	0.53 [0.16, 0.81]	0.82 [0.68, 0.92]	0.03 [-0.14, 0.18]	–
hypothesis-summary-1024	0.29 [-0.02, 0.52]	0.57 [0.33, 0.77]	0.14 [0.13, 0.16]	–
prior	–	–	–	0.00 [0.00, 0.00]
bayes	–	–	–	1.00 [1.00, 1.00]
logistic	–	–	–	0.95 [0.90, 0.98]
knn	–	–	–	0.46 [0.38, 0.54]
oracle	–	–	–	1.21 [1.15, 1.28]
extract-fit	0.93 [0.88, 0.97]	0.94 [0.88, 0.98]	–	–
letta	–	0.77 [0.55, 0.94]	–	–
mem0-8	–	-0.10 [-0.53, 0.25]	–	–
drift				
no-history	-0.31 [-0.62, -0.03]	0.00 [0.00, 0.00]	-0.02 [-0.08, 0.03]	–
full-history	-0.50 [-0.81, -0.19]	0.31 [0.11, 0.51]	-0.04 [-0.08, 0.00]	–
summary-512	-0.28 [-0.57, 0.01]	0.10 [-0.11, 0.32]	-0.09 [-0.28, 0.06]	–
summary-2048	-0.41 [-0.71, -0.09]	0.15 [-0.11, 0.43]	-0.12 [-0.26, 0.04]	–
profile-512	-0.51 [-0.91, -0.10]	-0.01 [-0.33, 0.32]	–	–
profile-2048	-0.49 [-0.83, -0.17]	-0.05 [-0.38, 0.24]	n/r	–
retrieval-8	-0.87 [-1.39, -0.43]	-0.63 [-1.04, -0.24]	-0.18 [-0.39, -0.00]	–
schema-summary-1024	0.47 [0.19, 0.73]	0.79 [0.71, 0.87]	-0.12 [-0.22, -0.02]	–
hypothesis-summary-1024	-0.23 [-0.57, 0.05]	0.25 [0.04, 0.45]	n/r	–
prior	–	–	–	0.00 [0.00, 0.00]
bayes	–	–	–	1.00 [1.00, 1.00]
bayes-stationary	–	–	–	0.48 [0.20, 0.75]
logistic	–	–	–	0.46 [0.23, 0.69]
knn	–	–	–	0.18 [0.07, 0.29]
oracle	–	–	–	1.08 [1.05, 1.11]
extract-fit	0.43 [0.18, 0.66]	0.45 [0.22, 0.68]	–	–
letta	–	0.57 [0.26, 0.82]	–	–

mem0-8	–	-0.51 [-1.08, -0.00]	–	–
contradiction				
no-history	-0.18 [-0.47, 0.05]	0.00 [0.00, 0.00]	0.01 [-0.01, 0.03]	–
full-history	-0.00 [-0.29, 0.27]	0.68 [0.51, 0.82]	0.03 [-0.01, 0.06]	–
summary-512	-0.16 [-0.54, 0.18]	0.43 [0.18, 0.64]	0.07 [-0.02, 0.14]	–
summary-2048	-0.04 [-0.28, 0.21]	0.53 [0.32, 0.69]	0.12 [-0.04, 0.26]	–
profile-512	-0.64 [-1.04, -0.17]	0.33 [-0.01, 0.63]	–	–
profile-2048	-0.27 [-0.64, 0.08]	0.47 [0.23, 0.68]	–	–
retrieval-8	-0.47 [-0.83, -0.12]	-0.21 [-0.59, 0.12]	-0.09 [-0.23, 0.05]	–
schema-summary-1024	0.61 [0.40, 0.82]	0.84 [0.76, 0.91]	0.08 [-0.21, 0.33]	–
hypothesis-summary-1024	0.23 [-0.07, 0.46]	0.61 [0.41, 0.79]	–	–
prior	–	–	–	0.00 [0.00, 0.00]
bayes	–	–	–	1.00 [1.00, 1.00]
logistic	–	–	–	0.94 [0.91, 0.97]
knn	–	–	–	0.44 [0.37, 0.49]
oracle	–	–	–	1.07 [1.04, 1.12]
extract-fit	0.92 [0.88, 0.96]	0.93 [0.88, 0.97]	–	–
letta	–	0.75 [0.54, 0.88]	–	–
mem0-8	–	-0.20 [-0.87, 0.29]	–	–
additive-skewed				
no-history	0.02 [-0.07, 0.13]	0.00 [0.00, 0.00]	–	–
full-history	0.50 [0.31, 0.68]	0.87 [0.79, 0.92]	–	–
summary-512	0.54 [0.28, 0.73]	0.66 [0.43, 0.81]	–	–
summary-2048	0.54 [0.29, 0.73]	0.76 [0.60, 0.87]	–	–
profile-512	0.64 [0.38, 0.85]	0.77 [0.57, 0.90]	–	–
profile-2048	0.36 [0.01, 0.61]	0.74 [0.55, 0.86]	–	–
retrieval-8	0.17 [-0.24, 0.49]	0.29 [-0.02, 0.56]	–	–
schema-summary-1024	0.81 [0.69, 0.90]	0.91 [0.80, 0.98]	–	–
prior	–	–	–	0.00 [0.00, 0.00]
bayes	–	–	–	1.00 [1.00, 1.00]
logistic	–	–	–	0.94 [0.91, 0.98]
knn	–	–	–	0.39 [0.32, 0.46]
oracle	–	–	–	1.08 [1.04, 1.14]
prior-predictive	–	–	–	0.37 [0.19, 0.52]

Table 9: Per-family information use ρ with 95% user-bootstrap intervals at $N = 128$, $A = 6$, track T1, real vocabulary, one block per family; these are the values pooled in Table 2. Extension families (additive-skewed) are shown for reference and do not enter any pooled value. n/r: denominator below the prespecified margin, not reported; –: condition not run for that backbone. Statistical references are unrestricted controls, not frontier-model rankings.

