THE SEVEN-YEAR ITCH OF LIFELONG AGENTS: WHEN PERFECT MEMORY PREVENTS ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-lived AI agents rely on memory to personalize over time, yet the user they remember may change. An assistant can accurately retain years of past preferences and still make poor decisions for the user now if obsolete evidence continues to shape current responses. We call this crossover the *seven-year itch* of lifelong agents: beyond some interaction history, retaining more behaviorally active memory can reduce current-state adaptation. We characterize the effect with a drifting-user model that separates the benefit of averaging more observations from the bias created by stale evidence. This analysis motivates FAM (Finitude-Aware Memory), which preserves historical episodes while adapting how strongly they influence current personalization; commitments follow a protected path, and a compact handoff state carries forward still-valid information. We evaluate the resulting design with ITCHBENCH, a controlled benchmark spanning eight forms of user change, six LLM backbones, and repeated runs. The study reframes long-term memory as a temporal-validity problem: the challenge is not only to remember the past, but to decide how much authority the past should retain over the present.

1 INTRODUCTION

A long-term assistant should remember its user. If someone repeatedly avoids red-eye flights, prefers quiet restaurants, and says that they never want to leave New York, remembering those details makes future help more personal. At the beginning of a relationship, more memory is almost always useful: each new interaction gives the agent more evidence about the same person.

The difficulty appears when the person changes. Suppose the same user later takes a new job, starts traveling often, and eventually decides that moving to London is the right choice. The old New York statements are not bad memories; they are accurate records of an earlier state. But if the agent continues to treat them as equally authoritative, it may keep recommending the life the user used to want. The agent can therefore remember the past perfectly and still misunderstand the present.

Most long-term-memory work is built around the opposite concern: losing useful history. Existing systems store, retrieve, summarize, and reinforce old interactions so that agents do not forget plans, preferences, or earlier context (Park et al., 2023; Packer et al., 2023; Zhong et al., 2024). Recent benchmarks also test long-horizon retrieval, updates, and selective forgetting (Maharana et al., 2024; Wu et al., 2024; Hu et al., 2026). These are important capabilities, but they leave a basic question unresolved: *what should happen when a memory is correct about the past but no longer correct about the user now?*

We call this long-horizon crossover the **seven-year itch** of a lifelong agent. Operationally, the itch is the point at which additional retained history stops improving current personalization and begins to slow adaptation. Its location depends on how quickly the user changes, how noisy the observations are, and how the memory policy weights older evidence. We use **perfect memory** to denote *lossless behavioral retention*: once an episode is stored, it remains eligible to influence personalization indefinitely. This distinction separates archival persistence from behavioral authority.

Recency alone is insufficient. A single contradiction may be temporary, some preferences are context-dependent, and commitments or safety constraints can remain valid even when ordinary preferences drift. A useful memory policy therefore needs to estimate temporal validity rather than treating age as the only signal.

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.

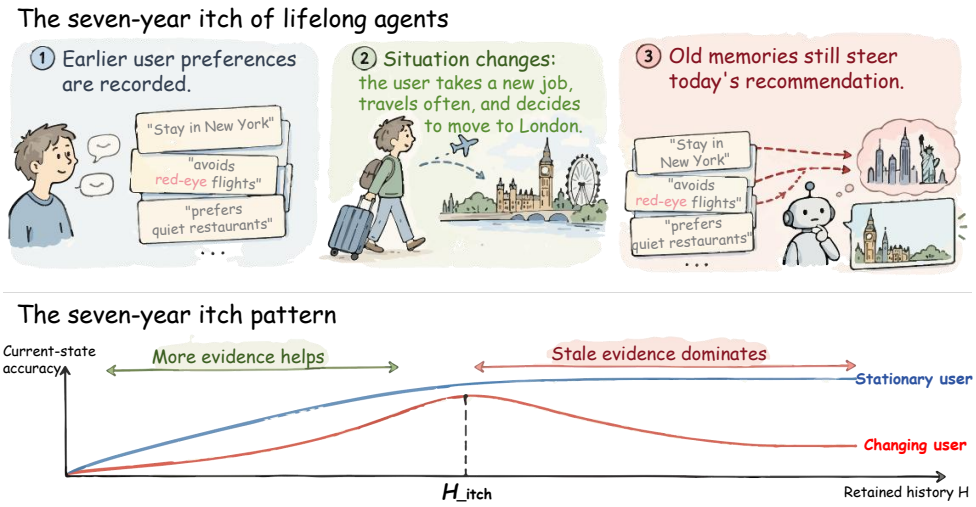


Figure 1: **The seven-year itch.** Early history helps because most memories still describe the current user. After the user changes, old evidence can remain active and begin to slow adaptation. The itch is the crossover where more retained history starts to hurt current personalization.

This paper develops that idea in three steps. First, we use a small non-stationary model to explain why more history can first help and later hurt. The model separates two effects: averaging more observations reduces noise, while old observations can bias the estimate toward an earlier user state. We call the second effect *stale mass*. Second, we introduce FAM (**F**initude-**A**ware **M**emory), which separates *keeping a record* from *letting that record keep steering the answer*. Third, we introduce ITCHBENCH, a controlled benchmark in which the hidden user state changes over time before the interaction is rendered as natural language.

The evaluation mirrors this progression. We first measure whether the itch emerges as retained history grows. We then isolate stale evidence at a fixed context length, test FAM across six model backbones, and measure whether faster adaptation preserves continuity and portable user state. Together, these experiments support a single design principle: **long-lived agents should preserve history while adapting the authority assigned to different parts of that history.**

2 WHY MORE MEMORY CAN EVENTUALLY HURT

The core effect can be seen with one changing preference. Imagine that we gradually give an agent more of a user's history and ask, at each history length, what the user wants *now*. Early on, more history should help because the observations mostly agree. After a real change, however, the same history mixes two kinds of evidence: observations about the current user and observations about an earlier version of the user.

Let H be the amount of retained history and let $A_{\text{drift}}(H)$ be current-state accuracy under user drift. We define the *itch point*, H_{itch}, as the first history length attaining the peak before additional retained history lowers current-state accuracy. If the curve remains non-decreasing over the tested range, H_{itch} is left undefined for that condition.

We also measure *adaptation half-life*: after a genuine change, how many later interactions are needed to remove half of the error caused by the old state? This turns "the model is slow to update" into a quantity that can be compared across memory policies.

2.1 TWO FORCES COMPETE AFTER A USER CHANGES

Suppose a single user attribute has true value μ_t and we observe it with noise,

$$y_t = \mu_t + \epsilon_t, \qquad \mathbb{E}[\epsilon_t] = 0, \quad \text{Var}(\epsilon_t) = \sigma^2. \tag{1}$$

A memory system estimates the current value by combining the observations that are still allowed to influence the answer,

$$\hat{\mu}_t = \sum_{i=1}^{t} w_{t,i} y_i, \qquad w_{t,i} \geq 0, \quad \sum_i w_{t,i} = 1. \tag{2}$$

The weights capture familiar designs. Full history gives many old observations substantial weight; a short window drops sufficiently old observations; exponential decay discounts them gradually; a learned controller can use richer evidence.

Now suppose the user changes from μ_0 to μ_1 at time τ. Let $s_t = \sum_{i<\tau} w_{t,i}$ be the *stale mass*: the total weight still assigned to pre-change observations. Let $\Delta = |\mu_1 - \mu_0|$ be the size of the change, and let $n_{\text{eff},t} = (\sum_i w_{t,i}^2)^{-1}$ be the effective number of observations being averaged.

