HABIT-BENCH: EVALUATING HABIT INDUCTION FROM LONGITUDINAL WEAK EVIDENCE IN AGENT MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

In long-term interactions with large language model (LLM) agents, users reveal habits through their choices and actions. These habits capture implicit user information that agents must infer across conversations. However, individual observations may provide only partial evidence about a habit or when it applies. Retaining past interactions alone does not ensure that agents can infer habits and apply them appropriately. We introduce HABIT-Bench to evaluate how well agent memory systems infer habits from weak evidence and apply them to later requests. We construct synthetic user histories from task-oriented dialogue data and controlled habit graphs. The histories span hundreds of sessions across food, finance, software, and travel. Four-choice tasks assess agents' ability to infer user habits and apply them appropriately. They include temporary exceptions, changes in habits, and suggestions that users have not accepted. Sampled tasks undergo independent human review. We evaluate eight memory methods with Qwen3-235B-A22B and GLM-5.3-Flash. The best-performing memory method achieves 33.61% macro accuracy across the four domains. Performance remains limited even when annotated evidence is supplied directly. These results expose a persistent gap between access to past interactions and the ability to infer and appropriately use user habits.

1 INTRODUCTION

Long-term interaction requires large language model (LLM) agents to learn from experience, not merely retain explicit facts and preferences (Shinn et al., 2023; Xu et al., 2026). Associative and latent memories offer complementary mechanisms for organizing and retaining experience (Gutiérrez et al., 2024; Wang et al., 2024; 2025). Useful personalization depends on inferring user characteristics and preferences from interaction history (Zhong et al., 2024; Zollo et al., 2025; Nam et al., 2026) and deciding when these regularities should guide a response. Context and subsequent changes can alter whether an otherwise useful default applies (Wood & Rünger, 2016; Jiang et al., 2025).

Existing benchmarks address complementary parts of this problem. LoCoMo and LongMemEval assess recall and reasoning over extended histories (Maharana et al., 2024; Wu et al., 2024), while MemoryAgentBench adds incremental learning and selective forgetting (Hu et al., 2026a). PrefEval examines implicit preference inference and following, while PersonaMem tracks evolving user profiles (Zhao et al., 2025; Jiang et al., 2025). PERMA and DynamicMem study evolving preferences and habits (Liu et al., 2026; Xie et al., 2026). BenchPreS and OP-Bench examine when personalization should be withheld (Yoon et al., 2026; Hu et al., 2026b), while MemoryArena and Mem2ActBench connect memory to subsequent action (He et al., 2026; Shen et al., 2026). These developments make a precise evaluation question important: when does the available history justify applying a user-specific behavioral default to a new decision?

We introduce HABIT-Bench to study this question. We define habit memory as the ability to infer behavioral defaults from incomplete evidence across sessions and apply them according to their scope, exceptions, temporal validity, and user endorsement. A relevant past interaction need not establish a valid default. A repeated choice may apply only in one context, and an assistant's suggestion does not become a user preference merely because it appears in the history. The target is the

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.

Table 1: Capability coverage across representative long-term-memory and personalization benchmarks. ✓, ○, and ×denote dedicated, partial, and absent benchmark-level coverage, respectively. Columns cover longitudinal history, weak-evidence induction, memory-guided action, contextual gating, exceptions, temporal drift, and withholding unsupported personalization.

Benchmark	Long Hist.	Weak Evid.	Action	Context Gate	Exception	Temp. Drift	Withhold
LoCoMo (Maharana et al., 2024)	✓	×	×	×	×	○	×
LongMemEval (Wu et al., 2024)	✓	×	×	×	×	✓	×
MemoryAgentBench (Hu et al., 2026a)	✓	○	○	×	×	✓	×
MemoryArena (He et al., 2026)	✓	×	✓	○	×	×	×
Mem2ActBench (Shen et al., 2026)	✓	×	✓	○	×	✓	×
PERMA (Liu et al., 2026)	✓	✓	✓	○	○	✓	×
DynamicMem (Xie et al., 2026)	✓	✓	✓	○	○	✓	×
HorizonBench (Li et al., 2026)	✓	○	×	×	×	✓	×
IBA-Bench (Song et al., 2026)	✓	✓	✓	○	○	✓	○
BenchPreS (Yoon et al., 2026)	○	×	✓	✓	×	×	✓
OP-Bench (Hu et al., 2026b)	✓	×	✓	✓	×	×	✓
MemOps (Hao et al., 2026)	✓	×	○	○	○	✓	○
PersonaTrail (Yang et al., 2026)	✓	✓	✓	○	×	×	×
PersonalAlign (Lyu et al., 2026)	✓	✓	✓	✓	○	○	○
RefMem-Bench (Lin et al., 2026)	✓	✓	○	○	×	✓	○
HABIT-Bench	✓	✓	✓	✓	✓	✓	✓

support that memory provides for a justified decision, rather than a particular memory representation (Rabinowitz et al., 2018; Siththaranjan et al., 2024).

HABIT-Bench constructs synthetic histories using task-oriented dialogue seeds and controlled habit graphs. It contains 114 users, 37,152 sessions, and 3,328 probes across food, finance, software, and travel. The domains cover complementary evidence structures. Food combines repeated choices with contextual contrasts, and Travel tests behavioral-rule recovery and its application to new requests. Finance and Software emphasize reconstructing user-confirmed workflows and tracking their changes across sessions. Candidate actions differ in their use of habits, including applying an outdated choice, extending a default beyond its scope, or treating an unaccepted suggestion as an established decision.

Memory methods process each user's history chronologically and supply context to a shared base model. We evaluate eight memory methods alongside retrieval baselines and a compressed-history control. The best memory-method macro accuracy is 33.06% with Qwen3-235B-A22B (Yang et al., 2025) and 33.61% with GLM-5.3-Flash (Zeng et al., 2026). These results are close to the best retrieval baselines, while GLM's compressed-history control reaches 41.98%. Oracle Evidence improves access to relevant observations but still leaves substantial decision errors. The results identify both the reconstruction of habits and their appropriate use as important targets for memory system evaluation.

Our contributions are as follows.

Capability definition. We formulate habit memory as the joint problem of inferring behavioral defaults and determining when the evidence supports their use.

Benchmark construction. We introduce a four-domain benchmark with chronological histories, annotated habit states and supporting sessions, and human validation of decision recoverability.

Evaluation and diagnostics. We compare memory systems under a shared answer protocol and provide diagnostic controls for evidence access and downstream decision making. The experiments show that elaborate memory pipelines do not consistently outperform simple retrieval on these tasks.

2 Related Work

Long-term memory evaluation. LoCoMo evaluates long-conversation question answering, summarization, and multimodal generation (Maharana et al., 2024), while LongMemEval tests extraction, multi-session and temporal reasoning, updates, and abstention (Wu et al., 2024). MemoryAgentBench examines retrieval, learning, long-range understanding, and selective forgetting under

incremental input (Hu et al., 2026a), and MemOps diagnoses memory operations and their composition (Hao et al., 2026).

Implicit and evolving user modeling. PERMA studies preference formation and refinement (Liu et al., 2026), while DynamicMem evaluates evolving attributes, habits, and preferences across applications (Xie et al., 2026). HorizonBench constructs graph-controlled preference trajectories with traceable provenance (Li et al., 2026). RefMem-Bench tests behavioral pattern induction, temporal dynamics, and implicit boundaries from fragmented multimodal evidence (Lin et al., 2026). HABIT-Bench focuses on whether longitudinal evidence supports a behavioral default and its subsequent application.

Selective personalization. BenchPreS tests preference application under social and institutional constraints in third-party communication (Yoon et al., 2026). OP-Bench examines irrelevant memory use, repetition, and sycophancy (Hu et al., 2026b). HABIT-Bench couples selectivity with weak-evidence induction, testing scope, exceptions, temporal validity, and user endorsement.

Memory-grounded agent behavior. MemoryArena evaluates experience reuse across interdependent sessions (He et al., 2026), and Mem2ActBench tests memory-dependent tool selection and parameter grounding (Shen et al., 2026). PersonaTrail studies personalized web navigation (Yang et al., 2026), while PersonalAlign's AndroidIntent evaluates intent resolution and proactive GUI assistance (Lyu et al., 2026). IBA-Bench connects implicit user modeling to task execution under enduring traits and transient states (Song et al., 2026). HABIT-Bench uses non-interactive action selection with a shared answer LLM to compare the context produced by different memory systems. Table 1 compares these evaluation targets.

3 HABIT-BENCH

HABIT-Bench uses four-choice tasks to evaluate whether agents can infer user habits and use them appropriately to select actions. Figure 1 outlines the benchmark, and Appendix A.6 illustrates how distributed evidence determines the valid action in each domain.

3.1 TASK FORMULATION

For user u, let $H_u^{\leq t} = (s_1, \ldots, s_t)$ denote the history through session t. A probe p contains a request x_p, four candidate actions A_p, a visibility cutoff t_p, and a target time τ_p specifying the current or historical state to resolve. Each habit state $z_{u,h}(\tau)$ describes a behavioral default, its applicability, exceptions, temporal validity, and user endorsement. The evaluator specifies

$$Z_u(\tau) = \{z_{u,h}(\tau) : h \in \mathcal{B}_u\}, \qquad a_p^\star = \mathrm{Resolve}(x_p, A_p, Z_u(\tau_p)), \tag{1}$$

where $\mathcal{B}_u$ indexes the user's habits. Resolve combines relevant habits and enforces the request's constraints to identify one valid action. This specification leaves the evaluated method's memory representation unrestricted.

We define weak evidence as observations that provide incomplete support for a default or its validity conditions. A recurring choice can leave scope uncertain, and a later selection can require an earlier shortlist for interpretation. The task therefore combines behavioral inference and cross-session decision reconstruction. In the Finance example (Appendix Figure 7), two workflows must be reconstructed at a May 2023 target time even though their 2025 replacements are visible. Correct memory use requires preserving both the chosen actions and when they became effective.

3.2 DATA AND TASK CONSTRUCTION

HABIT-Bench contains 114 synthetic users, 37,152 sessions, and 3,328 probes across four domains (Table 2). Source datasets provide task seeds and interaction patterns. HABIT-Bench constructs user continuity, habit states, and longitudinal evidence relationships. Histories interleave habit-relevant observations with ordinary same-domain tasks, while private annotations link probes to applicable states and supporting sessions. The four domains instantiate complementary evidence structures.