Condition	gpt-5.6-sol ρ [95% CI]	k/n	gpt-6-astra ρ [95% CI]	k/n	claude-sonnet-5 ρ [95% CI]	k/n	statistical ρ [95% CI]	k/n
no-history	$-0.29\,[-0.42, -0.17]$	3/3	$0.00\,[0.00, 0.00]$	3/3	$-0.01\,[-0.03, 0.01]$	3/3	–	
full-history	$-0.28\,[-0.40, -0.16]$	3/3	$0.48\,[0.39, 0.58]$	3/3	$-0.03\,[-0.05, -0.01]$	3/3	–	
summary-512	$-0.17\,[-0.31, -0.04]$	3/3	$0.15\,[0.05, 0.24]$	3/3	$-0.07\,[-0.13, -0.00]$	3/3	–	
summary-2048	$-0.18\,[-0.32, -0.05]$	3/3	$0.27\,[0.14, 0.39]$	3/3	$-0.06\,[-0.14, 0.01]$	3/3	–	
profile-512	n/r†	0/3	n/r†	0/3	–		–	
profile-2048	$-0.30\,[-0.48, -0.14]$	3/3	$0.22\,[0.10, 0.33]$	3/3	n/r†	0/3	–	
retrieval-8	$-0.72\,[-0.93, -0.53]$	3/3	$-0.29\,[-0.47, -0.14]$	3/3	$-0.10\,[-0.18, -0.02]$	3/3	–	
schema-summary-1024	$0.49\,[0.36, 0.61]$	3/3	$0.81\,[0.77, 0.86]$	3/3	$0.08\,[-0.06, 0.22]$	1/3	–	
hypothesis-summary-1024	$-0.03\,[-0.14, 0.08]$	3/3	$0.42\,[0.33, 0.50]$	3/3	n/r†	0/3	–	
extract-fit	$0.59\,[0.51, 0.67]$	3/3	$0.62\,[0.54, 0.71]$	3/3	–		–	
letta	–		$0.57\,[0.41, 0.72]$	3/3	–		–	
mem0-8	–		$-0.31\,[-0.60, -0.05]$	3/3	–		–	
logistic	–		–		–		$0.63\,[0.54, 0.72]$	3/3
knn	–		–		–		$0.35\,[0.31, 0.39]$	3/3

Table 10: Hard split: the same pooling as Table 2 restricted to the three structured families (context, drift, pairwise), where the misspecified logistic reference itself stays far from the exact posterior. Post hoc descriptive split (Appendix F); it enters no test.

Condition	gpt-5.6-sol $A=8$: ρ [95% CI]	($A=6$)	gpt-6-astra $A=8$: ρ [95% CI]	($A=6$)	claude-sonnet-5 $A=8$: ρ [95% CI]	($A=6$)	statistical $A=8$: ρ [95% CI]	($A=6$)
pairwise								
no-history	-0.27 [-0.51, -0.09]	(-0.28)	0.00 [0.00, 0.00]	(0.00)	–		–	
full-history	-0.29 [-0.54, -0.03]	(0.00)	0.53 [0.37, 0.69]	(0.59)	–		–	
summary-512	-0.20 [-0.44, 0.07]	(0.03)	0.22 [0.00, 0.43]	(0.28)	–		–	
summary-2048	-0.13 [-0.35, 0.09]	(0.18)	0.23 [0.04, 0.41]	(0.37)	–		–	
profile-512	-0.13 [-0.63, 0.40]	(0.49)	0.16 [-0.12, 0.44]	(0.32)	–		–	
profile-2048	-0.30 [-0.60, -0.01]	(0.05)	0.16 [-0.07, 0.40]	(0.45)	–		–	
retrieval-8	-0.28 [-0.58, -0.01]	(-0.33)	-0.04 [-0.30, 0.19]	(0.10)	–		–	
schema-summary-1024	0.57 [0.41, 0.70]	(0.68)	0.85 [0.78, 0.91]	(0.89)	–		–	
hypothesis-summary-1024	0.04 [-0.17, 0.26]	(0.24)	0.42 [0.26, 0.59]	(0.50)	–		–	
extract-fit	0.82 [0.74, 0.90]	(0.78)	0.83 [0.76, 0.89]	(0.81)	–		–	
logistic	–		–		–		0.83 [0.76, 0.90]	(0.82)
knn	–		–		–		0.29 [0.24, 0.34]	(0.39)
context								
no-history	-0.14 [-0.27, -0.03]	(-0.29)	0.00 [0.00, 0.00]	(0.00)	–		–	
full-history	-0.21 [-0.44, -0.01]	(-0.34)	0.51 [0.35, 0.64]	(0.55)	–		–	
summary-512	-0.18 [-0.46, 0.08]	(-0.27)	0.23 [-0.01, 0.45]	(0.07)	–		–	
summary-2048	-0.12 [-0.34, 0.08]	(-0.30)	0.30 [0.11, 0.44]	(0.28)	–		–	
profile-512	0.02 [-0.25, 0.24]	(-0.61)	0.13 [-0.16, 0.40]	(-0.00)	–		–	
profile-2048	-0.29 [-0.57, -0.03]	(-0.46)	0.34 [0.12, 0.55]	(0.25)	–		–	
retrieval-8	-0.60 [-0.99, -0.26]	(-0.96)	-0.37 [-0.72, -0.05]	(-0.34)	–		–	
schema-summary-1024	0.11 [-0.16, 0.35]	(0.33)	0.87 [0.78, 0.95]	(0.76)	–		–	
hypothesis-summary-1024	0.03 [-0.25, 0.29]	(-0.09)	0.47 [0.30, 0.60]	(0.51)	–		–	
extract-fit	0.57 [0.45, 0.67]	(0.57)	0.57 [0.46, 0.68]	(0.60)	–		–	
logistic	–		–		–		0.57 [0.46, 0.68]	(0.61)
knn	–		–		–		0.29 [0.21, 0.37]	(0.48)

Table 11: Harder regime (Amendment 6): pairwise and context families at $A = 8$ (3,515,625 enumerated states), $N = 128$, track T1, real vocabulary; per-family ρ with 95% whole-user bootstrap intervals, the $A = 6$ value in parentheses. Exploratory.