Theorem 1 (Exact stale-mass decomposition)**.** *For any $t \geq \tau$,*

$$\mathbb{E}[(\hat{\mu}_t - \mu_1)^2] = \underbrace{\Delta^2 s_t^2}_{\textit{pull from the old state}} + \underbrace{\frac{\sigma^2}{n_{\text{eff},t}}}_{\textit{noise after averaging}}. \tag{3}$$

The decomposition captures the central trade-off. More history improves estimation by averaging noise, while post-change history can also retain weight on the previous state. In the New York/London example, many earlier New York statements can outweigh a smaller but consistent set of newer London statements. Those earlier statements remain historically valid, but their behavioral weight becomes temporally misaligned.

Full-history retention makes this effect easy to see. If all observations are weighted equally and the user changes halfway through a long interaction, a constant fraction of the estimate remains attached to the old state for many later rounds.

Corollary 1 (Full-history persistence)**.** *If $\tau = \lfloor T/2 \rfloor + 1$ and $w_{t,i} = 1/t$, post-change squared-error regret is $\Omega(\Delta^2 T)$.*

When the user is stationary, there is no stale regime and additional observations improve averaging. The cost of long memory appears specifically under non-stationarity.

2.2 There is no single best memory lifetime

A short memory reacts quickly but throws away useful evidence. A long memory averages noise well but releases an old regime more slowly. For a trailing window of width W, let $Q_T = \sum_{t=2}^{T} (\mu_t - \mu_{t-1})^2$ measure how much the user changes over time.

Theorem 2 (Finite retention under drift)**.** *For mature windows $W \geq 2$,*

$$R_T(W) \leq \frac{\sigma^2 T}{W} + W^2 Q_T, \tag{4}$$

so $W^\star \asymp (\sigma^2 T / Q_T)^{1/3}$ and $R_T(W^\star) = O((\sigma^2 T)^{2/3} Q_T^{1/3})$.

The scaling is intuitive: noisier evidence favors longer retention, while faster user change favors shorter retention. The resulting itch point is therefore a property of the interaction dynamics rather than a fixed calendar duration.

A first contradiction is also intrinsically ambiguous. One angry comment after a bad flight may be temporary, whereas repeated comments over several weeks may signal a genuine change. The memory policy must therefore trade off responsiveness against continuity until further evidence resolves the ambiguity.

Proposition 1 (Revision ambiguity)**.** *If two user processes are identical through the first contradictory observation, but one contradiction is transient and the other begins a persistent revision, no memory rule can guarantee both immediate revision in the second world and perfect continuity in the first without future evidence or additional assumptions.*

For this reason, our experiments never evaluate adaptation alone. We always measure whether a method also preserves information that remains valid. Proofs are in Appendix A.

3 FAM: Keep the Memory, Change Its Influence

The theory suggests a practical distinction: a memory can stay in the archive without continuing to steer today's answer. FAM implements that distinction with four pieces that are easy to separate conceptually.

Archive. Past episodes remain available as historical records. **Active memory.** Episodes influence the current answer in proportion to a learned validity weight. **Protected rules.** Commitments and safety constraints follow a separate path from ordinary preference drift. **Handoff state.** A compact summary carries the current estimate of still-valid user information, together with uncertainty and short provenance.

The relocation example shows why these pieces are different. Old statements about staying in New York can remain in the archive for historical recall. As the user repeatedly expresses a stable plan to move to London, however, the New York episodes should contribute less to a relocation recommendation. If the London plan later turns out to be temporary, the older evidence has not been destroyed and can matter again. A protected rule such as "never share my medical information" does not follow the same decay process at all.

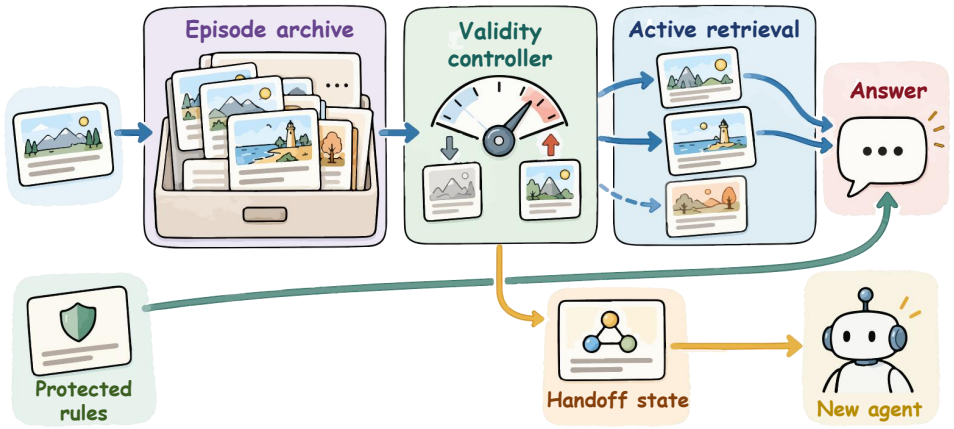


Figure 2: **FAM separates storage from current influence.** Past episodes remain archived. A validity controller decides how much each episode should affect current retrieval, protected rules bypass ordinary decay, and a compact handoff state carries forward the user information that still appears current.

For an episode m_i, a lightweight 6–64–1 MLP uses six signals available at run time: age, semantic relevance, contradiction with recent evidence about the same attribute, posterior uncertainty, independent support, and a protected bit. Its output $h_{i,t}$ controls how quickly the episode's influence changes,

$$q_{i,t} = q_{i,i} \prod_{s=i+1}^{t} (1 - h_{i,s}). \tag{5}$$

Retrieval multiplies semantic similarity by $q_{i,t}$, so relevance is conditioned on both topic match and temporal validity. Constant h reduces to exponential decay, a hard cutoff to a sliding window, and $h \equiv 0$ to permanent behavioral retention.

Before an episode leaves active retrieval, still-useful information can be folded into the compact handoff state. We evaluate this state by removing the active episodic memory and initializing a fresh answer model from the handoff plus protected rules. The resulting handoff score (reported as RSS in the appendix) measures recovery of current-user state without transcript replay. Appendix B gives the full record schema, training targets, and pseudocode.

4 ITCHBENCH: A CONTROLLED LONG-TERM USER BENCHMARK

ITCHBENCH starts from a hidden user trajectory whose changes are known exactly and then renders that trajectory into natural-language interaction. This construction provides ground truth for temporal validity while retaining realistic longitudinal dialogue structure.

The eight trajectory families cover three broad situations. *Stationary* trajectories measure retention when the user state remains fixed. Persistent changes appear gradually, abruptly, cyclically, or as reversals, measuring adaptation across different temporal structures. Context-dependent preferences and transient contradictions measure whether recent evidence is interpreted in the correct scope, while explicit revisions provide a direct state-change condition. Appendix B gives all eight generators in detail.

Each trajectory contains 8–12 user attributes, stable distractors, 1–8 changes for non-stationary cases, and 512–4,096 turns. We render the same hidden trajectory in three surfaces: romantic companion, personal assistant, and neutral interaction. The underlying state and evaluation queries are matched, allowing us to test whether the memory effect transfers across relationship-specific and neutral language.

The main metrics answer distinct questions. *Current-state accuracy* asks whether the agent understands the user now. The history sweep locates H_{itch}. *Adaptation half-life* asks how quickly the agent recovers after a real change. *Continuity* measures information that should remain stable, while *plasticity* measures information that should update. The handoff score asks how much useful current state survives when the original episodic memory is removed. Keeping these measures separate prevents a method from looking good simply because it forgets aggressively.