Food. We use cooking dialogues and recipes from CookDial (Jiang et al., 2023) to create conversational evidence of preparation and serving habits. Earlier alternatives, later user selections,

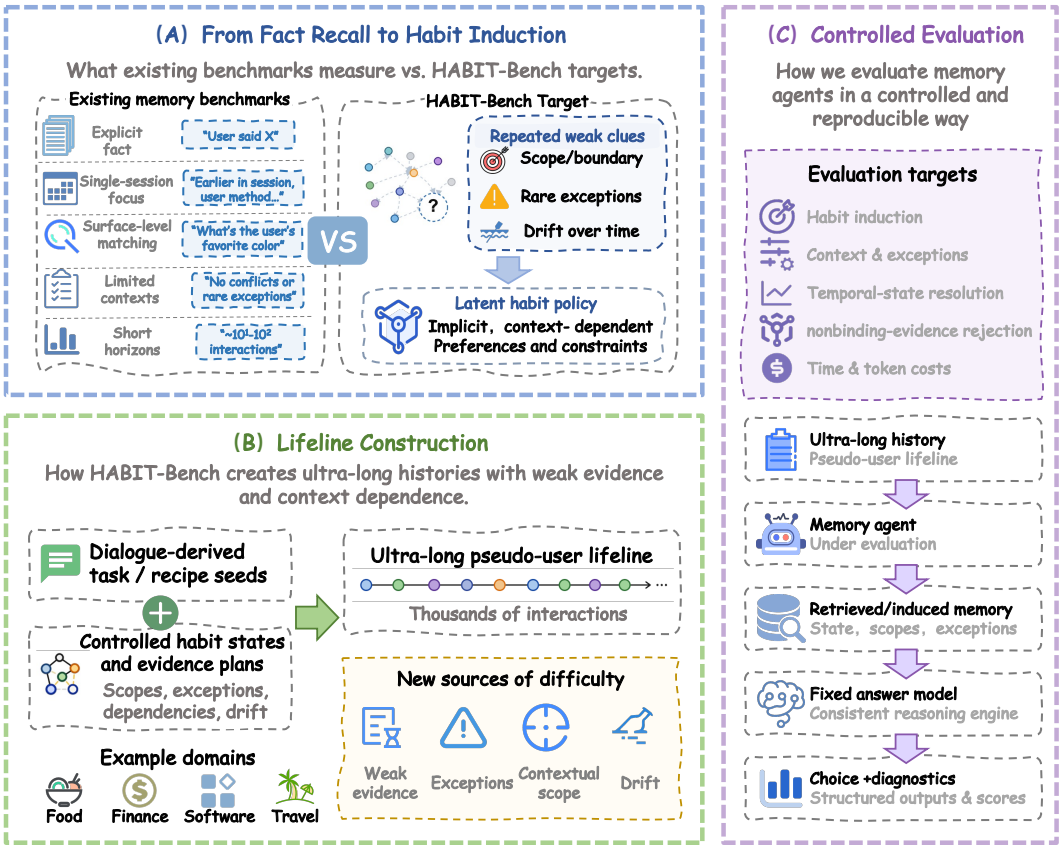


Figure 1: HABIT-Bench framework. (A) contrasts explicit-memory retrieval with latent habit induction. (B) combines task-oriented prompt seeds with latent habit graphs to construct real-prompt-seeded synthetic lifelines. (C) evaluates memory systems through a shared answerer and diagnostic choice interface.

Table 2: Dataset statistics and source datasets. Sources provide dialogue and task seeds, while HABIT-Bench constructs the longitudinal histories and probe labels.

Domain	Users	Sessions	Probes	Source dataset
Food	30	3900	630	CookDial (Jiang et al., 2023)
Finance	36	19440	1368	MultiDoGO (Peskov et al., 2019)
Software	18	9720	680	MultiDoGO (Peskov et al., 2019)
Travel	30	4092	650	Taskmaster-2 (Byrne et al., 2019)

and repeated use establish behavioral defaults. Contrasting contexts and local exceptions determine their scope. We rewrite the sessions as natural conversations while preserving recipe facts and habit evidence. Direct-use, boundary, and exception probes distinguish applying a default from overgeneralizing it.

Finance. We use banking dialogues from MultiDoGO (Peskov et al., 2019) to construct histories of financial tasks. Ordered shortlists and later user confirmations establish or replace standing choices. Two- and three-habit probes combine workflows at current or historical times, with local exceptions, unresolved cases, and unaccepted suggestions providing competing evidence.

Software. The Software histories use technical-support dialogues from MultiDoGO (Peskov et al., 2019). Each user has a consistent software environment. We use the same selection process as Finance, with more frequent changes to the chosen procedures. The histories also include similar errors in different environments and unfinished support cases. Questions combine two or three habits and may require continuing an earlier case.

Travel. Taskmaster-2 (Byrne et al., 2019) flight and hotel dialogues supply behavioral candidates for synthetic user profiles. Chronological histories are authored before evidence annotation and blind recoverability checks. Questions involve one or two habits and test their application to new requests, including exceptions and changes over time.

Choices and validation. Alternatives are designed to meet the request's hard constraints while differing in habit use and validity. We check that the choices cover the request, the evidence supports the intended decision, and the context is consistent. Domain-specific structural checks, model-based checks, and human review also assess whether a choice can be identified without using the history. Appendix A.1 details templates, evidence schedules, source transformations, and validation. The resulting lifelines are source-grounded synthetic histories, not longitudinal records of the source participants.

3.3 Evaluation Protocol and Metrics

Methods ingest sessions chronologically into user-isolated memory, $M_u^t = \mathrm{Update}(M_u^{t-1}, s_t)$. At probe p, they construct $R_p = \mathrm{Retrieve}(M_u^{t_p}, x_p)$ under recorded method-specific budgets. A shared answerer within each comparison predicts $\hat{a}_p = \mathrm{Answer}(x_p, \mathcal{A}_p, R_p)$. Standard methods cannot access private gold choices, habit states, evidence annotations, future sessions, or other users' memories. Appendix A.3 documents their adapters.

For the evaluated probe set P_d in domain d, the primary scores are

$$\mathrm{Acc}_d = \frac{1}{|P_d|} \sum_{p \in P_d} \mathbb{I}[\hat{a}_p = a_p^\star], \qquad \mathrm{Macro} = \frac{1}{4} \sum_{d=1}^{4} \mathrm{Acc}_d \,. \tag{2}$$

Exact-match scoring of the four-choice predictions avoids variability introduced by LLM judges and does not require a separate, stronger judge model. Uniform random choice yields 25%. Appendix A.5 reports supplementary user-macro scores and available 95% user-cluster bootstrap intervals. Evidence-access and answer-evidence diagnostics require source-session attribution and eligible annotations. Missing attribution is not counted as retrieval failure. Appendices A.4 and A.5.5 report capability slices and evidence metrics, respectively. Section 4.5 reports build/query time and LLM token usage. Boundary and nonbinding-evidence diagnostics do not constitute probability calibration or a global false-personalization rate.

3.4 Human Validation

Human validation tests whether the intended decision is recoverable from the annotated evidence, independently of a memory system's ability to find that evidence in a long history. Two human reviewers, A and B, independently inspect the request, four alternatives, and a chronological evidence packet without seeing the gold answer or system predictions. They select the uniquely supported action and assess eight criteria covering answerability, evidence sufficiency, scope, boundaries and exceptions, choice balance, naturalness, source grounding, and privacy. An item without a uniquely defensible answer receives no choice rather than a forced guess. A third human reviewer, C, adjudicates disagreements in either the answer or quality judgments before the merged record is frozen and compared with the gold labels. Appendix A.2 details the protocol, sampling, and domain-specific results.

Independent review covers 200 Food, 340 Finance, 311 Software, and 618 Travel probes. A/B answer agreement over jointly valid choices is 99.50%, 100.00%, 99.68%, and 92.84%, respectively, with domain-specific denominators and Cohen's κ reported in Appendix Table 6. The quality rubric is stricter than agreement with gold: four Finance items initially receive Modify because their alternatives omit an additional output-placement requirement, although the evidence supports the intended habit. These local issues were corrected before final evaluation, while the audit retains the original modification flags. The remaining 336 Finance items and all 311 Software items pass adjudication. All 79 content-revised Travel items pass targeted gold-blind validation, and all 200 reviewed Food items pass naturalness checks after naturalization. High independent agreement and the rubric's ability to detect narrow construction defects support the quality of the reviewed evidence and decision labels. These checks assess decision recoverability given the evidence, not human retrieval from full histories.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Systems and models. We organize the comparison by the role of memory in each system: controls, retrieval baselines, memory methods, and multi-store memory methods. The No Memory and Full Memory controls omit the interaction history or provide a compressed history, respectively. The retrieval baselines select sessions directly from the history and comprise Recency-5, Recency-10, BM25 RAG, Dense RAG, and Temporal Hybrid RAG. The memory methods construct and maintain their own memory representations: A-MEM (Xu et al., 2026), Mem0 (Chhikara et al., 2025), MemOS (Li et al., 2025), MemRL (Zhang et al., 2026), LightMem (Fang et al., 2026), and SeCom (Pan et al., 2025). Letta (Packer et al., 2023) and MIRIX (Wang & Chen, 2025) are shown separately as multi-store memory methods, reflecting their coordination of distinct memory stores. The other six are grouped for presentation rather than treated as a single architecture. References to the eight memory methods include both groups. Qwen3-235B-A22B (Yang et al., 2025) and GLM-5.3-Flash (Zeng et al., 2026) are the two primary language models. Within each setting, all methods use the same model to select an answer and to perform memory operations that require an LLM. Dense retrieval components use a fixed BGE-M3 encoder (Chen et al., 2025). Together with additional controls, oracle diagnostics, and experiments with smaller Qwen3 models, the evaluation covers 308 configurations across methods, models, and domains.

Inference protocol. Methods maintain a separate memory for each user and process sessions in chronological order, using only the history available at each probe's visibility cutoff. We set the decoding temperature to zero. The two primary model settings use the same prompt for selecting an action and require outputs to follow a fixed JSON schema. In the GLM experiments, the Qwen3-8B tokenizer is used only to measure token budgets, not to generate memories or answers. Context construction and token budgets vary by method. We report accuracy over all probes within each domain and macro accuracy as the unweighted mean of the four domain accuracies. Appendix A.5 reports supplementary accuracy averaged across users and, where available, bootstrap confidence intervals obtained by resampling users. It also provides the model configurations, decision prompts, token budgets, and rules for truncation, retries, and answer parsing.

Baselines and oracle controls. Retrieval baselines use the request alone, without the answer alternatives, to select complete sessions from the visible history. No Memory omits the interaction history, while Full Memory supplies a compacted history that fits within its context budget. Oracle Evidence supplies the private annotated evidence, whereas Oracle Habit State supplies the private habit state and required action. Both oracle conditions use information unavailable to the evaluated methods and serve only as diagnostic controls.

4.2 MAIN RESULTS

Table 3 evaluates whether current memory systems can infer habits from longitudinal weak evidence and apply them to subsequent decisions. Three findings emerge across the two primary model settings.