Family	bayes-active				logistic-active
	random	uncertainty	information	risk	uncertainty
additive (A=4)	0.13	0.16 (+0.04)	0.17 (+0.05)	0.17 (+0.04)	0.14 (+0.12)
additive (A=6)	0.08	0.09 (+0.02)	0.14 (+0.06)	0.15 (+0.07)	0.05 (-0.06)
additive (A=8)	-0.02	0.20 (+0.22*)	0.15 (+0.17)	0.15 (+0.17)	0.17 (+0.06)
pairwise	0.31	0.10 (-0.22)	0.28 (-0.03)	0.34 (+0.02)	0.23 (+0.05)
context	0.09	0.13 (+0.04)	0.09 (+0.00)	0.10 (+0.01)	-0.01 (-0.18*)
hidden-triple	0.45	0.20 (-0.25)	0.62 (+0.17)	0.60 (+0.14)	0.07 (-0.11)
drift	-0.22	0.21 (+0.44*)	0.26 (+0.48*)	0.15 (+0.38)	0.38 (+0.17)
contradiction	-0.01	0.09 (+0.11*)	0.17 (+0.18*)	0.15 (+0.16*)	0.05 (+0.02)

Table 12: T2 bounded acquisition: reduction in user-mean expected Brier ($\times 100$) after 5 acquisitions relative to step 0, paired within user, with the paired difference against the same system's random policy at step 5 in parentheses (positive: lower expected Brier than random). * marks a parenthesised difference whose 95% user-bootstrap interval excludes zero; the gains relative to step 0 are not marked. 16 users per row, $N = 128$, real vocabulary. bayes-active reads the exact posterior; logistic-active selects by its own uncertainty. Drift users change at N // 2; targets, probes and candidates are all post-change, so acquisitions inform the post-change state.

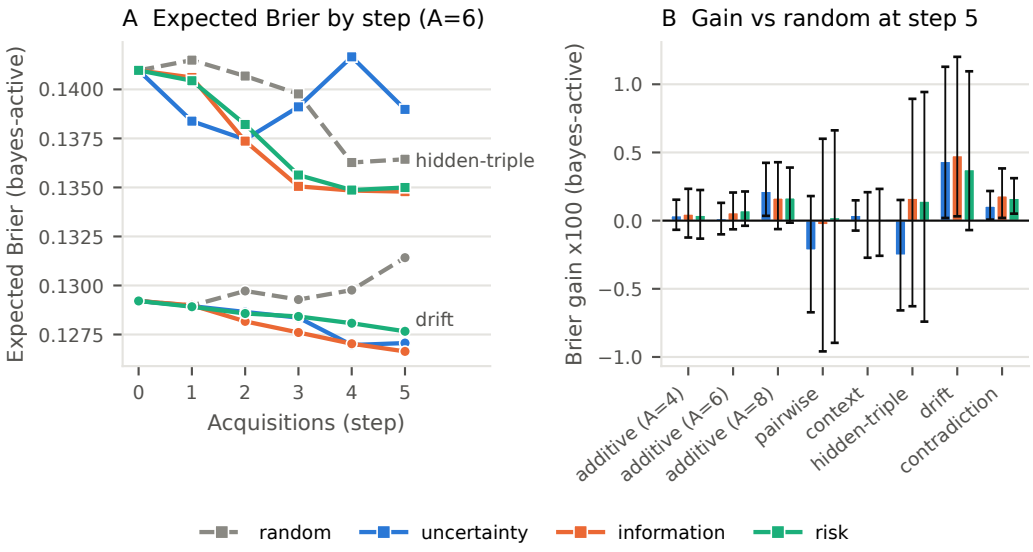


Figure 6: T2 bounded acquisition, exact-posterior reference (bayes-active), 16 users per family at $N = 128$. A: user-mean expected Brier by acquisition step at $A = 6$, one line per policy and family. B: paired gain against the random policy at step 5 ($\times 100$ expected-Brier units) for the uncertainty, information and risk policies, with descriptive 95% user-bootstrap intervals; positive is lower expected Brier than random.

Family	Full Bayes risk	Random-8 excess risk	Diverse-8 excess risk
Additive	0.1348	0.0675 [0.0411, 0.1010]	0.0695 [0.0543, 0.0895]
Context	0.1329	0.0820 [0.0639, 0.1054]	0.0704 [0.0536, 0.0908]
Equal-family pool	0.1338	0.0747 [0.0580, 0.0944]	0.0700 [0.0591, 0.0821]

Table 13: Post hoc exact selected-evidence reference, in raw Brier units. Full Bayes risk is $\mu_D(1-\mu_D)$; excess risk is $(\mu_S-\mu_D)^2$, conditioning both predictions on the full history for risk evaluation. Brackets are unadjusted 95% percentile intervals from 2,000 bootstrap resamples of 16 generating seeds, shared across families and controls; the pool weights families equally. The known-family reference receives gold selected observations. No LLM reader predictions or semantic-retrieval claims are included.