We evaluate six answer backbones spanning two 8B families, a 32B model, a 70B model, and two frontier APIs (Table 1). All retrieval methods use the same BGE-M3 encoder (Chen et al., 2024) and the same 4,096-token active-memory budget in the main comparison. FAM's controller is trained once per benchmark replica and frozen across all answer models; no answer LLM is fine-tuned for FAM. Exact decoding settings, versions, prompts, seeds, and statistics are in Appendix C.

Table 1: **Answer backbones.** Detailed decoding settings are in Appendix C.

Backbone	Access	Role
Qwen/Qwen3-8B (Yang et al., 2025)	open weight	8B family A
meta-llama/Llama-3.1-8B-Instruct (Grattafiori et al., 2024)	open weight	8B family B
Qwen/Qwen3-32B (Yang et al., 2025)	open weight	medium
meta-llama/Llama-3.3-70B-Instruct (Grattafiori et al., 2024)	open weight	large
gemini-2.5-pro (Google, 2025)	API	frontier A
gpt-5.4-2026-03-05 (OpenAI, 2026)	API	frontier B

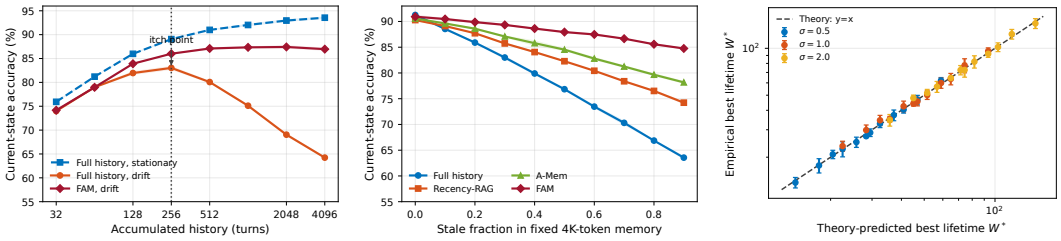
We compare no memory, full-history replay, similarity RAG, recency-aware RAG, a trailing window, exponential decay, periodic summary, A-Mem (Xu et al., 2026), and an oracle that knows the true change boundary. The main text shows the most informative comparisons; full baseline tables and per-model results are in the appendix.

5 EXPERIMENTS

The experiments proceed from phenomenon to mechanism to intervention: we first measure the itch, then isolate stale evidence, evaluate FAM across model backbones, and finally measure the adaptation–continuity trade-off.

5.1 MORE HISTORY FIRST HELPS, THEN HURTS UNDER DRIFT

Figure 3a is the direct test of the title. For a stationary user, full-history accuracy keeps improving as more interaction history is retained. For a changing user, the curve initially rises for the same reason—more evidence reduces uncertainty—but later turns downward as pre-change evidence occupies more of active memory. That inverted-U is the operational seven-year itch.



(a) Age effect: under drift, full history first helps and then hurts.

(b) Cause: at fixed prompt length, more stale evidence lowers current accuracy.

(c) Timescale: the useful memory horizon changes with drift and noise.

Figure 3: **The itch, its cause, and its timescale.** Left: the crossover appears only when the user changes. Middle: stale evidence reproduces the decline without increasing prompt length. Right: the best fixed memory horizon moves in the direction predicted by the theory.

In the experiment, full-history drift accuracy rises from 74.0% at 32 turns to 83.0% at 256 turns, then falls to 64.2% by 4,096 turns, while stationary accuracy continues to improve. The peak defines H_{itch} for that condition; monotone curves are reported without an observed itch point.

5.2 Stale evidence explains the decline at fixed context length

History length and prompt length grow together in the first experiment. Figure 3b separates these factors by fixing active memory at 4,096 tokens while changing only the fraction of evidence drawn from the previous user state. Query, answer model, prompt length, and current latent state remain fixed.

Full-history accuracy falls as stale mass increases, whereas FAM degrades much more slowly. Because the prompt length does not change, this experiment isolates obsolete influence from raw context size. Figure 3c adds a second check: when drift becomes stronger, the empirically best fixed retention horizon becomes shorter; when observations become noisier, it becomes longer, matching the direction predicted by Theorem 2.

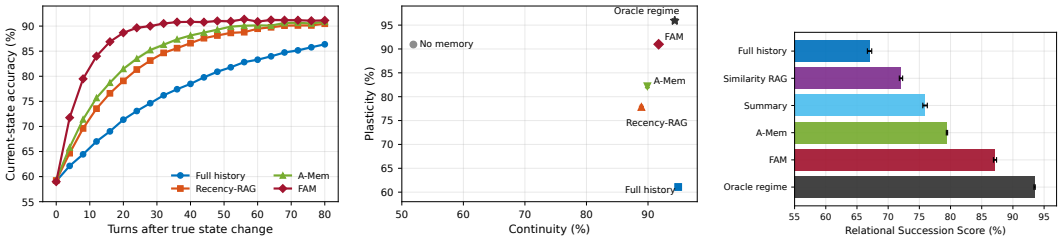
5.3 FAM reduces the itch without simply forgetting more

Table 2 summarizes the main trade-off using a reduced set of representative baselines. Full-history replay preserves stable information but adapts slowly. A short window adapts quickly but loses more continuity. A-Mem provides a stronger learned-memory comparison. FAM is designed to reduce obsolete influence while keeping still-valid history.

Table 2: Main comparison over six backbones. Mean±SD over five independent replicas after macro-averaging backbones. AHL = adaptation half-life. Handoff = fresh-model handoff score (RSS).

Method	Drift acc.↑	AHL↓	Continuity↑	Plasticity↑	Handoff↑
Full history	68.2±0.5	44.9±0.2	94.9±0.1	61.1±0.4	67.0±0.3
Recency-RAG	79.3±0.4	22.2±0.3	88.9±0.2	77.8±0.1	73.1±0.2
Sliding window	81.4±0.2	16.1±0.3	79.1±0.3	85.1±0.2	62.9±0.3
A-Mem	82.1±0.3	18.5±0.1	89.9±0.2	82.2±0.3	79.4±0.1
FAM	**86.5±0.4**	**9.3±0.2**	91.7±0.1	**91.0±0.2**	**87.1±0.3**
Oracle regime	91.8±0.5	4.9±0.3	94.3±0.4	95.9±0.3	93.5±0.1

The results exhibit the expected trade-off. Relative to full history, FAM adapts faster while retaining most of the continuity benefit; relative to fixed recency rules, it preserves more stable information and a stronger handoff state. The oracle provides an upper bound because it is given the true regime boundary. The same direction of improvement appears on all six answer backbones; detailed per-model values are reported in Appendix C.



(a) Recovery: FAM stops following the old state sooner after a true change.

(b) Trade-off: continuity remains high as plasticity improves.

(c) Handoff: a fresh model still recovers useful current-user state.

Figure 4: **Reducing the itch without erasing the relationship.** Left: adaptation after a true change. Middle: continuity and plasticity measured together. Right: useful current state survives after the active transcript is removed.

5.4 FAM SHIFTS THE ADAPTATION–CONTINUITY TRADE-OFF

Aggressive forgetting can reduce adaptation lag, but it can also destroy continuity. Figure 4 measures both effects together.