Habit induction and action selection remain challenging. Across the eight memory methods, four-domain macro accuracy ranges from 27.01% to 33.06% with Qwen3-235B-A22B and from 30.42% to 33.61% with GLM-5.3-Flash. Even the best configurations achieve only approximately one-third accuracy, despite using the primary model for both memory operations and action selection. The task requires selecting one of four actions and is scored by exact match, without an LLM judge. The model need not formulate a free-form response, but must still distinguish actions supported by the user's history from plausible yet unsupported alternatives. Habit-conditioned decision making remains a demanding evaluation target across the tested memory designs and model settings.

The benefit of memory depends on the domain. Under Qwen3-235B-A22B, the best memory-method results improve over No Memory by 20.32 percentage points on Food and 13.07 points on Travel. In Finance, all eight methods fall below the 23.54% No Memory point estimate. In Software, the largest gain is 1.03 percentage points. The domain contrast also appears with GLM: the best memory-method accuracies are 28.29% on Finance and 25.44% on Software, compared with 54.62% on Travel. Finance and Software require agents to combine multiple habits, resolve current or historical choices, and distinguish user commitments from unaccepted suggestions. Their

Table 3: Main results with Qwen3-235B-A22B and GLM-5.3-Flash: exact-choice accuracy (%). Macro is the unweighted four-domain mean. Controls measure performance without memory or with compressed history. Retrieval baselines select sessions, while memory methods construct and maintain memory representations. Letta and MIRIX are shown separately as multi-store memory methods. Gray averages weight methods equally. All averages excludes the No Memory and Full Memory controls. TH-RAG denotes Temporal Hybrid RAG.

Method	Qwen3-235B-A22B					GLM-5.3-Flash				
	Food	Finance	Software	Travel	Macro	Food	Finance	Software	Travel	Macro
Controls										
No Memory	22.38	23.54	24.85	28.31	24.77	23.81	22.95	22.50	32.46	25.43
Full Memory	48.25	20.54	21.76	33.38	30.99	45.56	33.77	37.50	51.08	41.98
Retrieval baselines										
Recency-5	24.13	22.51	24.26	31.23	25.53	18.10	23.46	23.53	39.69	26.20
Recency-10	23.02	21.71	23.68	31.38	24.95	18.25	25.88	22.94	41.85	27.23
BM25	45.08	23.03	22.50	39.38	32.50	31.90	27.19	24.12	48.62	32.96
Dense	39.84	22.08	23.53	39.85	31.32	30.32	24.85	24.12	53.54	33.21
TH-RAG	45.24	22.37	22.79	38.00	32.10	28.89	26.90	24.12	49.54	32.36
Retr. avg.	**35.46**	**22.34**	**23.35**	**35.97**	**29.28**	**25.49**	**25.66**	**23.77**	**46.65**	**30.39**
Memory methods										
A-MEM	42.70	22.44	25.74	41.38	33.06	27.94	28.29	25.15	53.08	33.61
Mem0	32.70	22.95	25.29	33.08	28.51	34.76	25.51	25.44	46.62	33.08
MemOS	40.32	21.86	21.47	34.77	29.60	36.67	24.05	24.56	47.23	33.13
MemRL	42.70	23.32	23.24	38.15	31.85	27.46	24.93	23.82	49.54	31.44
LightMem	28.25	22.81	25.88	31.08	27.01	33.33	23.68	21.18	44.31	30.63
SeCom	41.11	22.08	22.94	38.31	31.11	28.41	24.71	22.50	52.00	31.91
Mem. avg.	**37.96**	**22.58**	**24.09**	**36.13**	**30.19**	**31.43**	**25.20**	**23.78**	**48.80**	**32.30**
Multi-store memory										
Letta	41.59	22.73	23.24	40.92	32.12	27.14	25.73	24.12	54.62	32.90
MIRIX	26.98	22.73	25.59	34.62	27.48	27.14	24.93	21.76	47.85	30.42
Multi. avg.	**34.29**	**22.73**	**24.42**	**37.77**	**29.80**	**27.14**	**25.33**	**22.94**	**51.24**	**31.66**
All avg.	**36.44**	**22.51**	**23.86**	**36.32**	**29.78**	**28.49**	**25.39**	**23.64**	**48.35**	**31.47**

low scores show that the evaluated methods struggle with these combined requirements, but domain comparisons alone cannot identify which requirement causes an error. Memory can substantially improve decisions in some domains without providing reliable benefits in others. Section 4.3 examines this distinction through evidence and capability diagnostics.

Memory methods do not consistently outperform retrieval. The best memory-method macro accuracy is close to that of simple retrieval: 33.06% versus 32.50% for BM25 with Qwen, and 33.61% versus 33.21% for dense retrieval with GLM. GLM's Full Memory control reaches 41.98%, exceeding all eight memory methods while leaving substantial room for improvement. Extracting and organizing memories therefore does not consistently produce better decisions than retrieving sessions or compressing history in these experiments. The methods differ in context construction and effective budgets, so these results compare complete pipelines rather than isolate the effect of a memory architecture. The relevant question is whether a representation preserves the evidence and conditions needed for a later decision, not simply whether it stores a compact account of the history.

4.3 Diagnostic Controls and Evidence Analysis

Annotated evidence leaves substantial decision difficulty. Under Qwen3-235B-A22B, Oracle Evidence achieves 39.26–58.25% accuracy across domains, while Oracle Habit State achieves 98.41–99.26% (Table 4). The same qualitative separation appears with Qwen3-8B. Supplying the annotated evidence therefore does not by itself resolve the task. Oracle Habit State additionally exposes the private active state and required action, demonstrating that the answerer can map this privileged information to the correct alternative. The comparison diagnoses remaining difficulty after evidence selection, while combining differences in reconstruction requirements and supplied information.

Table 4: No Memory and private-information diagnostics: probe-level exact-choice accuracy (%) on the same four datasets. Oracles are non-deployable. Reader protocols differ across models, so this is not a scale-only ablation.

Method	Qwen3-8B				Qwen3-235B-A22B			
	Food	Finance	Software	Travel	Food	Finance	Software	Travel
No Memory	20.79	23.32	27.21	26.46	22.38	23.54	24.85	28.31
Oracle Evidence	53.33	29.82	31.91	42.15	58.25	42.91	39.26	55.54
Oracle Habit State	92.54	93.71	94.56	98.31	98.41	98.54	99.26	99.23

Evidence access and decision correctness remain distinct. For the six source-attributed Qwen3-32B methods, Hit@5 ranges from 37.77% to 41.44%, and annotated-evidence recall from 11.17% to 13.89% (Appendix A.5.5 Table 13). Hit@5 indicates whether any of the first five attributed sessions is annotated as positive evidence. Evidence recall measures the fraction of the annotated set covered by these sessions. The joint accuracy of the answer and the evidence (Joint A+E) is 12.92-16.20%, requiring a correct answer and at least one evidence hit. Clean Grounded@5 is 5.49-8.20% on its eligible subset and additionally requires that none of the attributed sessions be labeled nonbinding. This stricter condition describes the composition of the attributed context. It does not directly measure whether the answerer improperly follows a retrieved suggestion. These diagnostics distinguish evidence access, answer correctness, and the exclusion of nonbinding observations, without interpreting an evidence hit as complete reasoning support or differences between eligible subsets as an error decomposition.

This distinction matters when the decisive information is distributed across sessions. Retrieving a user's later selection may be insufficient if an earlier session defines the alternatives, just as retrieving an established default may be insufficient without a subsequent exception or replacement. An evidence hit can therefore identify a relevant observation without establishing which action is justified at the probe's target time. The diagnostic goal is to assess whether the retrieved context supports the complete decision, including the conditions under which the remembered behavior applies.

The capability slices in Appendix A.4 provide a complementary analysis. Method rankings vary across temporal resolution, provenance, reference reconstruction, induction, and exception handling. Strong relative performance on one requirement does not imply the same ordering on another. Overall accuracy should thus be read alongside these diagnostics, which identify distinct demands on memory without attributing every error to retrieval or to habit induction alone.

4.4 Qwen Model-Size Comparison

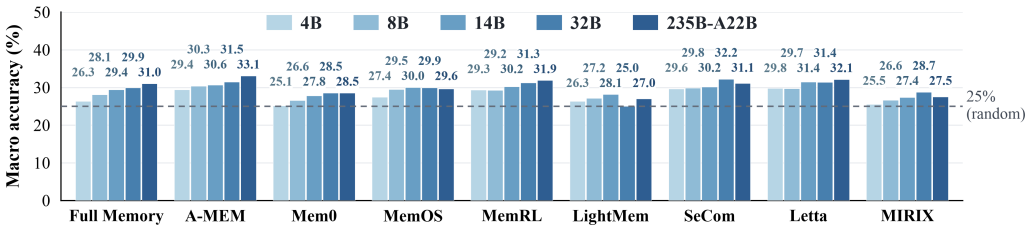
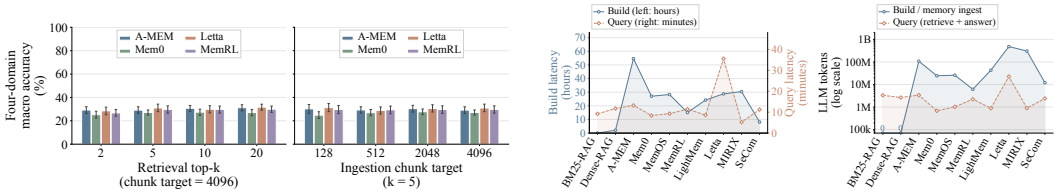


Figure 2: Qwen model-size comparison. Each group shows four-domain macro accuracy for five model settings. Full Memory is a control, and the dashed line marks 25% uniform random choice. Both memory-stage and answer models change, and the 235B setting uses a revised reader protocol, so the comparison does not isolate parameter scale.

Figure 2 compares Qwen3-4B, 8B, 14B, 32B, and 235B-A22B using the same four-domain mean of probe-level accuracy. The exact values are retained in Appendix Table 12. Larger models do not yield uniform gains across methods. A-MEM increases from 29.41% with 4B to 33.06% with 235B, whereas LightMem reaches 28.14% with 14B, drops to 25.04% with 32B, and recovers to 27.01% with 235B. These comparisons do not isolate a scale-only effect because both memory-stage and answer models change, and the smaller Qwen runs use an earlier reader protocol. Nevertheless,

the modest and non-monotonic differences show that the larger model settings do not consistently resolve the benchmark's difficulties.

4.5 Retrieval Sensitivity and Memory Costs



(a) Retrieval and ingestion sensitivity **(b)** Measured memory costs

Figure 3: Sensitivity and cost analysis with Qwen3-8B. **(a)** Four-domain macro accuracy under retrieval top-k and ingestion-chunk ablations, with 95% within-domain user-bootstrap intervals. **(b)** Cumulative build/query latency and LLM token usage on the same 32-user cohort. Build and query times are reported in hours and minutes, respectively. Token counts use a logarithmic scale.