C AN EXACT REFERENCE FOR CONTROLLED SELECTED EVIDENCE

This post hoc offline diagnostic isolates evidence discarded by two outcome-independent selectors. On the existing additive and context cohorts ($A = 6$, $N = 128$, real vocabulary, sixteen shared generating seeds and 24 T1 targets per family), we compare the exact full-history forecaster μ_D with μ_S, the exact forecaster given eight selected observations. Random-8 uses a deterministic hash of public user and episode identifiers; diverse-8 greedily maximizes lexical Jaccard diversity after removing the observed-choice suffix. Neither selector reads outcomes, target labels or latent parameters. Both selectors are query independent. The evaluator uses the known family and gold observed features; these are statistical references, not language-model results.

Conditional on the public design X, a selector $S = f(X)$ contributes no additional parameter likelihood. Thus $\pi(\theta \mid X, S, Y_S) \propto \pi(\theta)\prod_{i\in S} p(Y_i \mid \theta, X_i)$. This relies on the generator's state-independent feature, identifier and rendering streams. The risk of the subset forecaster, evaluated under the *full-history* posterior, is

$$\mathbb{E}[(Y-\mu_S)^2 \mid D, X] = \mu_D(1-\mu_D) + (\mu_S-\mu_D)^2. \tag{5}$$

The second term is a nonnegative predictive loss from discarding evidence. It differs from $\mu_S(1-\mu_S) - \mu_D(1-\mu_D)$, which subtracts risks under different conditioning sets and can be negative for an individual history. It is also distinct from true-generator expected Brier loss and from mutual information.

Across the two equally weighted families, full-history posterior Bayes risk is 0.134. The subset-induced excess risk is 0.075 $[0.058, 0.094]$ for random-8 and 0.070 $[0.059, 0.082]$ for diverse-8 (Table 13). Intervals are descriptive percentiles from 2,000 synchronized resamples of the sixteen actual generating seeds. All 768 recomputed full-history predictions exactly reproduce their stored references. So even an ideal reader that knows the family loses predictive information when given only eight episodes. The diagnostic does not measure a language model's use of those episodes, compare equal byte budgets, or explain the existing semantic-retrieval result. In particular, semantic selection can depend on observed labels; treating only its returned episodes as the conditioning information would require modelling that selection process.

Evidence-count curve. We extend the same two selectors to nested subsets with $k \in \{1, 2, 4, 8, 16, 32, 64, 128\}$, holding the users and targets fixed (Figure 7). Pooled excess risk falls from $0.075/0.070$ at eight episodes to $0.028/0.028$ at 32 and $0.010/0.009$ at 64 (random/diverse); at 128 it is exactly zero by construction. The 1,536 target/control values at $k = 8$ exactly reproduce Table 13. These are descriptive finite-cohort curves, with no monotonicity imposed. The selected prose artifacts average about 3.6 kB at eight episodes, 14.5 kB at 32 and 57.9 kB for the full history, counting actual UTF-8 session text and excluding prompts and queries. Episode count determines selection; these sizes do not define matched byte budgets or an optimal compression frontier.

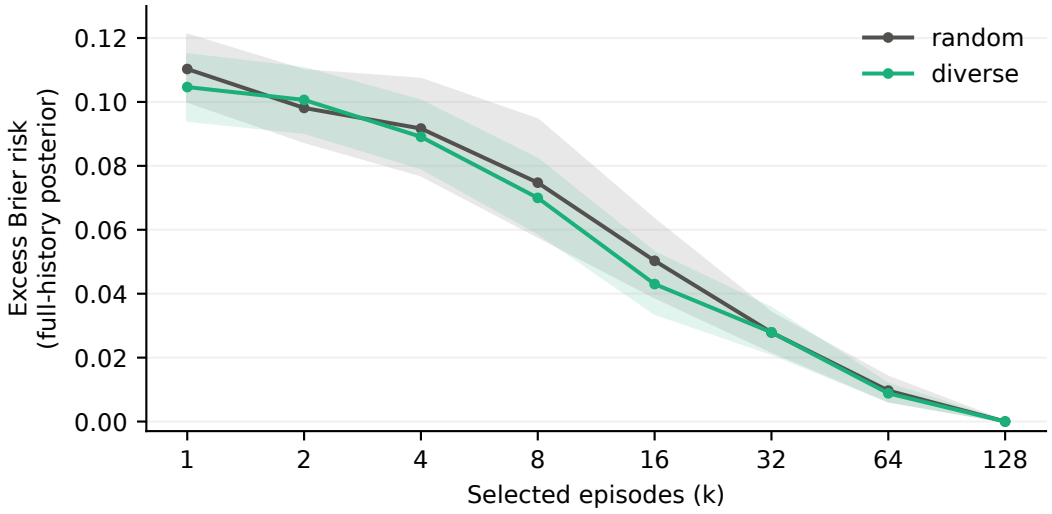


Figure 7: Exact predictive loss from retaining fewer observations, pooled with equal weight over additive and context. Each point averages 24 targets within each user and sixteen generating seeds within each family. Shading shows pointwise, unadjusted 95% intervals from 2,000 shared-seed bootstrap resamples, synchronized across both selectors and all evidence counts. These are known-family statistical forecasters, not LLM readers; the endpoint uses all 128 observations.