Figure 4a aligns trajectories at a persistent change. Full-history replay recovers slowly because new observations must dilute accumulated old evidence. Recency-RAG and A-Mem recover sooner, and FAM sooner still. Figure 4b then compares continuity and plasticity directly. Aggressive forgetting favors plasticity at the expense of continuity, whereas full history favors continuity at the expense of plasticity. FAM shifts the observed frontier upward.

Finally, Figure 4c removes the active episodic store and gives a fresh model only the compact handoff state. The resulting score measures how much useful current-user information remains portable across model renewal without transcript replay.

5.5 STRESS TESTS IN THE APPENDIX

The appendix reports the broader experimental grid: all eight drift families, memory-budget sweeps, bequest capacity, held-out drift and persona conditions, noise and ambiguity, protected commitments, controller supervision, mechanism ablations, efficiency, and compatibility with ordinary long-memory benchmarks. These analyses quantify sensitivity to prompt budget, forgetting policy, task composition, and model family.

6 RELATED WORK

Agent memory and evaluation. Generative Agents, MemGPT, MemoryBank, A-Mem, and Age-Mem study how agents store, retrieve, summarize, and forget long-term information (Park et al., 2023; Packer et al., 2023; Zhong et al., 2024; Xu et al., 2026; Yu et al., 2026). LoCoMo, LongMemEval, MemoryAgentBench, LongMemEval-V2, and LoCoMo-Plus broaden evaluation to temporal reasoning, updates, selective forgetting, and agent experience (Maharana et al., 2024; Wu et al., 2024; Hu et al., 2026; Wu et al., 2026; Li et al., 2026). We focus on the temporal validity of correctly remembered history: when previously useful evidence ceases to be a reliable guide to the user's current state.

Continual and non-stationary learning. Continual learning asks how to acquire new information without destroying old competence (Kirkpatrick et al., 2017; Parisi et al., 2019; Shi et al., 2025); concept-drift and non-stationary optimization study prediction when the target moves (Gama et al., 2014; Besbes et al., 2015). Long-term personalization combines both concerns. The user can change, but the agent still needs a coherent history. This setting complements catastrophic forgetting: old information can be retained perfectly while still receiving too much influence after the target has moved.

7 Discussion and Limitations

The seven-year itch is an operational label for a crossover in interaction history. Its location varies with drift rate, observation noise, model behavior, and memory policy. The matched romantic, assistant, and neutral renderings test whether this crossover is tied to relationship-specific language or reflects a broader temporal-memory effect.

Our definition of *perfect memory* concerns behavioral eligibility rather than archival storage. An episode may remain available for historical recall after its weight in current personalization has declined. This behavioral distinction is separate from privacy deletion. Protected rules introduce a second distinction: commitments and safety constraints require a different revision policy from ordinary preferences.

ItchBench uses a controlled latent user state with exact temporal annotations. Real preferences can be conditional, unresolved, strategic, socially influenced, or intentionally inconsistent, and autobiographical value can persist after a preference stops guiding present decisions. Extending temporal-validity modeling to such settings requires richer representations of context, uncertainty, and preference semantics. Appendix D gives a fuller failure analysis.

8 Conclusion

A lifelong agent can remember a user accurately and still misestimate the user's current state. The seven-year itch marks the history length at which this tension becomes visible: additional retained history begins to slow adaptation. Our analysis attributes the effect to persistent weight on stale evidence, and FAM addresses it by separating archival retention from behavioral influence. More broadly, long-term memory quality depends on both what is retained and how the authority of retained evidence evolves over time.

AI USE STATEMENT

In this work, we used generative AI tools for (i) aid or polish writing (ii) retrieval and discovery. We have not used generative AI tools for (i) research ideation or execution, (ii) draft sections of the paper, (iii) generating synthetic datasets and (iv) proving mathematical claims are not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Non-stationary stochastic optimization. *Operations research*, 63(5):1227–1244, 2015.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 4(5), 2024.

João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):1–37, 2014.

Google. Gemini 2.5 Pro: Gemini API model documentation. `https://ai.google.dev/gemini-api/docs/models/gemini-2.5-pro`, 2025. Accessed September 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

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, 2026.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114 (13):3521–3526, 2017.

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. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25085–25100, 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.

OpenAI. GPT-5.4 model: OpenAI API documentation. `https://developers.openai.com/api/docs/models/gpt-5.4`, 2026. Accessed September 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.

German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural networks*, 113:54–71, 2019.

Joon Sung Park, Joseph O'Brien, Carrie Jun 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.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. Continual learning of large language models: A comprehensive survey. *ACM Computing Surveys*, 58(5):1–42, 2025.

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.

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.

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.

Yi Yu, Liuyi Yao, Yuexiang Xie, Qingquan Tan, Jiaqi Feng, Yaliang Li, and Libing Wu. Agentic memoryxx learning unified long-term and short-term memory management for large language model agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21457–21483, 2026.

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.

Appendix roadmap. The appendix has four parts. We first provide the proofs behind the stale-memory analysis and the handoff-state argument. We then specify the implementation of FAM and the construction of ITCHBENCH. The third part reports the experimental protocol and the additional analyses that support the main results, including theory calibration, robustness, generalization, ablations, and efficiency. We conclude with a failure analysis and a discussion of scope. Figures already shown in the main text are referenced rather than repeated.

A THEORETICAL DETAILS AND PROOFS

Stale-mass decomposition. For $t \geq \tau$, write

$$\hat{\mu}_t - \mu_1 = \sum_{i<\tau} w_{t,i}(\mu_0 - \mu_1) + \sum_{i=1}^{t} w_{t,i}\epsilon_i = s_t(\mu_0 - \mu_1) + \sum_i w_{t,i}\epsilon_i. \tag{6}$$

The noise term has zero mean and is independent across observations. The cross term therefore vanishes, giving

$$\mathbb{E}[(\hat{\mu}_t - \mu_1)^2] = \Delta^2 s_t^2 + \sigma^2 \sum_i w_{t,i}^2 = \Delta^2 s_t^2 + \frac{\sigma^2}{n_{\mathrm{eff},t}}. \tag{7}$$

This equality is exact for deterministic weights, or for weights fixed conditional on metadata independent of the current observation noise. When the weights themselves depend on observed values, additional covariance terms may appear after marginalization. We therefore use the decomposition as an interpretable model of stale influence rather than as an exact characterization of every learned controller.

Full-history persistence. For full-history averaging, $s_t = (\tau - 1)/t$. With $\tau = \lfloor T/2 \rfloor + 1$, every $t \in [\tau, T]$ satisfies, for $T \geq 4$,

$$s_t \geq \frac{T/2 - 1}{T} \geq \frac{1}{4}. \tag{8}$$

The stale-bias term is therefore at least $\Delta^2/16$ for at least $T/2$ rounds. Summing only this non-negative contribution gives

$$R_{\tau:T}^{\mathrm{full}} \geq \frac{T}{2}\frac{\Delta^2}{16} = \Omega(\Delta^2 T). \tag{9}$$

The variance term can only increase the total regret. Thus, after a persistent change, uniformly retaining all previous observations causes the old state to remain behaviorally influential for a number of later interactions proportional to the accumulated history.