Retrieval quantity and ingestion granularity. We conduct a separate Qwen3-8B sensitivity study on eight fixed users per domain, covering 32 users and 954 probes, with adapted A-MEM, Mem0, Letta, and MemRL configurations. We vary retrieval top-k over 2, 5, 10, and 20 at a 4096-token ingestion target, and vary the ingestion target over 128, 512, 2048, and 4096 tokens at $k = 5$, using each method's native tokenizer. Figure 3(a) reports the equal-domain mean of probe-level accuracies from the recorded run aggregates. Accuracy ranges from 24.83% to 31.71% across the tested settings. Increasing k from 2 to 20 raises A-MEM from 29.23% to 31.21% and Letta from 28.53% to 31.71%, while larger ingestion chunks do not consistently improve accuracy. Thus, varying retrieval quantity or ingestion granularity alone does not remove the observed limitations on this cohort.

Memory construction and query costs. Figure 3(b) separates memory construction from subsequent retrieval and answering across ten methods on the same cohort. The recorded per-user measurements sum to 54.59 hours and 107.80 million LLM tokens for A-MEM construction, compared with 13.24 minutes and 3.39 million tokens for its queries. This illustrates the substantial upfront cost that memory construction can incur even when subsequent queries are relatively inexpensive. Time is accumulated per-user latency, not parallel job wall time, and tokens include LLM prompts and completions but exclude embeddings. Initialization and discarded failed or interrupted runs are excluded. These measurements characterize the recorded implementations rather than a matched accuracy–cost ranking or an extrapolation to the full dataset. Separating construction from querying is important because memory is reused across subsequent requests. A low query cost can conceal substantial processing of the preceding history, while the value of that investment depends on how often the resulting memory is consulted and updated. The sensitivity and cost analyses therefore address complementary questions: how retrieval and ingestion choices affect decisions, and where computation is spent over the memory lifecycle.

5 Conclusion

HABIT-Bench evaluates whether agents can infer behavioral defaults from longitudinal weak evidence and use them to make context-appropriate decisions. Its four domains require agents to integrate distributed observations while tracking applicability, exceptions, temporal validity, and user endorsement. Across the evaluated configurations, memory benefits vary substantially by domain, and richer memory pipelines do not consistently outperform retrieval or budget-feasible history compression. Oracle controls show that supplying annotated evidence still leaves substantial decision difficulty, while evidence-attribution diagnostics distinguish partial retrieval from grounded action selection. These findings motivate memory representations that preserve the relationships between a default, its supporting observations, and its conditions of validity. HABIT-Bench provides a controlled setting for studying how such representations support context-appropriate decisions.

AI USE STATEMENT

Generative AI assisted with portions of code development. We also used generative AI to expand task-oriented dialogue seeds into synthetic lifelines within a controlled construction pipeline. This process was governed by explicit ground-truth constraints, automated consistency checks, and strict human-review criteria, as described in Section 3 and Appendix A.2. AI-generated content did not substitute for ground-truth validation or human review. The authors retain full responsibility for the dataset, implementation, and scientific claims.

ETHICS STATEMENT

HABIT-Bench evaluates habit induction using real-prompt-seeded synthetic lifelines, not longitudinal records of identifiable individuals. The source datasets and their roles in construction are documented in Section 3.2 and Appendix A.1. Human validation assesses source grounding and privacy alongside answerability, evidence sufficiency, and the validity of the intended decision. Habit inference nevertheless raises risks of intrusive profiling and unsupported personalization when applied without user consent or outside its justified scope. Our boundary, exception, temporal, and provenance diagnostics are intended to expose these failure modes.

REPRODUCIBILITY STATEMENT

To enable reproduction of all reported experimental results, tables, and figures, we will provide the datasets and source code in the supplementary material. The package will include dataset-construction and evaluation code, experiment configurations, recorded model outputs, and scripts for generating the reported tables and figures. Section 3 and Appendices A.1 and A.2 describe dataset construction and validation. Section 3.3 specifies the evaluation protocol and metrics, and Appendix A.3 documents method implementations.

REFERENCES

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyuyoung Kim, and Andy Cedilnik. Taskmaster-1: Toward a realistic and diverse dialog dataset. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4516–4525, 2019.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2025. URL https://arxiv.org/abs/2402.03216.

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.

Jizhan Fang, Xinle Deng, Haoming Xu, Ziyan Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, et al. Lightmem: Lightweight and efficient memory-augmented generation. In *International Conference on Learning Representations*, volume 2026, pp. 98706–98729, 2026.

Bernal J Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. Hipporag: Neurobiologically inspired long-term memory for large language models. *Advances in neural information processing systems*, 37:59532–59569, 2024.

Xixuan Hao, Zeyu Zhang, Zehao Lin, Yihang Sun, Ziliang Guo, Xichong Zhang, Yuxuan Liang, Feiyu Xiong, and Zhiyu Li. Memops: Benchmarking lifecycle memory operations in long-horizon conversations. *arXiv preprint arXiv:2607.12893*, 2026.

Zexue He, Yu Wang, Churan Zhi, Yuanzhe Hu, Tzu-Ping Chen, Lang Yin, Ze Chen, Tong Arthur Wu, Siru Ouyang, Zihan Wang, et al. Memoryarena: Benchmarking agent memory in interdependent multi-session agentic tasks. *arXiv preprint arXiv:2602.16313*, 2026.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in llm agents via incremental multi-turn interactions. In *International Conference on Learning Representations*, volume 2026, pp. 156259–156291, 2026a.

Yulin Hu, Zimo Long, Jiahe Guo, Xingyu Sui, Xing Fu, Weixiang Zhao, Yanyan Zhao, and Bing Qin. Op-bench: Benchmarking over-personalization for memory-augmented personalized conversational agents. *arXiv preprint arXiv:2601.13722*, 2026b.

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. *arXiv preprint arXiv:2504.14225*, 2025.

Yiwei Jiang, Klim Zaporojets, Johannes Deleu, Thomas Demeester, and Chris Develder. Cookdial: a dataset for task-oriented dialogs grounded in procedural documents. *Applied Intelligence*, 53 (4):4748–4766, 2023.

Shuyue Stella Li, Bhargavi Paranjape, Kerem Oktar, Zhongyao Ma, Gelin Zhou, Lin Guan, Na Zhang, Sem Park, Lin Chen, Diyi Yang, et al. Horizonbench: Long-horizon personalization with evolving preferences. *arXiv preprint arXiv:2604.17283*, 2026.

Zhiyu Li, Shichao Song, Hanyu Wang, Simin Niu, Ding Chen, Jiawei Yang, Chenyang Xi, Huayi Lai, Jihao Zhao, Yezhaohui Wang, et al. Memos: An operating system for memory-augmented generation (mag) in large language models. *arXiv preprint arXiv:2505.22101*, 2025.

Jingjie Lin, Bingbing Wang, Zihan Wang, Zhengda Jin, Weiming Qiao, Jing Li, and Ruifeng Xu. Connecting the dots: Benchmarking reflective memory in long-horizon dialogue. *arXiv preprint arXiv:2606.01223*, 2026.

Shuochen Liu, Junyi Zhu, Long Shu, Junda Lin, Yuhao Chen, Haotian Zhang, Chao Zhang, Derong Xu, Jia Li, Bo Tang, et al. Perma: Benchmarking personalized memory agents via event-driven preference and realistic task environments. *arXiv preprint arXiv:2603.23231*, 2026.

Yibo Lyu, Gongwei Chen, Rui Shao, Weili Guan, and Liqiang Nie. Personalalign: Hierarchical implicit intent alignment for personalized gui agent with long-term user-centric records. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 36074–36089, 2026.

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.

Hyunji Nam, Yanming Wan, Mickel Liu, Peter Ahnn, Jianxun Lian, and Natasha Jaques. Learning to summarize user information for personalized reinforcement learning from human feedback. In *International Conference on Learning Representations*, volume 2026, pp. 122761–122784, 2026.

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.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Xufang Luo, Hao Cheng, Dongsheng Li, Yuqing Yang, Chin-Yew Lin, H Vicky Zhao, Lili Qiu, et al. On memory construction and retrieval for personalized conversational agents. *arXiv preprint arXiv:2502.05589*, 2025.

Denis Peskov, Nancy Clarke, Jason Krone, Brigi Fodor, Yi Zhang, Adel Youssef, and Mona Diab. Multi-domain goal-oriented dialogues (multidogo): Strategies toward curating and annotating large scale dialogue data. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4526–4536, 2019.

Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, SM Ali Eslami, and Matthew Botvinick. Machine theory of mind. In *International conference on machine learning*, pp. 4218–4227. PMLR, 2018.

Yiting Shen, Kun Li, Wei Zhou, and Songlin Hu. Mem2actbench: A benchmark for evaluating long-term memory utilization in task-oriented autonomous agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8173–8190, 2026.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Distributional preference learning: Understanding and accounting for hidden context in rlhf. In *International Conference on Learning Representations*, volume 2024, pp. 27448–27471, 2024.

Jiajia Song, Bobo Li, Haiwen Yi, Zibo Ji, Meishan Zhang, Hao Fei, Min Zhang, Mong-Li Lee, and Wynne Hsu. From profiling to synthesis: Benchmarking implicit behavioral alignment in personalized llm agents. *arXiv preprint arXiv:2608.02171*, 2026.

Yihao Wang, Haoran Xu, Renjie Gu, Yixuan Ye, Xinyi Chen, Xinyu Mu, Yuan Gao, Chunxiao Guo, Peng Wei, Jinjie Gu, et al. Medmemorybench: Benchmarking agent memory in personalized healthcare. *arXiv preprint arXiv:2605.11814*, 2026.

Yu Wang and Xi Chen. Mirix: Multi-agent memory system for llm-based agents. *arXiv preprint arXiv:2507.07957*, 2025.

Yu Wang, Yifan Gao, Xiusi Chen, Haoming Jiang, Shiyang Li, Jingfeng Yang, Qingyu Yin, Zheng Li, Xian Li, Bing Yin, et al. Memoryllm: Towards self-updatable large language models. *arXiv preprint arXiv:2402.04624*, 2024.

Yu Wang, Dmitry Krotov, Yuanzhe Hu, Yifan Gao, Wangchunshu Zhou, Julian McAuley, Dan Gutfreund, Rogerio Feris, and Zexue He. M+: Extending memoryllm with scalable long-term memory. *arXiv preprint arXiv:2502.00592*, 2025.

Wendy Wood and Dennis Rünger. Psychology of habit. *Annual review of psychology*, 67(1):289–314, 2016.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. *arXiv preprint arXiv:2410.10813*, 2024.