D ONE-STEP PREDICTIVE INFORMATION VALUE

For candidate probabilities r_θ under the current weights w, mean $\bar{r} = \sum_\theta w_\theta r_\theta$, and a public probe distribution $\mathcal{P}$ over target-support queries, the selector scores each candidate by

$$V(r) = \mathbb{E}_{z\sim\mathcal{P}}\left[\frac{\mathrm{Cov}_w(p_\theta(z), r_\theta)^2}{\bar{r}(1-\bar{r})}\right], \tag{6}$$

which is zero for deterministic candidate outcomes. For a fixed probe let $P = p_\theta(z)$, $R = r_\theta$, and $C = \mathrm{Cov}_w(P, R)$. On candidate answer one, the posterior probe mean changes by $C/\bar{r}$; on answer zero, by $-C/(1-\bar{r})$. The expected squared change is $C^2/[\bar{r}(1-\bar{r})]$. Conditional variance identifies this with the expected reduction in Bayes Brier risk at that probe. Averaging probes gives Equation 6. The argument conditions on the current history and a selection policy using only permitted observations; hidden-state-dependent selection would require modelling the selection information as well. The greedy policy is not an optimal five-step policy, and hypothesis information gain need not maximize predictive improvement.

T2 results (Section 6). T2 starts from the exact posterior at $N = 128$, spends five acquisitions from a public pool of 24 candidates and rescores T1 (16 users per family, statistical references only) with random, uncertainty, information-gain and one-step expected-Brier (Equation 6) policies (uncertainty for the logistic learner). The posterior is already concentrated (Table 12, Figure 6, Appendix B): five answers change expected Brier by well under 0.01 in most families. Informed policies beat random where the history leaves most unresolved, drift ($+0.44$ to $+0.48$, $\times 100$ units) and contradiction ($+0.11$ to $+0.18$), with intervals excluding zero; on hidden-triple no policy separates from random, and the logistic learner gains less.

E EARLIER PILOTS

Two earlier development harnesses precede the v3 design. Their generators, families and rendering differ from Section 3, and their retained numbers (Tables 14 and 15, Figures 8 and 9) are not evidence about the v3 families or about language models.

Schema-disclosed finite families. A finite-family pilot used four public attributes with sign-only coefficients $\theta_j \in \{-1, +1\}$, temperature 2, and four families: linear (16 states), pairwise with $2x_0x_1$ and $2x_2x_3$ (64), context with cx_0 and cx_1 (64), and a hidden triple $2x_0x_1x_2$ (32). Histories activated at most two attributes and targets at least three, with 32 users per family, prefixes of 16, 64 and 128 observations, 24 targets, and artifact caps of 256, 1,024 and 4,096 bytes; systems received structured features and the public family. The completed local run comprised 128 synthetic users and 483,840 of 497,664 intended prediction records, with 576 user/condition instances of measured memory infeasibility (13,824 targets) retained rather than imputed. At 128 observations, mean unresolved predictive uncertainty ranged from 0.0024 to 0.0041 in the three fully exposed families, versus 0.0209 in the hidden-triple family (descriptive 95% user-bootstrap interval $[0.0194, 0.0222]$); the hidden-triple posterior still improved true-generator expected Brier loss over the prior by 0.0174 ($[0.0124, 0.0230]$), because additive preferences remain learnable, while a MAP control committed to a hidden sign despite unchanged posterior odds. Its acquisition track compared random selection, predictive uncertainty, hypothesis information gain and Equation 6 from a 1,024-byte posterior artifact (Table 14 and Figure 8, which derive from hash-verified outputs of that run).

Additive linear diagnostic. The earliest schema-v1 pilot sampled $\beta \in \{-2, -1, 1, 2\}^d$, predicted $\mathrm{sigmoid}(\beta^\top x/2)$, and used two nonzero signed attributes in experience versus three at test, with $d \in \{4, 8, 12, 16\}$, 24 stationary and 24 independent drift users per dimension, 80 observations, 12 targets per track, and a 1,024-byte cap; drift flipped half the coefficients after observation $N/2$, its active heuristic used uncertain predictions with a single logistic update, and rendering checks preserved structured features. In that completed run, compact logistic expected-Brier improvement over the constant predictor was 0.0906 (descriptive 95% user-bootstrap interval $[0.0744, 0.1060]$) at $d = 4$ and 0.0742 ($[0.0627, 0.0860]$) at $d = 16$, and exemplar controls also improved on average in every dimension. The additive signal was recoverable; the pilot says nothing about nonlinear families or about retrieval.

Family	Unresolved	Posterior gain	Linear excess	kNN excess
linear	0.0024	0.0268	0.0048	0.0311
pairwise	0.0041	0.0483	0.0247	0.0444
context	0.0037	0.0367	0.0151	0.0327
hidden-triple	0.0209	0.0174	0.0038	0.0253

Table 14: Finite-family development controls at $N = 128$ and $B = 1024$ bytes. Unresolved is posterior target-probability variance. Posterior gain is true-generator expected Brier improvement over the prior. Excess is squared distance from history-conditioned Bayes, under the public family/prior. Full-history kNN has unrestricted storage. These are statistical controls, not frontier-model rankings.