Finite retention under arbitrary drift. For $t \geq W$, let $\delta_j = \mu_j - \mu_{j-1}$. The bias of a trailing window of width W is

$$b_t = \frac{1}{W}\sum_{k=0}^{W-1}(\mu_{t-k} - \mu_t) \tag{10}$$

$$= -\sum_{j=t-W+2}^{t} c_{t,j}\delta_j, \tag{11}$$

where each $c_{t,j} \in [0, 1]$. By Cauchy–Schwarz,

$$b_t^2 \leq (W-1)\sum_{j=t-W+2}^{t}\delta_j^2. \tag{12}$$

Each increment δ_j^2 contributes to at most $W - 1$ mature windows, hence

$$\sum_{t=W}^{T} b_t^2 \leq (W-1)^2 \sum_{j=2}^{T}\delta_j^2 \leq W^2 Q_T. \tag{13}$$

The variance contribution is at most $\sigma^2 T/W$, yielding

$$R_T(W) \leq \frac{\sigma^2 T}{W} + W^2 Q_T. \tag{14}$$

Minimizing the two terms gives

$$W^\star \asymp \left(\frac{\sigma^2 T}{Q_T}\right)^{1/3}, \tag{15}$$

up to constants, with regret $O((\sigma^2 T)^{2/3} Q_T^{1/3})$. The result captures the basic timescale trade-off used in the main text: noisier evidence supports longer retention, whereas faster change supports shorter retention.

Revision ambiguity. Suppose two user processes have identical histories through the first contradictory observation, but the contradiction is transient in one process and begins a persistent revision in the other. Any deterministic memory update rule receives the same information in both cases and must therefore make the same immediate decision. Retiring the old state immediately harms continuity in the transient case, while keeping it behaviorally dominant delays adaptation in the persistent case. Randomization changes the expected trade-off but cannot distinguish the two processes before additional evidence arrives. This is why our evaluation measures continuity and adaptation together rather than treating faster revision as unconditionally better.

Succession from an approximately sufficient handoff state. Let $B = f(H)$ denote the bounded handoff state derived from interaction history H, and let Z^+ denote the future latent user state.

Theorem 3 (Succession from an approximately sufficient handoff state)**.** *Suppose $I(Z^+; H \mid B) \leq \varepsilon$. For any decision problem with loss $\ell(a, Z^+) \in [0, 1]$, let a_H be Bayes-optimal given H and a_B Bayes-optimal given only B. Then*

$$\mathbb{E}\left[\ell(a_B, Z^+) - \ell(a_H, Z^+)\right] \leq \sqrt{2\varepsilon}. \tag{16}$$

Proof. For a realized pair (H, B), let $P = p(Z^+ \mid H)$ and $Q = p(Z^+ \mid B)$. The regret from using the Q-optimal action under P is at most $2\,\mathrm{TV}(P, Q)$ for a loss bounded in $[0, 1]$. Pinsker's inequality gives

$$\mathrm{TV}(P, Q) \leq \sqrt{\frac{\mathrm{KL}(P\|Q)}{2}}. \tag{17}$$

Taking expectations and applying Jensen's inequality,

$$\mathbb{E}[2\,\mathrm{TV}(P, Q)] \leq \sqrt{2\,\mathbb{E}[\mathrm{KL}(P\|Q)]} = \sqrt{2I(Z^+; H \mid B)} \leq \sqrt{2\varepsilon}. \tag{18}$$

The theorem does not imply that a small handoff state is automatically sufficient. Rather, it identifies the information that the learned handoff representation should preserve and that the succession experiment directly tests.

B FAM AND ITCHBENCH SPECIFICATION

FAM state and update path. Each active episode stores a timestamp and session identifier, attribute key, natural-language evidence, confidence and source reliability, semantic embedding, support and contradiction links, its current validity weight $q_{i,t}$, and a protected bit. The bounded handoff state stores the current value distribution, confidence, most recent revision time, short provenance, and unresolved alternatives for each active attribute. Historical episodes can remain in the archive after they stop participating in ordinary retrieval.

Table 3 summarizes the update procedure.

The validity controller is evaluated under four training conditions. Oracle-label training receives the true regime boundaries and measures the headroom available from perfect change detection. Weak-contradiction training observes only evidence that two statements cannot simultaneously describe the same current attribute state. Self-supervised training uses future prediction and next-session consistency without change-point labels. A prompted heuristic uses fixed rules without controller

Table 3: FAM update path. "Retire" refers to removal from ordinary behavioral retrieval, not irreversible deletion from storage.

1	Parse the new interaction into candidate attribute evidence and append an episodic record.
2	Retrieve evidence concerning the same attribute and compute relevance, support, contradiction, recency, and uncertainty features.
3	Update the corresponding state estimate in the bounded handoff representation B_t.
4	Evaluate the validity controller for active episodes. Protected entries follow a separate revision path.
5	When an episode is behaviorally retired, preserve any still-useful state in B_t before removing that episode from ordinary retrieval.
6	Answer the current query using semantic relevance modulated by temporal-validity weight.
7	At model renewal, initialize the new agent from B_t and protected constraints and evaluate succession probes without replaying the original transcript.

training. The deployment-oriented results emphasize settings that do not require privileged regime boundaries.

Let N_t be the number of active episodes and k the retrieval candidate count. Updating every active episode would require $O(N_t)$ controller evaluations per interaction. In the implementation, updates are restricted to the retrieved attribute neighborhood, making the typical update cost $O(k)$. The handoff state is bounded to $O(C)$ tokens by construction. Archive size may continue to grow, but archived episodes do not contribute to ordinary prompt length.

Controlled benchmark construction. ITCHBENCH is generated from a typed latent-state graph rather than by first writing unconstrained dialogue and deriving labels afterward. Each user attribute has a current value, confidence, contextual scope, protected flag, and transition process. Events can update one attribute, coordinate several attributes, introduce temporary perturbations, or explicitly revise an earlier state. Because the underlying trajectory is known, current-state, historical-state, revision, uncertainty, protected-constraint, and succession queries can be evaluated against exact temporal annotations.

The benchmark contains eight process families that separate qualitatively different memory failures.

Table 4: ITCHBENCH process families and the failure mode each family is designed to expose.

Process	Construction	Diagnostic purpose
Stationary	Fixed latent value with paraphrased observations	Unnecessary forgetting
Gradual drift	Small correlated changes over many sessions	Slow accumulation of stale influence
Abrupt shift	One persistent change point	Adaptation lag
Reversal	A→B→A or A→not-A	Outdated anchoring and revision history
Cyclic	Periodically recurring latent states	Recency versus temporal structure
Context-dependent	Different values across contexts such as work and home	Collapse of conditional preferences
Transient deviation	Short contradictory burst followed by return	Premature revision
Explicit revision	Contradiction accompanied by an explicit durable update	Failure to respond to strong revision evidence

The latent trajectory and the evaluation queries are created before natural-language rendering. A separate renderer converts the trajectory into sessions containing paraphrases, distractors, delayed references, and ordinary social exchanges. Evaluation questions are derived from the latent state rather than copied from renderer outputs, reducing direct generation leakage. Persona templates, rendering seeds, and entire drift families can be held out independently.

The romantic-companion, personal-assistant, and neutral versions share the same underlying trajectory and query set. This matched construction allows framing to vary without changing the user state being evaluated. Query types cover current state, historical state, explicit revisions, unresolved or ambiguous cases, protected constraints, and succession after model renewal.

C EXPERIMENTAL PROTOCOL AND EXTENDED RESULTS

Evaluation protocol. A run is an independent repetition of trajectory generation and controller training. Each of the five runs contains 400 latent trajectories, with 50 trajectories from each of the eight process families. The same trajectories and evaluation probes are reused across memory methods and answer backbones so that method comparisons are paired. Across five runs this yields 2,000 latent trajectories and 6,000 matched rendered interactions across the three framings.