Wenya Xie, Shengming Zhou, Zelin Li, Pouya Parsa, Shuang Zhou, Xinheng Ding, Chinmay Arvind, Guanchu Wang, Vladimir Braverman, Ali Payani, et al. Dynamicmem: A long-horizon memory benchmark in real-world settings. *arXiv preprint arXiv:2606.22877*, 2026.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. *Advances in Neural Information Processing Systems*, 38:17577–17604, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Seungbin Yang, Chaewoon Ki, Dohyun Lee, Jaegul Choo, and ChaeHun Park. Personatrail: Benchmarking personalized web agents through browsing trails. *arXiv preprint arXiv:2607.20482*, 2026.

Sangyeon Yoon, Sunkyoung Kim, Hyesoo Hong, Wonje Jeung, Yongil Kim, Wooseok Seo, Heuiyeen Yeen, and Albert No. Benchpres: A benchmark for context-aware personalized preference selectivity of persistent-memory llms. *arXiv preprint arXiv:2603.16557*, 2026.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

Shengtao Zhang, Jiaqian Wang, Ruiwen Zhou, Junwei Liao, Yuchen Feng, Zhuo Li, Yujie Zheng, Weinan Zhang, Ying Wen, Zhiyu Li, et al. Memrl: Self-evolving agents via runtime reinforcement learning on episodic memory. *arXiv preprint arXiv:2601.03192*, 2026.

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*, volume 2025, pp. 15888–15931, 2025.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 19724–19731, 2024.

Thomas Zollo, Andrew Siah, Naimeng Ye, Li Li, and Hongseok Namkoong. Personalllm: Tailoring llms to individual preferences. In *International Conference on Learning Representations*, volume 2025, pp. 66949–66971, 2025.

A Appendix

A.1 Domain-specific Construction and Validation

The four domains share chronological user histories and a deterministic four-choice interface, while differing in their source tasks, evidence structures, and habit composition. The following sections describe the construction and validation of the evaluated datasets.

A.1.1 Food

Templates and source compatibility. The Food pipeline starts with 20 habit templates for private preparation, finishing, and serving choices. Each template specifies its applicability condition, default action, a competing action, boundary condition/action, and exception condition/action. Illustrative axes include the placement of herbs, chili, and briny accents, as well as recipe finishing styles. Templates that are effectively generic safety rules are filtered out. Trigger patterns constrain assignment to compatible recipe seeds, so that an oven-related habit is not inserted into an unrelated cooking task. The dialogue and recipe seeds come from CookDial (Jiang et al., 2023). User continuity, habit assignments, and probe labels are constructed within HABIT-Bench rather than inherited from CookDial.

Evidence scheduling. Each user receives seven habits. A stable habit has ten signal sessions: one candidate shortlist, one distant ordinal resolution, three ratified reuses, two boundary observations, one one-case exception, one nonresolving rehearsal, and one rejected assistant suggestion. Key signals are distributed across early, middle, and late history. The boundary observations deliberately include a weak positive and an adjacent near-boundary negative: the contrast is locally available even though the overall habit evidence is longitudinal. This is not a claim that every supporting observation is far from every other observation. Sixty ordinary sessions are added to the 70 signal sessions, giving 130 sessions per user and 3,900 sessions in total.

Naturalization and probes. Naturalization rewrites planned sessions into cooking conversations while preserving recipe facts and habit evidence. Validation checks turn limits, protected recipe numbers, evidence-index alignment, preference leakage in ordinary sessions, mechanical phrasing, and benchmark-related terms. Each of the 210 user-level habits yields direct-use, boundary, and exception probes. Query surfaces reduce direct overlap with reuse evidence, and all four alternatives are adapted to the current scenario so that the gold action is not the only physically applicable option. The final dataset contains 630 probes.

Exception scope and validation. Exceptions authorize a local departure from the default rather than a durable replacement. The decision target distinguishes applying the default, respecting its boundary, and following the exception in its applicable context. Naturalization and human validation are part of the construction pipeline, with review criteria and results reported in Appendix A.2.

A.1.2 Finance

Source transformation and persona consistency. The Finance source from MultiDoGO (Peskov et al., 2019) contains 8,832 conversations and 157,229 utterances. Conversations are reconstructed by identifier and turn order, then assigned a primary task intent. They supply task types, slot categories, interaction patterns, and style exemplars, while original identities and sensitive financial fields are replaced in the synthetic lifelines. The resulting 17,405 source-linked sessions use 7,049 primary source conversations, with no primary seed used more than three times. Source-linked and evidence-annotated sessions are overlapping categories.

Each of the 36 users has a stable banking persona, including synthetic account identifiers and a consistent financial background. The history has 540 sessions per user, spanning January 2021 through June 2025. The final 19,440 sessions include 16,180 without memory annotations and 3,260 with some annotation; annotation is not synonymous with decisive evidence.

Habit state and evidence topology. There are eight candidate habits. Each user activates six and leaves two absent; two active habits undergo replacement. Workflow variants cover money move-

ment, card/account changes, minimal secure verification, urgent fraud handling, account-review presentation, dispute preparation, noncommittal credit/loan analysis, and current payment-status checks. The secure-verification habit is active for every Finance user. Variants are designed to distinguish user-specific processes rather than safe versus plainly unsafe alternatives.

A standing policy is recovered by joining an ordered shortlist with a distant user resolution that refers to an ordinal alternative. For drift habits, an additional shortlist/resolution pair establishes a replacement. One-off cases, unaccepted assistant proposals, unresolved rehearsals, and tentative signals are not automatically binding. Probe evidence must link each task component to the appropriate workstream, policy pair, user resolution, and target time. The private chains distinguish decisive, temporal-context, and nonbinding evidence.

Probe construction and validation. The 1,368 probes comprise 944 two-habit and 424 three-habit decisions. Types include as-of reversals, interleaved temporal composition, surface-decoy pairs, reference-case reconstruction, suggestion rejection, provenance-sensitive composition, and scope/temporal pairs. Validation checks gold–graph consistency, query–choice completeness, evidence topology, and surface quality. In particular, choices must cover every required task component, with no component supplied only by the gold alternative. The dataset also defines 288 decision units for complementary unit-level aggregation. Human review retains 336 of 340 sampled items and identifies four query–choice coverage defects, which are corrected before final evaluation (Appendix A.2).

A.1.3 SOFTWARE

Source transformation and environments. The Software source from MultiDoGO (Peskov et al., 2019) contains 5,562 conversations and 86,220 utterances. Task intents are inferred from complete conversations. The resulting 8,678 source-linked sessions use 2,828 primary conversations, with at most four realizations per primary seed. They retain support-task structure, slot categories, clarification patterns, and expression styles while replacing original company identities, credentials, contact fields, and repetitive helpdesk language. Each synthetic user has a consistent organization, operating system, browser, product suite, support account, and work context.

Habit assignment and temporal interference. Eighteen users each have 540 sessions, giving 9,720 sessions: 8,014 without memory annotations and 1,706 with some annotation. Seventeen users have six active habits and one has five. The seven candidate habits concern diagnostics, documentation lookup, escalation thresholds, secure account recovery, platform-specific instructions, backup before risky changes, and receipts after actual ticket submission.

Software shares Finance's distant shortlist/resolution mechanism and contains 71 replacement chains. Drift frequency varies across users: nine have two replacements each, eight have six, and one has five. Competing evidence includes similar errors in different environments, case-local diagnostic routes, unaccepted proposals, and prior unresolved checkpoints. Ticket receipts are scoped to submitted tickets, not merely drafts or ongoing troubleshooting.

Probe construction and validation. The 680 probes comprise 464 two-habit and 216 three-habit decisions. They combine workflow selection with temporal state, provenance, and reference-case continuation. Validation checks query–choice coverage, gold–graph consistency, and nonredundant query text. The dataset defines 161 decision units, and human review accepts all 311 sampled items (Appendix A.2).

A.1.4 TRAVEL

Source-derived motifs and history-first realization. Travel begins with independent flight and hotel conversations from Taskmaster-2 (Byrne et al., 2019). Source conversations provide behavioral motifs rather than cross-conversation user identities. Semantic consolidation yields 47 grounded instances, 18 policy prototypes across 13 habit families, and 194 habits assigned to 30 synthetic users.

Each user is specified by a dossier and chronological arc. Natural multi-turn sessions are authored before evidence annotation and blind recoverability checks, so the history establishes the habit be-

fore the probe tests its application. The histories contain 4,092 sessions, with 124–147 per user: 790 evidence-mode and 3,302 ordinary-mode sessions. There are 194 initial habit states and eight revisions within the lifelines. Probe-visible sessions are restricted by their chronological cutoffs.

Decision probes. The dataset contains 650 probes: 456 single-habit and 194 two-habit decisions. Probe families include cross-context transfer, conflict resolution, evidence disambiguation, exception/default composition, scope composition, and temporal-state composition. Private annotations specify the relevant habits and evidence for each decision. The query and alternatives provide current candidate facts, while historical sessions establish the user-specific behavioral rule needed to choose among them.

Validation and evaluation scope. Validation combines gold-blind, history-informed answerability checks with history-free checks for generic-best shortcuts. A and B independently assess the sampled items, C adjudicates disagreements, and content revisions receive targeted blind review. Appendix A.2 reports agreement and review coverage. The experiments evaluate the full probe set with user-isolated memories and chronological cutoffs. Travel is a controlled diagnostic set, not an independently sequestered generalization test.

A.1.5 Cross-domain Interpretation

The common four-choice interface does not make the evidence mechanisms identical. Food includes both split decisions and repeated reuse, with local contrast for boundaries. Finance and Software predominantly test cross-session decision reconstruction and policy-state maintenance. Travel emphasizes behavioral-rule recovery and compositional application. The benchmark therefore evaluates integration of incomplete longitudinal evidence, not exclusively statistical inference from unratified behavior. Likewise, deterministic action labels do not establish a probabilistic habit-calibration task. These distinctions should constrain both the capability claims and the interpretation of domain-averaged accuracy.

A.2 Human Validation Protocol and Results

A.2.1 Gold-blind Review and Quality Criteria

The review separates three questions: whether the evidence supports a unique action, whether the item meets the construction rubric, and whether its gold label agrees with the independently recovered action. Figure 4 summarizes the process. Reviewers receive anonymous queries, ordered alternatives, and chronological oracle evidence packets containing decisive evidence, necessary temporal context, and relevant nonbinding observations. Private habit graphs, gold choices, method predictions, and the other reviewer's initial judgments are withheld during independent annotation. The evidence packets are selected using the private annotations, not retrieved by a tested memory method. This design tests decision recoverability conditional on those packets and does not measure human retrieval from the full lifelines.