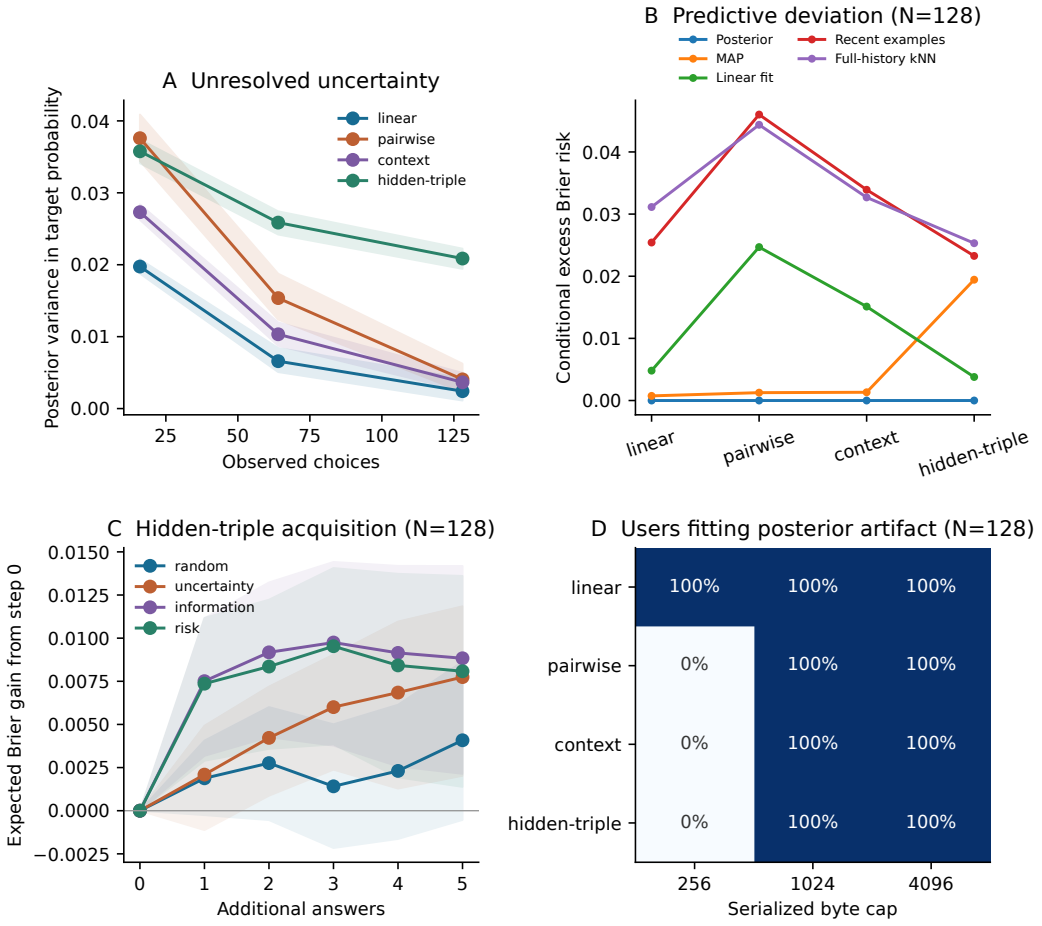


Figure 8: Finite-family development results: 32 users per family, with repeated prefixes and target predictions. Shading in A and C gives descriptive 95% user-bootstrap intervals (2,000 resamples); B shows means and D shows coverage. B–D use 128 observations. Posterior in B is the 1,024-byte quantized reference; full-history kNN is unrestricted. C uses the hidden-triple family and 1,024-byte memory, with paired true-generator expected Brier gains. Intervals are unadjusted for multiplicity.

F Prespecified Analysis Plan

The analysis plan was prespecified in the repository before the main matrix and hash-committed; no independent time-stamped registration exists. Fixed then: the six families and grids; $\tau = 1.0$, chosen on development seeds (48 users per cell, $N = 64$) as the temperature whose additive $A = 6$ oracle expected accuracy is nearest 0.8 subject to an exact-Bayes gain over the prior of at least

Control	Expected acc.	Expected Brier	Mean bytes
No experience	0.500	0.250	19
Recent exemplars	0.617	0.230	1014
Diverse exemplars	0.670	0.223	1013
Compact logistic	0.766	0.163	104
Recency-weighted logistic	0.704	0.202	105
Full-history logistic	0.766	0.163	104
Full-history nearest neighbor	0.658	0.216	3301
Permuted labels	0.505	0.276	105
Generator oracle (privileged)	0.788	0.146	0

Table 15: Synthetic development controls at $d = 8$, $N = 80$, and $B = 1024$ bytes. Full-history controls are unrestricted; the oracle knows the latent process. Full-history and compact logistic fits use the same estimator and coincide when its artifact fits the cap. Bytes denote mean serialized prediction-artifact size, not consolidation workspace. Oracle zero indicates no serialized artifact, not zero privileged information.

0.02 in every evaluated cell; the information-use margin of 0.01; the nine conditions with byte caps, reasoning efforts and completion caps; the three frozen prompts and the family-description template, with digests in every manifest; the no-retry failure policy; the estimand ρ on expected Brier with whole-user paired bootstrap intervals; and held-out seeds generated after freezing, separate from the development seeds.

Hypotheses. H1 (unused information): $\rho(\text{full-history}) < 1$ with the 95% upper bound below 0.9 in the pooled-family mean; refuted if the point estimate is ≥ 0.9 or the per-family upper bounds straddle 1 in a majority of the six families. H2 (compression loses signal): $\rho(\text{summary-2048}) < \rho(\text{full-history})$ as a paired difference on the same resampled users; refuted if the interval contains zero or lies above it. H3 (schema disclosure closes part of the gap): $\rho(\text{schema-summary-1024}) > \rho(\text{summary-2048})$, same construction. *Near-zero rule:* the phrase is permitted for a condition only if $\rho < 0.25$ and its upper bound is below 0.4, pooled and in every family that enters. *Holm family, per backbone:* the six paired comparisons against full-history (summary-512, summary-2048, profile-512, profile-2048, retrieval-8, schema-summary-1024) plus the H3 contrast, corrected together at $\alpha = 0.05$ (seven after Amendment 2 withdrew full-history-high). Everything else is descriptive and unadjusted.

Amendments. Amendment 1 (after calibration, before any main-matrix dispatch) raised the retrieval completion cap from 1,024 to 2,048 tokens; Amendment 2 (after the first two batches started) withdrew full-history-high on failure statistics alone and raised the transport timeout from 120 s to 300 s; Amendment 3 raised it to 600 s and added an operational resume mode. Amendment 4 (after the main matrix, before submission): paired contrasts involving retrieval-8 are restricted to the targets both conditions predicted (the first twelve of each track), since retrieval-8 predicts twelve targets per user and every other condition twenty-four; the pre-amendment values on unmatched targets are retained in the report. Local hashes and commitments support integrity checks; they do not substitute for independent time-stamped registration.