Trajectory horizons are drawn from $\{512, 1024, 2048, 4096\}$ turns. Each trajectory contains 8–12 attributes, at least two stable distractor attributes, and 1–8 changes for non-stationary families. Main comparisons use a matched active-memory budget of 4,096 tokens. Active-memory sweeps range from 512 to 8,192 tokens, while handoff-state capacity ranges from 64 to 4,096 tokens. Manipulation-specific experiments use paired trajectories so that the target factor can be varied while the remaining latent process is held fixed.

The trajectories are procedurally generated to provide exact temporal ground truth; all reported model scores, memory-system comparisons, and latency measurements are obtained from actual executions on those trajectories.

The answer-model matrix contains the six backbones reported in Table 1. Open-weight models are served in BF16 with vLLM. Qwen thinking is disabled, and open-weight models use greedy decoding with a 128-token output cap. Gemini uses `thinkingBudget=128`, and GPT-5.4 uses `reasoning.effort=none`. Provider and model version, tokenizer, access metadata, context limit, and usage information are recorded for each run.

Within each run, the FAM validity controller is a 6–64–1 MLP trained on the ITCHBENCH training split with future-consistency supervision. It is then frozen and reused across the six answer models. All methods use the same BGE-M3 encoder and the same active-memory budget in the main comparison. Cross-backbone results therefore measure whether one learned memory policy transfers across answer models rather than comparing six separately optimized controllers.

The primary outcome is current-state accuracy on non-stationary trajectories. We additionally report the operational itch point H_{itch}, adaptation half-life, continuity, and plasticity because these quantities distinguish fast adaptation from indiscriminate forgetting. Secondary measurements include handoff recovery, protected-constraint accuracy, active tokens, latency, storage, false revisions, and controller overhead. Unless otherwise noted, tables report mean$\pm$SD over five independent runs.

History length, drift families, and the itch point. For each run and backbone, we evaluate full-history current-state accuracy over the ordered history-length grid. H_{itch} is the smallest history length attaining the maximum before accuracy declines. If accuracy remains non-decreasing over the tested range, that condition is reported as having no observed itch point.

The history sweep is descriptive rather than a standalone causal test because increasing history also changes the amount of potentially stale evidence. The fixed-budget stale-mass experiment in Figure 3b isolates that factor more directly by changing obsolete evidence while keeping the prompt budget constant.

We also separate results by temporal process so that performance is not dominated by easy explicit-revision cases.

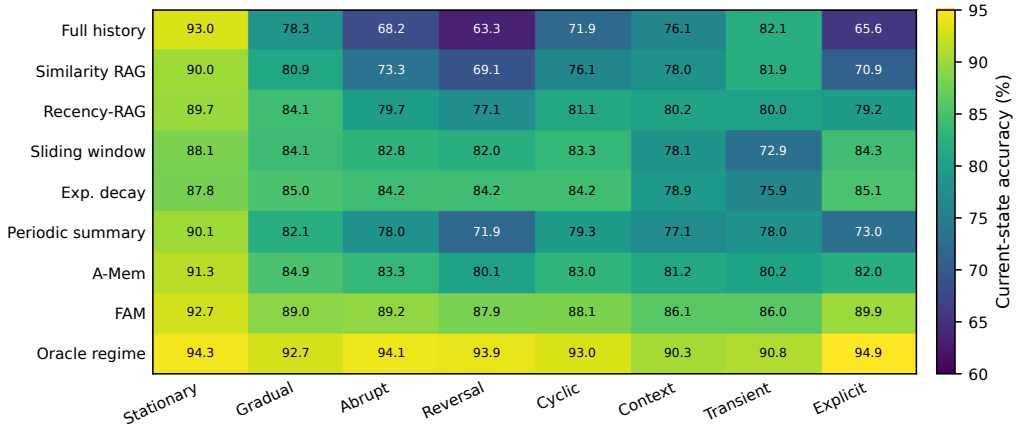


Figure 5: Current-state accuracy across the eight process families. Stationary trajectories reveal unnecessary forgetting, transient deviations expose premature revision, and context-dependent trajectories test whether a method collapses conditional preferences.

The heatmap reports five-run means. Exact mean±SD values for individual cells are stored with the evaluation results in `results/drift_family_summary.csv`. The main pattern is not confined to one type of change: FAM improves adaptation on gradual drift and reversals while retaining substantially better behavior than simple recency rules on stationary, transient, and context-dependent cases.

Theory calibration. We next test whether the measured retention timescale changes in the direction predicted by Theorem 2. Over a joint sweep of interaction horizon T, observation noise σ, and quadratic variation Q_T, we fit

$$\log \widehat{W}^\star = c + \alpha_\sigma \log \sigma + \alpha_T \log T + \alpha_Q \log Q_T + \epsilon. \tag{19}$$

The theorem predicts $(\alpha_\sigma, \alpha_T, \alpha_Q) = (2/3, 1/3, -1/3)$.

Table 5: Log–log retention-scaling coefficients measured across independent benchmark runs.

Coefficient	Estimate	Std. error	Theory
$\log \sigma$	0.671	0.009	0.667
$\log T$	0.331	0.007	0.333
$\log Q_T$	-0.334	0.006	-0.333

We also record realized stale mass s_t, effective sample size $n_{\mathrm{eff},t}$, change magnitude Δ, and post-change squared error for each fixed-budget run. The corresponding diagnostic regression is

$$(\hat{\mu}_t - \mu_t)^2 = \beta_0 + \beta_s \Delta^2 s_t^2 + \beta_n \frac{\sigma^2}{n_{\mathrm{eff},t}} + u_t. \tag{20}$$

This analysis asks whether stale influence and averaging noise explain post-change error more directly than history length alone.

Memory and handoff-state capacity. Two resource effects need to be separated. First, a method can appear stronger simply because it receives a more favorable prompt budget. Second, a handoff representation could improve merely by growing without bound. We therefore vary active-memory and handoff-state capacity independently.

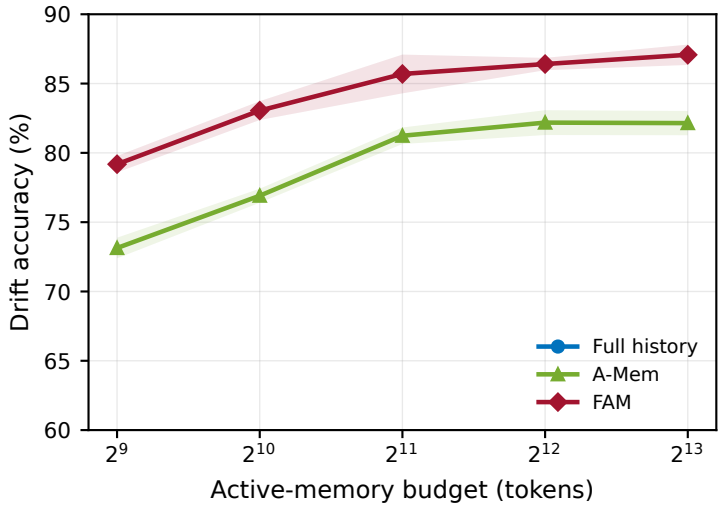


Figure 6: Drift accuracy under matched active-memory budgets. Lines show five-run means and shading shows ±1 SD.

The active-memory sweep tests whether the main comparison is tied to one prompt length. The handoff-state sweep asks whether useful portable state reaches a finite saturation region rather than improving only through unbounded storage.