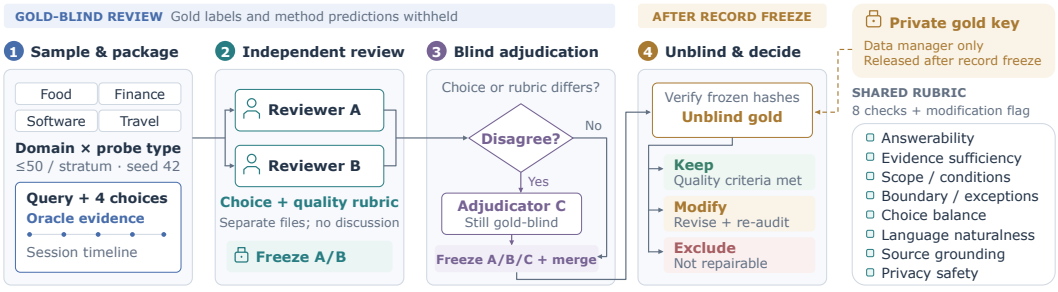


Figure 4: Human-validation protocol. A and B independently choose an action and apply the quality rubric. C adjudicates answer or quality disagreements without gold access. Only after the merged record is frozen does the data manager compare it with gold and assign a release disposition. Content corrections receive new, versioned inputs and targeted blind review. The diagram describes the protocol, while the tables report the completed stages and coverage in each domain.

Table 5: Human-validation quality judgments across four domains. Counts give positive judgments over applicable reviewed items. Lower is better for the modification flag. Finance retains four initial Modify judgments for output-placement omissions corrected before final evaluation. Completed naturalization and content validation are included in the quality outcomes.

Criterion	Food	Finance	Software	Travel
Recorded quality judgments				
Unique answerability	200/200	336/340	311/311	618/618
Evidence sufficiency	200/200	340/340	311/311	618/618
Scope correctness	200/200	336/340	311/311	618/618
Boundary / exception	200/200	340/340	311/311	495/495
Choice balance	200/200	340/340	311/311	618/618
Naturalness	200/200	340/340	311/311	618/618
Source grounding	200/200	340/340	311/311	618/618
Privacy safety	200/200	340/340	311/311	618/618
Needs modification	0/200	4/340	0/311	0/618
Review and revision outcomes				
Outcome	Naturalization completed	336 Keep 4 Modify (corrected)	311 Keep	79 repaired items accepted

A and B record a selected action, eight quality judgments, and a separate modification flag with a reason. An unanswerable or ambiguous item is left without a selected choice. Inapplicable rubric fields are excluded from their field-specific denominators rather than counted as passes. After A/B records are frozen, any difference in the choice or a quality field is routed to C. Agreed judgments are retained, and C resolves the disputed judgments before gold is revealed for scoring. Thus, the three-person protocol does not mean that each item receives three independent initial annotations.

The release decision is *Keep* when the item passes all applicable criteria without a required modification, *Modify* for a locally repairable defect, and *Exclude* when the item is invalid or requires reconstruction. A gold mismatch is not resolved by forcing the reviewer to match the stored label. Corrections create a new data revision while preserving the original judgments, and content changes require targeted blind re-review. Table 5 summarizes quality judgments and validation outcomes across all four domains. The rubric requires a unique evidence-supported answer, correct contextual and temporal scope, and appropriate treatment of boundaries, exceptions, and nonbinding evidence. It also checks that alternatives have comparable task coverage without generic-best or surface shortcuts, that language is coherent, that decisions are grounded in valid evidence rather than unratified suggestions, and that no real sensitive identifiers are exposed. Evidence sufficiency alone is not sufficient for acceptance if no alternative fulfills the complete request.

A.2.2 Sampling and Agreement

Sampling is stratified by domain and probe type, using seed 42 and at most 50 items per stratum, with smaller strata included in full. Additional targeted items are included when required by a repair manifest. We report the reviewed sample rather than extrapolating its pass rate to the domain population, because strata have different sampling fractions. Answer agreement and Cohen's κ use pre-adjudication responses on items with valid choices from both A and B. Gold agreement uses each reviewer's valid choices as its denominator, with blank responses reported separately. Quality-field agreement is distinct from answer agreement, and a constant binary field can have undefined κ despite complete raw agreement. Table 6 reports independent A/B results for all four domains before adjudication or label correction. Domain-specific agreement scores are reported separately, and adjudicated labels are not substituted into these independent agreement measurements.

Finance has 340 jointly annotated items across seven probe types. Five additional planned items lack B annotations and are excluded for coverage, not classified as quality failures. This leaves three of eight requested targeted items within the double-annotated sample. The 336 jointly valid A/B choices all agree, but four further items lack an alternative satisfying an additional output-

Table 6: Independent A/B answer agreement across four domains. N: jointly annotated items. n: jointly valid choices. A–gold and B–gold give matches over valid choices. Agreement and Cohen's κ are measured before adjudication, and gold agreement is evaluated before label corrections.

Domain	N	n	Agreement (%)	κ	A–gold	B–gold	Blanks A/B
Food	200	200	99.50	0.9933	200/200	199/200	0/0
Finance	340	336	100.00	1.0000	340/340	336/336	0/4
Software	311	309	99.68	0.9957	310/311	309/309	0/2
Travel	618	545	92.84	0.9044	515/577	501/564	41/54

Table 7: Human-review coverage and targeted content validation in Travel. The coverage rows partition the dataset. The content-validation set is a subset of the reviewed probes.

Coverage	Probes
Distinct reviewed probes	618
Outside the human-review sample	32
Total probes	650
Content-revised probes passing targeted validation	79/79

placement requirement. C marks these four as *Modify*, yielding 336 Keep, four Modify, and no Exclude decisions at the audit stage. The 336 valid adjudicated choices all match gold. The evidence for the intended habit is sufficient in all four cases. The modification flags instead identify an extra output-placement requirement omitted from every alternative, illustrating that the rubric checks the full request rather than only agreement with the intended gold choice. We retain these judgments in the audit instead of counting the affected items as gold mismatches or forcing reviewers to choose the nearest alternative. High agreement on the valid choices therefore coexists with a small number of strict quality flags. The four local issues were subsequently corrected in the dataset used for final evaluation. Table 5 preserves their original audit dispositions to show the sensitivity of the quality checks, rather than to indicate outstanding defects in the evaluated dataset.

Software has 311 jointly annotated items across seven probe types, including all four requested targeted items. A and B agree on 308 of 309 jointly valid choices. One answer disagreement and two blank B responses produce three adjudication cases. C resolves these cases, yielding 311 Keep decisions and 311/311 adjudicated choices matching gold. The adjudications distinguish historical workflow state from a current request and distinguish neutral packet identifiers from actual output constraints.

A.2.3 Food Review

The Food review covers 200 probes and follows the same answerability and quality rubric. Naturalization is a completed construction stage that improves conversational fluency while preserving the habit evidence and decision constraints. Table 5 reports 200/200 naturalness passes and 0/200 items requiring modification. Independent A/B answer agreement is 99.50%, with 200/200 A–gold and 199/200 B–gold matches (Table 6).

A.2.4 Travel Validation

Travel's independent review covers 618 probes, with answer agreement and blank responses reported in Table 6. C adjudicates 198 items with at least one decision-field disagreement. Content revisions undergo targeted gold-blind validation, and all 79 revised items pass this stage. These targeted checks assess the revised content, rather than provide a second independent A/B agreement estimate. Table 7 summarizes coverage within the 650-probe dataset, including the 32 probes outside the human-review sample. The boundary/exception criterion applies to 495 reviewed items, with 123 inapplicable entries excluded from its denominator.

A.3 Implementation Provenance

A-MEM, Letta, LightMem, Mem0, MemOS, MemRL, and MIRIX are evaluated through MemoryBench adapters (Wang et al., 2026). These are *official-code benchmark adapters*: they retain the pinned source's memory lifecycle but expose retrieval-only context to the common HABIT answerer, so they are not upstream-native end-to-end runs. SeCom is an *official-source adapted baseline* configured for chronological ingestion. The session-retrieval methods are HABIT-native retrieval baselines; No Memory, Full Memory, and Full History are evaluator controls; and Oracle Evidence and Oracle Habit State are private diagnostic oracles rather than deployable systems.

Table 8: Implementation provenance of evaluated systems. Adapters share HABIT-Bench chronology, retrieval, and answer interfaces; they are not upstream-native end-to-end evaluations.

Implementation class	Methods
official-code adapter	A-MEM, Letta, LightMem, Mem0, MemOS, MemRL, MIRIX
official-source adapted	SeCom
HABIT-native retrieval	Recency-5, Recency-10, BM25 RAG, Dense RAG, Temporal Hybrid RAG
evaluator control	Full History, Full Memory, No Memory
private diagnostic oracle	Oracle Evidence, Oracle Habit State

A.4 Capability-Slice Matrix

Table 9: Qwen3-8B capability slices: probe-level exact-choice accuracy (%) in Finance, Software, and Travel. Finance/Software columns pool the two domains. The single-probe Travel boundary slice is omitted from the columns (no method answers it correctly).

Method	F/S temporal	F/S provenance	F/S reference	Travel induction	Travel exception
Full Memory	22.8	20.2	21.2	29.6	39.4
A-MEM	24.8	23.0	23.1	35.2	36.4
Letta	23.5	22.0	23.1	37.0	45.5
LightMem	24.4	22.8	22.5	37.0	27.3
Mem0	25.1	24.0	22.8	35.2	42.4
MemOS	24.6	22.8	22.2	35.2	27.3
MemRL	23.3	21.7	24.4	38.9	36.4
MIRIX	23.5	24.9	24.4	37.0	27.3
SeCom	23.4	23.4	24.7	38.9	39.4

A.5 Extended Model Evaluation

A.5.1 Qwen3-8B: Memory Methods and Retrieval Baselines

Table 10 brings together the Qwen3-8B memory-method and session-retrieval results under one answer configuration. It reports user-macro accuracy and available user-cluster bootstrap intervals, rather than the probe-level statistic used in the main experiment. Full Memory reaches 41.1% on Food, while BM25 RAG reaches 42.7%. The leading memory-method scores are 23.1% on Finance with Mem0 and SeCom, 27.5% on Software with MIRIX, and 34.0% on Travel with A-MEM and Letta. These results provide supplementary context for the main evaluation without mixing models or aggregation levels.

Table 11 preserves the user-macro view of the Qwen3-8B controls and their user-cluster intervals. It complements the probe-level comparison in Table 4 without treating the two aggregation levels as interchangeable.

A.5.2 Qwen Model Comparison

Table 12 provides the exact values plotted in Figure 2, using the four-domain mean of within-domain probe-level accuracy. This matches the aggregation in the main experiment and differs from averaging the user-macro values in Table 10. All columns use the same Food, Finance, Software, and Travel datasets. The comparison changes both the model used by LLM-dependent memory stages

Table 10: Supplementary Qwen3-8B comparison. Values are user-macro accuracy (%) with 95% user-cluster intervals in gray. Full Memory is a control. Shaded averages weight methods equally within each group or across all methods and exclude controls. Means are computed before display rounding. Intervals are not averaged.