Amendment 5 (exploratory extensions; written before any extension call). Added on the primary cell only (T1, real vocabulary, $A = 6$, $N = 128$, the same users, the frozen prediction prompt): hypothesis-summary-1024 and extract-fit on both backbones; the sign-skewed additive family under all eight main conditions on both backbones, with the prior-predictive forecaster as an extra reference; Mem0 and Letta on `gpt-6-astra`; and a third backbone from another provider, which was prespecified with its own confirmatory block and was run as Claude Sonnet 5 through its provider's API on a reduced matrix (notes on the third backbone below). None of these enters the Holm families; intervals are descriptive. Expectations stated so that they could fail: E1, both systems fall between retrieval-8 and full-history: the matched contrasts do not establish this ordering. Mem0 minus retrieval-8 is $+0.006$ $[-0.146, 0.174]$; Letta minus full-history is -0.015 $[-0.076, 0.049]$ (Table 8). E2, hypothesis-summary-1024 exceeds summary-2048 and approaches schema-summary-1024: the first half holds on both backbones (0.14 and 0.51 against -0.07 and 0.39; differences of pooled values $+0.21$ and $+0.12$, descriptive), the second does not (0.55 and 0.82). E3, extract-fit approaches the logistic reference: met on both backbones (0.76 and 0.78 against 0.79, all six families; the paired

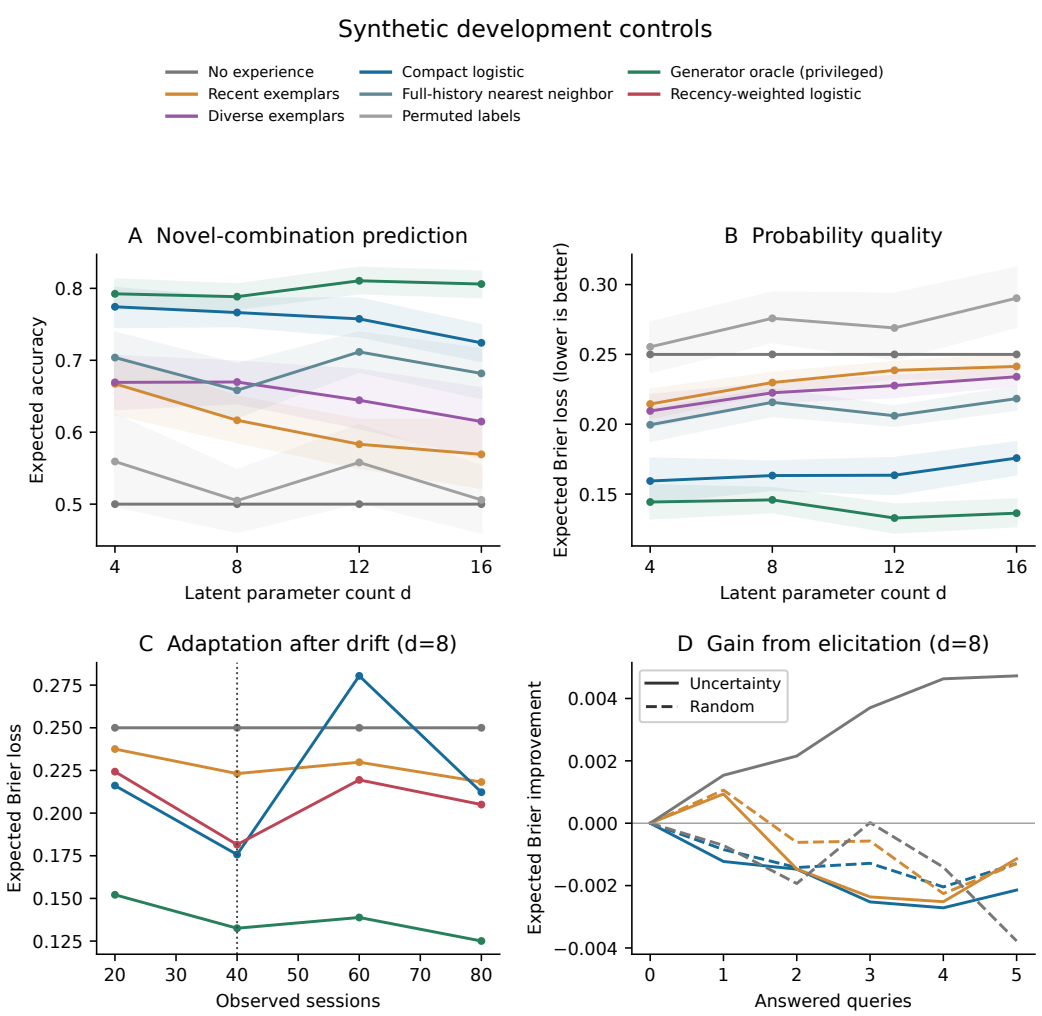


Figure 9: Earlier schema-v1 additive diagnostic, not the v3 experiment. Panels A–B show descriptive user-bootstrap intervals; C–D show mean trajectories at $d = 8$.