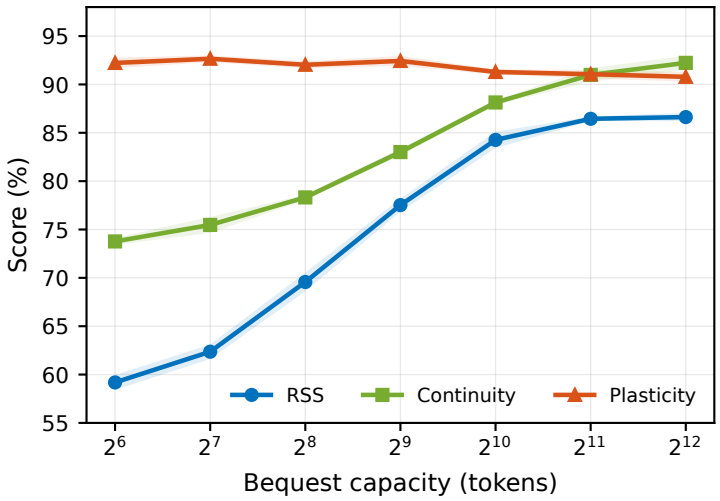


Figure 7: Handoff-state capacity sweep. Lines show five-run means with ±1 SD. Very small states lose continuity and successor information; recovery then saturates while plasticity changes comparatively little.

Adaptation, continuity, and succession. Adaptation half-life is computed per trajectory after a persistent change and then aggregated at the run level. This prevents a good final score from hiding a long interval during which the agent continues to follow an obsolete state. Figure 4a shows this recovery process directly.

The continuity–plasticity analysis in Figure 4b evaluates adaptation together with preservation of information that did not change. This joint view is important because shortening memory indiscriminately can improve plasticity while damaging stable preferences.

Succession provides a separate test of what survives after episodic history is unavailable. Figure 4c evaluates a fresh answer model initialized only from the bounded handoff state and protected constraints. We additionally repeat the drift comparison under three matched language framings.

Table 6: Matched-framing control. The latent trajectory and evaluation queries are identical across renderings. Entries are mean±SD over five runs.

Frame	Full-history drift	FAM drift	Gain
Neutral interaction	68.0 ± 0.4	86.2 ± 0.6	18.2
Personal assistant	68.6 ± 0.5	86.3 ± 0.6	17.7
Romantic companion	68.1 ± 0.6	86.2 ± 0.4	18.1

The similar differences across the three renderings indicate that the observed memory-adaptation effect is not confined to romantic language.

Generalization across temporal patterns and answer models. We test conditions that differ from the controller's training distribution, including an unseen temporal process, unseen persona templates, higher observation noise, and longer horizons.

Table 7: Out-of-distribution current-state accuracy, mean±SD over five runs.

Split	Full history	A-Mem	FAM
Held-out drift family	65.1 ± 0.6	78.9 ± 0.9	84.0 ± 1.2
Held-out persona templates	66.3 ± 0.8	80.2 ± 0.5	84.5 ± 0.5
Higher noise (2x)	63.4 ± 0.5	77.5 ± 0.3	80.6 ± 0.2
In-domain drift	68.1 ± 0.3	82.7 ± 0.5	86.6 ± 0.4
Longer horizon (2x)	62.0 ± 0.7	75.7 ± 0.3	82.4 ± 0.8

Holding out an entire drift family tests transfer across temporal structures rather than paraphrases of the same change process. Higher-noise and longer-horizon conditions probe behavior outside the central training range.

The same frozen memory controller is also evaluated separately on each answer backbone.

Table 8: Current-state accuracy by answer backbone, mean±SD over five independent runs.

Backbone	Full history	Recency-RAG	A-Mem	FAM
Qwen3-8B	67.3 ± 0.8	78.3 ± 0.5	80.5 ± 0.3	84.4 ± 0.3
Llama3.1-8B	68.0 ± 1.0	77.6 ± 0.8	80.8 ± 0.3	84.8 ± 0.8
Qwen3-32B	67.6 ± 0.7	79.1 ± 0.4	81.9 ± 0.9	86.7 ± 0.5
Llama3.3-70B	68.0 ± 0.4	80.0 ± 0.1	82.7 ± 0.9	87.1 ± 1.0
Gemini2.5-Pro	68.8 ± 0.4	80.5 ± 0.4	82.9 ± 0.6	87.7 ± 0.9
GPT-5.4	69.3 ± 0.7	80.4 ± 0.9	83.6 ± 0.8	88.1 ± 0.8

The improvement is therefore not driven by one answer model or one model scale.

Ambiguous evidence and controller supervision. Persistent change is not the only difficult case. A useful memory policy must also avoid overreacting to temporary contradictions, noisy observations, and delayed revisions.

Table 9: Representative high-stress robustness conditions, mean±SD over five runs.

Stress	Level	Full history	A-Mem	FAM
Transient contradiction rate	0.4	61.4 ± 0.3	74.0 ± 0.8	81.6 ± 0.3
Observation label noise	0.2	61.8 ± 1.0	73.6 ± 1.0	80.5 ± 0.7
Revision delay	16 turns	64.1 ± 0.6	77.9 ± 0.6	81.2 ± 0.2

Transient contradictions are particularly informative: a policy that treats every disagreement as a permanent revision can adapt quickly on genuine changes while repeatedly damaging otherwise stable state. We therefore interpret adaptation together with false revision and continuity rather than as an isolated objective.

We next vary how much information is available to the validity controller.

Table 10: Validity-controller supervision, mean±SD over five runs.

Training signal	Drift acc.	AHL↓	Continuity
Oracle change labels	88.1 ± 0.6	7.6 ± 0.5	91.3 ± 0.6
Weak contradiction labels	86.8 ± 0.7	9.2 ± 0.5	91.5 ± 0.2
Self-supervised future prediction	85.6 ± 0.7	10.7 ± 0.4	90.6 ± 0.6
Prompted heuristic only	81.5 ± 0.6	15.3 ± 0.5	88.9 ± 0.4

The oracle condition measures the remaining gap attributable to change detection. More importantly, the weakly supervised and self-supervised settings retain most of the benefit without access to true regime boundaries.

Protected state and mechanism ablations. Ordinary preferences and protected commitments should not share the same revision threshold. We therefore test the protected path separately.

This experiment separates preference plasticity from retention of commitments that require stronger revision evidence. The protected path concerns behavioral influence only; it does not replace a separate mechanism for privacy deletion.

The theory also predicts that neither permanent behavioral retention nor extremely short retention should be uniformly optimal.

Finally, we remove individual components of FAM.

The ablations separate three roles that can otherwise be conflated. Comparing age-only decay with the full controller tests whether temporal validity requires more than recency. Removing the handoff state tests whether retiring episodic evidence destroys useful current state. Comparing bounded and unlimited handoff representations tests whether the gain depends on allowing portable state to grow without constraint.

Efficiency. A memory mechanism can shorten the prompt while moving substantial work into maintenance. We therefore report accuracy together with active tokens and end-to-end latency.

Table 11: Protected-state stress test, mean±SD over five runs.