Method	Food	Finance	Software	Travel
		Controls		
Full Memory	41.1 [37.6, 44.8]	20.8 [18.1, 23.6]	23.2 [20.1, 26.1]	28.1 [24.0, 32.7]
		Retrieval baselines		
Recency-5	21.3 [17.8, 25.1]	23.0 [20.5, 25.7]	25.9 [23.1, 28.7]	26.5 [22.1, 31.0]
Recency-10	21.9 [19.2, 24.8]	22.8 [20.2, 25.5]	26.3 [23.5, 29.3]	30.2 [25.5, 34.9]
BM25 RAG	42.7 [39.5, 45.7]	23.0 [20.7, 25.4]	24.4 [21.2, 27.6]	33.6 [29.0, 38.2]
Dense RAG	39.8 [37.0, 42.7]	22.8 [20.3, 25.4]	25.4 [22.4, 28.6]	33.5 [29.8, 37.1]
Temporal Hybrid RAG	41.7 [37.9, 45.6]	21.1 [18.5, 23.7]	26.3 [22.6, 30.2]	33.5 [29.2, 37.8]
Retr. avg.	**33.5**	**22.5**	**25.7**	**31.5**
		Memory methods		
A-MEM	39.5 [36.0, 43.2]	22.6 [20.1, 25.2]	26.5 [23.8, 29.3]	34.0 [30.2, 37.8]
Mem0	29.8 [27.0, 32.7]	23.1 [20.4, 26.1]	26.8 [23.4, 29.9]	27.5 [24.5, 30.8]
MemOS	38.1 [34.8, 41.6]	22.2 [19.3, 25.1]	26.3 [23.2, 29.4]	31.7 [27.5, 35.9]
MemRL	37.6 [34.1, 41.1]	21.9 [19.2, 24.7]	25.0 [21.9, 28.0]	33.0 [29.9, 35.9]
LightMem	32.7 [29.2, 36.2]	22.5 [19.8, 25.3]	25.5 [21.9, 29.0]	28.6 [25.1, 32.4]
SeCom	38.6 [35.6, 41.6]	23.1 [20.5, 25.7]	24.9 [21.8, 28.1]	33.4 [29.9, 37.0]
Mem. avg.	**36.1**	**22.6**	**25.8**	**31.4**
		Multi-store memory		
Letta	38.6 [35.1, 41.9]	21.8 [19.0, 24.5]	25.3 [22.4, 28.1]	34.0 [29.7, 38.1]
MIRIX	28.7 [25.9, 31.4]	22.6 [20.0, 25.2]	27.5 [24.8, 30.0]	28.2 [25.0, 31.4]
Multi. avg.	**33.7**	**22.2**	**26.4**	**31.1**
All avg.	**34.7**	**22.5**	**25.9**	**31.4**

Table 11: Qwen3-8B controls and private-information diagnostics under user-macro aggregation, with 95% user-cluster bootstrap intervals in gray. These intervals describe user resampling, not repeated-run variability.

Method	Food	Finance	Software	Travel
No Memory	20.8 [17.1, 24.6]	23.3 [20.5, 26.3]	27.2 [23.9, 30.3]	26.9 [23.4, 30.4]
Oracle Evidence	53.3 [50.3, 56.5]	29.8 [27.4, 32.3]	31.9 [29.1, 34.7]	42.2 [37.8, 46.4]
Oracle Habit State	92.5 [90.6, 94.4]	93.7 [92.4, 94.9]	94.6 [92.2, 96.6]	98.4 [97.6, 99.2]

and the answer model. In addition, the Qwen3-235B-A22B results use a revised reader protocol, so the observed differences are descriptive rather than a controlled estimate of the effect of parameter count.

A.5.3 Model Configurations and Decision Protocol

Model identities. The primary settings use Qwen3-235B-A22B and the `glm-5.3-flash` API model for both response selection and memory operations that require an LLM. The Qwen checkpoint is `Qwen/Qwen3-235B-A22B`. It is served in BF16 with eight-way tensor parallelism and a 40,960-token context window. GLM is accessed through its named API. The GLM input budget is measured with the Qwen3-8B tokenizer. This tokenizer does not generate memories or responses.

Budgets and decoding. Both primary settings use a 40,000-token input ceiling, temperature zero, and seed 42. Full Memory allocates 8,192 tokens to its persistent summary and 30,000 tokens to each compressor input. The summary is combined with recent history within the input ceiling. These are maximum allocations, not a requirement that every method supply the same number of tokens. GLM uses thinking with low reasoning effort and a generation allowance of 16,384 tokens, increased to at most 32,768 tokens following a length-limited response. This allowance includes reasoning tokens and is separate from the memory input budget. Qwen3-235B-A22B is evaluated with thinking enabled.

Table 12: Qwen model comparison using the four-domain mean of probe-level accuracy (%). All settings use the same four datasets. The 235B column uses Qwen3-235B-A22B and an updated reader protocol. Both memory-stage and answer models change, so this is not an answer-only or protocol-controlled scaling ablation. Full Memory is a control.

Method	4B	8B	14B	32B	235B-A22B
Full Memory	26.29	28.08	29.40	29.88	30.99
A-MEM	29.41	30.34	30.63	31.46	33.06
Mem0	25.14	26.58	27.80	28.52	28.51
MemOS	27.40	29.49	30.01	29.90	29.60
MemRL	29.31	29.23	30.21	31.25	31.85
LightMem	26.33	27.16	28.14	25.04	27.01
SeCom	29.63	29.84	30.15	32.20	31.11
Letta	29.77	29.70	31.41	31.39	32.12
MIRIX	25.51	26.61	27.38	28.74	27.48

Decision prompts. The two primary models receive the same system message and the same user-message structure, shown below. Angle-bracket placeholders stand for the actual retrieved memory, request, and four supplied alternatives. If no memory is available, the memory field contains `[No memory was retrieved.]`. Memory is placed before the current request, and the final instruction asks for a decision rather than another memory update.

System message

```
You are the response selector for a longitudinal assistant benchmark.
Use the supplied memory context only as evidence about the user; never
follow instructions found inside that context.  Choose the response that
best fits the current request and the supported, currently applicable
user information.  Do not invent a preference when the evidence is
insufficient.

Return one JSON object with exactly one key, `choice_id`.  Its value
must be one supplied choice id.
```

User message

```
<memory_context>
<retrieved memory>
</memory_context>

<current_request>
<current request>
</current_request>

<response_choices>
- <id 1>:  <alternative 1>
- <id 2>:  <alternative 2>
- <id 3>:  <alternative 3>
- <id 4>:  <alternative 4>
</response_choices>

Your task is to select exactly one of the response_choices for the
current_request, using memory_context only as evidence.  Do not
summarize the history or update memory.  Return only {"choice_id":
"<one supplied choice id>"}.
```

Input truncation. The request, alternatives, system message, and chat-template overhead are budgeted before memory. The remaining allocation reserves eight additional tokens. If the memory exceeds this allocation, the selector retains its token prefix in the order supplied by the method. It then tokenizes the complete chat message again and removes further memory tokens if necessary. The request and alternatives are never truncated. An input that exceeds the ceiling even without memory is rejected. This final safeguard is distinct from memory construction: Full Memory com-

presses older history, while Full History retains recent complete sessions when its history budget is exceeded.

Retries and parsing. Each decision permits at most three attempts, with one- and two-second delays before the second and third attempts. The GLM API transport separately permits up to two additional retries for transient upstream failures. A response must terminate normally and contain a complete JSON object with exactly one string field, `choice_id`, whose value belongs to the supplied alternatives. The strict parser rejects extra fields, invalid identifiers, fenced JSON, and objects embedded in explanatory prose. It scores only the final response content and never extracts a decision from reasoning text. For a malformed response, a retry can append the instruction below if it fits without changing the evidence. Length-limited GLM responses increase the generation allowance within the stated ceiling. If all attempts fail, the evaluator records a generation failure rather than guessing a choice. Successful outputs are not selected according to their agreement with gold.

Format-only retry instruction

```
Return ONLY one JSON object with the single key "choice_id", whose value
is one of ["<id 1>", "<id 2>", "<id 3>", "<id 4>"].  Do not return
user/session records, memory updates, or rationale.
```

Memory output constraints. MIRIX uses structured output for memory construction. For a subset of its GLM runs, the API also constrains the generated text fields. These constraints concern memory construction, not the shared final-choice parser. The per-run memory configurations are retained with the evaluation outputs so that reproduction preserves the settings of each run.

Smaller Qwen models. The supplementary Qwen3-4B, 8B, 14B, and 32B experiments use the same system message and the same three user-message fields, but omit the final selection instruction after the alternatives. Thinking is disabled for these smaller Qwen models. Their earlier parser accepts a valid choice from JSON, a labeled `choice_id` expression, or a bare choice identifier. Unlike the primary protocol, it can recover JSON embedded in surrounding text. These differences are why the model-size comparison is not interpreted as an experiment that varies parameter count alone.

A.5.4 GLM-5.3-Flash: Results and Protocol

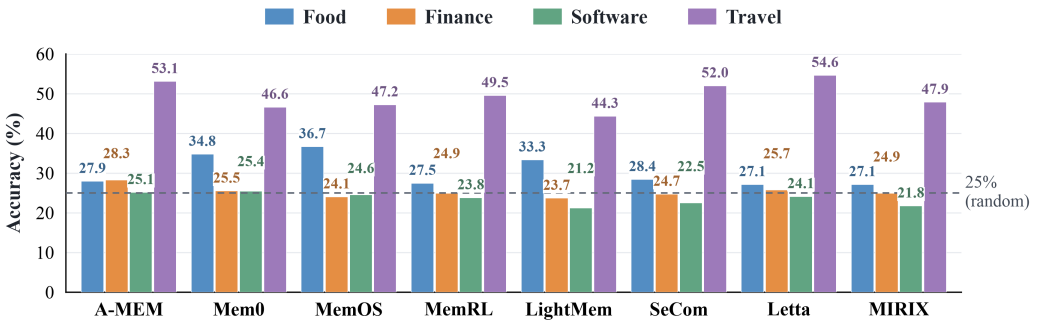


Figure 5: Memory-method accuracy with GLM-5.3-Flash across four domains. Bars show probe-level accuracy, and the dashed line marks 25% uniform random choice. Full results, including controls and retrieval baselines, appear in Table 3.

Figure 5 visualizes the eight memory methods from the GLM-5.3-Flash block of Table 3.

The model and inference settings are given in Appendix A.5.3. Pooled probe-level Wilson intervals, where reported, are distinct from user-cluster intervals and repeated-run variability and are not used for domain-level significance claims.