residual gap is reported in Table 8, rather than inferred from overlapping intervals). E4, full-history ρ is higher under the skewed prior on both backbones: met on sol ($+0.45$ $[0.11, 0.76]$), positive with an interval covering zero on astra ($+0.13$ $[-0.02, 0.29]$). E5 (H1 on the third backbone): met, trivially; pooled full-history $\rho = -0.007$ $[-0.023, 0.008]$ on `claude-sonnet-5`. *Dated notes, each written before the affected dispatch or from failure statistics alone.* (i) Mem0's extraction prompt of about 8,000 tokens per call made one call per session unaffordable, so both systems ingest eight sessions per call and run on the first eight users of each family, with the coverage rule applied to eight. (ii) Letta: the final open-source server release 0.16.8 is pinned by image id with the default agent, prompt and tools; model calls use the Azure Responses endpoint at low effort because chat completions reject function tools for the reasoning deployments; the default 30,000-token context window is kept; persistent bytes are core memory plus the in-context message buffer. (iii) Transport-retry rule: an internal call of either system that fails in transport is retried once by the metering proxy, both attempts metered; a second failure fails the user; the benchmark's own calls are never retried; completed users are kept and attempts are merged with per-user provenance. (iv) Extract-fit rerun rule: a cell whose only failure is `call-failed` is rerun once under the same frozen config; parsing or fitting failures stay failed. (v) The per-user caps stay frozen (USD 8 Mem0, USD 12 Letta, in conservative ledger terms): three Letta users whose agents used memory tools made 32 model calls instead of about 17, reached the cap without any transport failure, and are reported as beyond the cap, not rerun; six infrastructure failures are rerun once.

Third backbone (dated notes of 2026-09-18 and 2026-09-19). The cloud route prespecified for the third backbone refused every request on the provider account, and Claude Opus 5 was not available otherwise. By author decision, recorded before dispatch, the third backbone is Claude Sonnet 5 (`claude-sonnet-5`) called directly through its provider's API on the approved reduced matrix: seven conditions (no-history, full-history, summary-512 and summary-2048, profile-512 and profile-2048, schema-summary-1024; no retrieval), T1 for all and T4 for no-history, full-history, summary-2048 and profile-2048, six families, the same sixteen users, $A = 6$, $N = 128$, effort low, 600 s timeout; only unbilled provider refusals (rate limit, overload) are retried, up to three times. Two measures apply to this backbone only, decided from parse status alone: prediction calls use the provider's structured-output feature with one required number per query id, converted to the list format the frozen prompt requests before the unchanged strict parser runs (the raw text is logged), and prediction and consolidation outputs are capped at 8,192 tokens, because its thinking counts toward the cap. Without them the model mis-copied the 20-character target ids in about a third of prediction calls, emitted invalid JSON, exhausted the 2,048-token cap on 8 of 30 calls or, with ids constrained but no key required, omitted one of 24 predictions in about one call in five. These two stopped attempts (30 and 68 prediction calls), and the earlier cloud launches that produced only refused calls, are not reported. Coverage is given in Appendix B. Outcome: H1 holds (E5 above); none of the four available Holm comparisons rejects (summary-512, summary-2048 and schema-summary-1024 against full history, and H3; profile-512, profile-2048 and retrieval-8 retain $p = 1$), and summary-512, summary-2048 and schema-summary-1024 meet the near-zero rule. Its full-history forecasts lie within 0.05 of $1/2$ for 82% of targets (sd 0.05), as in the first attempt without structured output (96%, 24 targets), so the flat result is not produced by the output constraint; with the 2,048-byte summary its forecasts spread (sd 0.13) and correlate 0.24 with the exact posterior, yet pooled ρ stays at 0.02. Retrieval-8 and hypothesis-summary-1024 were added later on this backbone (author-approved; interrupted runs completed with the harness resume mode, request ids unchanged): retrieval-8 pools to -0.07 $[-0.14, 0.02]$, not different from full history (-0.058 $[-0.142, 0.034]$, Holm $p = 1$); hypothesis summaries exceed 1,024 bytes for most users, so no family meets the coverage rule. Extract-fit was not run on this backbone.

Amendment 6. (a) The pool over the three structured families (context, drift, pairwise; Table 10) is a post hoc descriptive split, chosen after the draft combined report showed the other three families nearly saturated by parse-then-regress; it is computed like the six-family pool, enters no test and replaces no prespecified number. (b) A harder regime was prespecified before dispatch: pairwise and context at $A = 8$ (3,515,625 states each), $N = 128$, T1, the same sixteen seeds and both backbones, under the eight main conditions plus hypothesis-summary-1024 and extract-fit; expectation E6: the logistic reference and extract-fit fall further below the exact reference than at $A = 6$, and astra's full-history ρ is lower than at $A = 6$. Outcome (Table 11): E6 is not met for pairwise (logistic 0.83 at $A = 8$ against 0.82 at $A = 6$; extract-fit 0.82 and 0.83 against 0.78 and 0.81) and only weakly for context (logistic 0.57 against 0.61; astra's full-history 0.51 against 0.55, intervals overlapping): these point estimates do not support a uniform degradation from the 25-fold larger space. Interval overlap is not an equivalence test. Coverage: of 288 user-condition instances per text batch, 247 (pairwise) and 237 (context) completed, 22 and 28 failed (all but one on request transport), 16 each were budget-infeasible (14 and 13 profile-512, the rest profile-2048) and 3 and 7 were partial (retrieval-8); extract-fit completed 30 and 32 of 32 after the transport rerun. *Amendment 7 (post-review, 2026-09-18).* Pooled intervals now resample actual generating seeds jointly across families, replacing incorrect coupling by positions in sorted opaque-ID arrays. The original cell eligibility and point estimands are preserved; unavailable Holm hypotheses retain $p = 1$, and two-sided bootstrap tails use a plus-one correction with 2,000 resamples, so the resulting p-values are approximate, not exact randomization tests. Secondary comparisons additionally match committed query features and context. New paired pipeline contrasts, the posterior decomposition and actual-prior sensitivity are descriptive post hoc analyses. Historical predictions and the original report are retained. The affordability polarity and drift-prefix descriptions are corrected without changing datasets. No matched schema/representation or random/diverse language-model control results are reported here. The additional post hoc exact-reference diagnostic for outcome-independent selected evidence appears in Appendix C; it contains no new model calls.