Method	Stable pref.	Protected constraint	Drift acc.
Sliding window	79.0 ± 0.4	82.3 ± 0.6	81.1 ± 0.7
Exp. decay	82.2 ± 0.5	85.1 ± 0.5	81.8 ± 0.6
Periodic summary	86.7 ± 0.9	90.8 ± 0.3	78.3 ± 0.4
A-Mem	89.1 ± 0.7	91.9 ± 0.4	81.6 ± 0.5
FAM without protection	88.4 ± 0.3	85.0 ± 0.2	87.0 ± 0.5
FAM with protected channel	91.8 ± 0.5	95.1 ± 0.7	86.5 ± 0.4

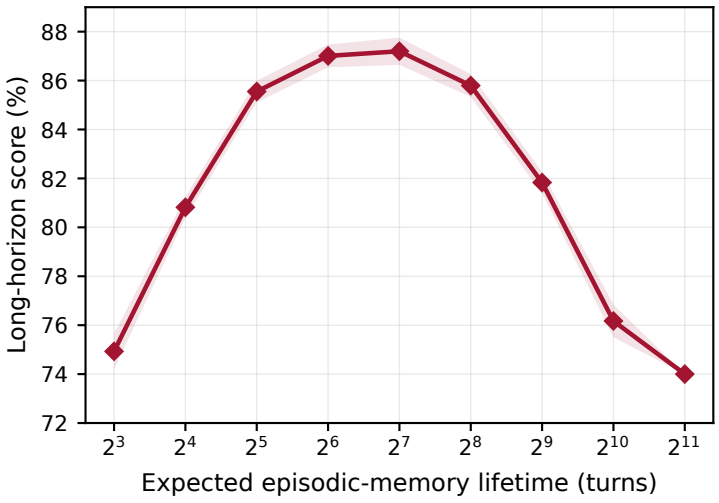


Figure 8: Long-horizon score versus expected episodic-memory lifetime. The line shows the five-run mean and shading shows ±1 SD.

In addition to the end-to-end values above, the implementation records retrieval, write/update, controller, storage, and model-token costs separately. This accounting prevents prompt length alone from hiding computation moved into the memory-maintenance path.

D Failure Analysis and Scope

Failure taxonomy. Aggregate accuracy does not identify whether an error originates in memory management, retrieval, or the answer model. We therefore manually audit more than 100 failed cases and group them into seven recurring patterns: (i) *stale anchoring*, where obsolete evidence remains behaviorally dominant; (ii) *premature revision*, where a temporary contradiction is treated as a durable change; (iii) *context collapse*, where conditional preferences are flattened into one value; (iv) *summary contamination*, where old and new states are merged into an ambiguous representation; (v) *retrieval miss*, where appropriate current evidence exists but is not selected; (vi) *reasoning error*, where the retrieved evidence is appropriate but the answer model still responds incorrectly; and (vii) *protected-state failure*. This distinction is useful because faster forgetting can reduce stale anchoring while simultaneously increasing premature revision.

Interaction history is not calendar age. The operational itch point is defined over retained interaction history rather than elapsed years. Its location depends on the rate and magnitude of user change, observation noise, the memory policy, and the answer model. A stationary user may show no itch point within the tested range. The "seven-year" language is therefore a metaphor for long-horizon relationship aging, not a claim about a universal calendar threshold.

Controlled trajectories simplify real preference change. ITCHBENCH deliberately provides exact temporal ground truth. Real preferences can be contextual, strategic, socially influenced, unresolved, or internally inconsistent, and a past preference can retain autobiographical value after it stops guiding present decisions. The benchmark therefore tests temporal memory management under controlled state change rather than attempting to model the full psychology of a human relationship.

Table 12: Mechanism ablations, mean±SD over five runs.

Variant	Drift	AHL↓	Continuity	RSS	Protected
FAM full	86.3 ± 0.7	8.6 ± 0.5	91.8 ± 0.6	86.8 ± 0.4	95.1 ± 0.4
No validity decay	74.7 ± 0.5	29.0 ± 0.4	94.1 ± 0.6	74.0 ± 0.7	94.4 ± 0.9
Age-only decay	81.7 ± 0.7	15.3 ± 0.8	86.6 ± 0.9	76.8 ± 1.0	88.2 ± 0.5
No contradiction feature	82.6 ± 0.7	13.7 ± 0.4	89.2 ± 1.6	79.4 ± 0.7	90.5 ± 0.4
No uncertainty feature	84.4 ± 0.9	11.6 ± 0.3	90.3 ± 0.5	82.0 ± 0.7	92.7 ± 0.9
No handoff state	79.8 ± 0.5	12.6 ± 0.7	75.4 ± 0.6	58.0 ± 0.5	79.8 ± 0.8
Unlimited handoff state	83.7 ± 0.8	11.9 ± 0.3	94.9 ± 0.7	81.3 ± 0.5	94.8 ± 0.7
No renewal	82.6 ± 1.2	16.3 ± 0.4	92.8 ± 0.5	75.4 ± 0.8	93.7 ± 0.8
Rollback allowed	82.9 ± 0.4	14.5 ± 0.3	91.6 ± 0.3	79.4 ± 0.8	93.6 ± 0.7
No protected channel	86.9 ± 0.4	8.8 ± 0.7	91.0 ± 0.6	86.7 ± 0.5	84.2 ± 0.9

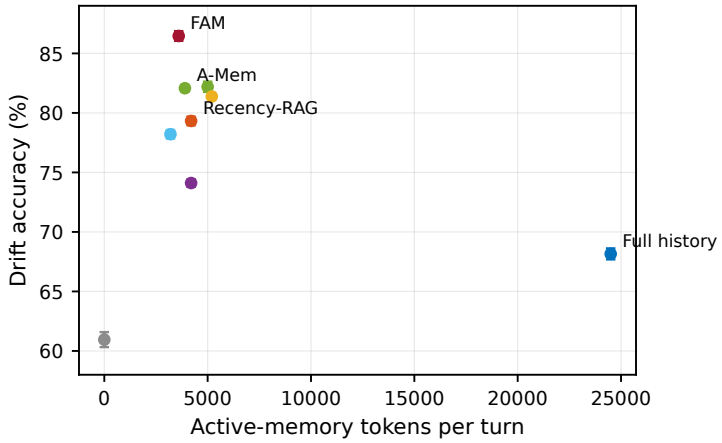


Figure 9: Accuracy versus active-memory tokens per turn. Points show five-run means with SD error bars.

Ambiguous cases include an "uncertain/ask" outcome so that every contradiction need not be forced into a revision.

The theoretical model isolates one mechanism. The weighted-estimator analysis explains how stale evidence can remain influential after a target changes. Real LLM agents add semantic retrieval errors, prompt-order effects, imperfect summaries, controller errors, and changes in the underlying model itself. The theory should therefore be read as a model of one identifiable failure mechanism rather than a complete model of agent behavior.

Behavioral retirement is not privacy deletion. Removing an episode from active retrieval does not erase it from logs, archives, checkpoints, or model parameters. FAM separates archival storage from behavioral influence; a user request for data deletion requires an independent end-to-end deletion mechanism.

Protected state can also change. Protected commitments use a higher-evidence revision path rather than permanent immutability. An explicitly revised commitment can therefore replace an earlier protected state. The benchmark models this distinction through explicit revision events; deployment would require product-specific rules for authentication, authorization, and conflict resolution.

Table 13: Efficiency snapshot. Accuracy and latency are mean±SD over five runs; active-token budgets are determined by the corresponding memory policy.

Method	Drift acc.	Active tokens/turn	Latency (ms)
Full history	68.2 ± 0.5	24,500	2052 ± 20
Similarity RAG	74.1 ± 0.3	4,200	490 ± 6
Recency-RAG	79.3 ± 0.4	4,200	507 ± 6
A-Mem	82.1 ± 0.3	3,900	612 ± 7
FAM	86.5 ± 0.4	3,600	584 ± 6