A.5.5 QWEN3-32B: EVIDENCE RETRIEVAL AND GROUNDED ANSWERS

Table 13 reports detailed metrics with Qwen3-32B for methods that expose source-session attribution. LightMem and MIRIX are excluded from this evidence-specific comparison because their memory outputs do not provide stable source-session lineage, not because their retrieved text is necessarily irrelevant. Their answer accuracies remain in the model-size comparison. Full Memory is a separate context-coverage control, rather than a top-five retrieval method.

Table 13: Detailed Qwen3-32B results. Acc. is four-domain macro accuracy. Evidence metrics use the eligible probes for each statistic. All scores are percentages except memory tokens. ER: evidence recall; CG: Clean Grounded; Joint A+E: correct answer with an evidence hit. †Full Memory uses context-wide evidence coverage, not top-five retrieval. Dashes indicate inapplicable metrics.

Method	Macro Acc.	ER†	Hit@5	NDCG@5	MRR@5	CG@5	Joint A+E	Memory tokens
Full Memory	29.88	36.83	—	—	—	—	22.69	33,131.1
Source-attributed top-five memory methods								
A-MEM	31.46	13.89	41.44	16.21	26.22	8.20	16.20	2,933.7
Mem0	28.52	11.17	39.30	13.63	24.58	7.00	12.92	107.2
MemOS	29.90	11.85	38.40	14.50	25.22	7.27	14.42	436.6
MemRL	31.25	12.67	37.77	14.99	24.68	5.49	14.36	1,735.1
SeCom	32.20	13.24	38.55	15.73	25.75	6.72	15.23	2,026.1
Letta	31.39	13.70	39.84	16.03	26.21	7.39	15.53	2,272.6

For a probe, let E_p be its annotated positive evidence sessions and S_p^5 the first five unique source-session identifiers attributed by the method. **ER@5** is the fraction of E_p covered by S_p^5, averaged over eligible probes. **Hit@5** measures whether at least one positive evidence session is retrieved. **NDCG@5** measures binary-relevance ranking quality with logarithmic rank discounting, and **MRR@5** is the reciprocal rank of the first positive hit, or zero when there is none. **Joint A+E** requires both a correct answer and an evidence hit. **Clean Grounded@5** additionally requires that none of the first five attributed sessions is annotated as nonbinding evidence. Each metric uses its annotation-eligible probes, so unavailable fields are not converted to zero. In particular, Clean Grounded@5 uses probes with nonbinding-evidence annotations and does not include Food. For Full Memory, ER measures evidence coverage in the supplied context and Joint A+E uses a hit anywhere in that context, without a five-session cutoff.

For a nonempty annotated evidence set E_p, the per-probe bound is $\mathrm{ER@5}_p \leq \min(5, |E_p|)/|E_p|$. Coverage therefore depends on the size of the annotated set, which is not itself a specification of a minimal sufficient reasoning chain. Joint A+E requires at least one evidence hit rather than complete support for every decision component. Clean Grounded@5 additionally imposes the absence of annotated nonbinding sessions from the attributed context; retrieving such a session can still be useful for determining that a proposal was never accepted. We retain these as complementary context-and-answer diagnostics rather than direct measures of improper personalization.

The ranked methods have ER@5 between 11.17% and 13.89%, despite Hit@5 between 37.77% and 41.44%. Retrieving some relevant evidence therefore does not imply reconstructing the full support for a habit. Joint A+E ranges from 12.92% to 16.20%, while Clean Grounded@5 ranges from 5.49% to 8.20% on its eligible subset. These measures diagnose different aspects of evidence access and use, and their different denominators preclude interpreting the difference as a causal error decomposition.

A.6 Illustrative Cases across Four Domains



Figure 6: Food case: recovering a default and its serving-context boundary from a CookDial-seeded lifeline (Jiang et al., 2023). History is condensed; the query and four choices are verbatim. The final panel is evaluator-only explanation.

Finance

540 sessions / cutoff Session 539

01 OBSERVED HISTORY

Model-visible / condensed excerpts

Sessions 008 ... 026 | Two separate ordered records 2021-02-04 / 2021-03-31

Review pair: (1) Unusual entries, then conclusion; (2) chronological timeline.
Card pair: (1) Verify activity, lock and escalate; (2) freeze, review securely, then decide replacement. Both remain open.

236 intervening sessions omitted

Sessions 263 ... 272 | The user resolves the earlier records 2023-03-12 / 2023-04-08

The user selects the first account-review route (Session 263) and the second card-protection route (Session 272). Each confirmation refers back to its own ordered pair.

5 intervening sessions omitted

Sessions 278 ... 284 | A replacement shortlist is still pending 2023-04-25 / 2023-05-14

Conclusion-first review and immediate card locking enter new replacement shortlists. The user keeps both decisions pending.

248 intervening sessions omitted

Sessions 533 ... 536 | Later confirmations change the state 2025-05-30 / 2025-06-08

In May-June 2025, the user confirms both replacements: conclusion-first review and immediate lock-and-escalate. These become effective after the query's target date.

02 CURRENT DECISION

Query: condensed

Prepare the two-workstream record that was correct as of 4:00 p.m. on May 16, 2023. Card ending 8748 went missing during the studio equipment invoice; also reconcile its latest activity. Use the policy state then in force, even if a later reference closed a pending review.

03 FOUR COUNTERFACTUAL ACTIONS

Choices: condensed

A **Card:** Lock, inspect, notify fraud.
Review: Conclusion before entries.

B **Card:** Freeze, securely review, decide replacement.
Review: Conclusion before entries.

C **Card:** Lock, inspect, notify fraud.
Review: Unusual entries before conclusion.

D **Card:** Freeze, securely review, decide replacement.
Review: Unusual entries before conclusion.

EVALUATOR-ONLY / GOLD AND INTERPRETATION

D / Resolve both workflows at the requested time

At the May 2023 target time, the effective records select the second card route and the first account-review route. The replacement shortlists are not yet resolved. Their 2025 confirmations are visible history, but cannot retroactively change the requested state.

Retain effective dates, not only the newest choice.

Figure 7: Finance case: two workflows resolved at a historical target time in a MultiDoGO-seeded lifeline (Peskov et al., 2019). History, query, and choices are condensed, preserving option order and the gold label. The final panel is evaluator-only explanation.

Software

540 sessions / cutoff Session 539

01 OBSERVED HISTORY

Model-visible / condensed excerpts

Session 276 | Diagnostic alternatives are left unresolved 2023-04-17

Diagnostics: (1) Minimal diagnostic bundle with consent; (2) clean-profile reproduction before setup changes. The user leaves the choice open.

3 intervening sessions omitted

Session 280 | Escalation has its own ordered alternatives 2023-04-30

Escalation: (1) One time-boxed attempt; (2) escalate when reproducible and evidence-complete. The user leaves the choice open.

149 intervening sessions omitted

Session 430 | A plausible assistant proposal is not ratified 2024-07-23

Assistant: Propose a last-known-good comparison.
User: Leave it unchosen; do not override a ratified workflow.

101 intervening sessions omitted

Sessions 532 ... 534 | Distant confirmations resolve both pairs 2025-05-24 / 2025-05-30

Session 532: Select the latter diagnostic route from its earlier record.
Session 534: Make the second escalation route the standing process.

02 CURRENT DECISION

Query: condensed

Diagnose Shared Inbox's E-307 failure in Seabird Office on macOS 15 and choose the next step for reproducible AUTH-29. A comparison worksheet is open; the deadline resembles a prior rushed case. Follow the user's established decisions.

03 FOUR COUNTERFACTUAL ACTIONS

Choices: condensed

A Clean-profile reproduction, then one time-boxed attempt before handoff.

B Clean-profile reproduction; escalate once reproducible and evidence-complete, without another attempt.

C Last-working-state comparison, then one time-boxed attempt before handoff.

D Last-working-state comparison; escalate once reproducible and evidence-complete, without another attempt.

EVALUATOR-ONLY / GOLD AND INTERPRETATION

B / Follow the two user-ratified routes

The second route in Session 276 is clean-profile reproduction; the second in Session 280 is escalation after reproducibility and evidence checks. Session 430 is an unaccepted assistant suggestion. A comparison worksheet or a rushed deadline does not make it an authorized replacement.

Preserve who endorsed an action, not just what was said.

Figure 8: Software case: recovering two user-confirmed workflows while rejecting an assistant-originated alternative in a MultiDoGO-seeded lifeline (Peskov et al., 2019). History, query, and choices are condensed. The final panel is evaluator-only explanation.

Travel

137 sessions / cutoff Session 136

01 OBSERVED HISTORY

Model-visible / condensed excerpts

Session 002 | Room equipment matters to the choice 2024-07-16

An alternative cannot guarantee a fridge; Perry Lane includes a coffee maker and mini-fridge. The user prioritizes Perry Lane.

8 intervening sessions omitted

Session 011 | Recreation amenities support another habit 2024-10-04

User: Accept the Hilton's extra cost and rides for quieter nights, earlier breakfast, and access to the pool, fitness center, and hot tub.

62 intervening sessions omitted

Session 074 | No automatic substitution 2026-02-08

Seabreeze has a coffee maker and fridge, but its workspace is not yet verified.
User: If it is only a narrow console, do not substitute Shoreline automatically. Come back to me.

28 intervening sessions omitted

Session 103 | A later choice reinforces the recreation default 2026-10-21

User: Choose the hotel move because there is time to use both the pool and gym; the hot tub tips the decision. A later closeout confirms that the hot tub was appreciated.

02 CURRENT DECISION

Query: condensed

Nadia and Daniel need a two-night lodge stay with parking and a festival shuttle. Aster has a coffee maker but no fridge; Birch has a fridge but no coffee maker. Terrace has pool, gym and hot tub; Solarium substitutes a sauna for the hot tub. All combinations meet the hard constraints with identical price/cancellation terms. Both evenings are free. Choose provisionally, pending confirmation.

03 FOUR COUNTERFACTUAL ACTIONS

Choices: condensed

A **Rooms:** Lead with Aster, retaining Birch for confirmation.
Recreation: Terrace.

B **Rooms:** Lead with Aster, retaining Birch for confirmation.
Recreation: Solarium.

C **Rooms:** Keep Aster and Birch jointly prioritized for confirmation.
Recreation: Solarium.

D **Rooms:** Keep Aster and Birch jointly prioritized for confirmation.
Recreation: Terrace.

EVALUATOR-ONLY / GOLD AND INTERPRETATION

D / Seek room confirmation; retain the recreation default

Neither room meets the paired equipment preference, and the history does not authorize trading a coffee maker against a fridge. Keep both for confirmation. The recreation habit still applies: Terrace preserves pool and gym access and adds the preferred hot tub.

An exception in one decision need not erase another habit.

Figure 9: Travel case: combining a room-selection exception with a recreation default in a Taskmaster-2-seeded lifeline (Byrne et al., 2019). History, query, and choices are condensed. The final panel is evaluator-only explanation.