ProcessPrefBench: Evaluating Cross-Task Transfer of Individual Process Preferences from Naturalistic Human–AI Interaction

Anonymous ACL submission

Abstract

Personalized assistants must learn how a user expects recurring work to be performed. We study *individual process preferences*: user-specific standards for evidence, depth, analytical focus, procedure, and delivery, expressed through corrections, prescriptions, or self-descriptions. ProcessPrefBench tests cross-task transfer using three months of natural interactions from two knowledge workers and 66 source–target cases built through separate judgments of source evidence and target applicability. Three language models answer each target with no memory, the full source interaction, or a distilled preference; target-specific 0–2 rubrics measure functional realization, and a separate task gate flags severe failures. Full-history and imperative distilled-preference prompting improve preference realization over target-only prompting by 0.204 and 0.244, while distillation shows no reliable advantage over full history. Both memory conditions increase fabricated-evidence failures, from 5.8% to 10.4% under imperative injection. These results show that the tested memory conditions improve cross-task realization of individual process preferences but also increase fabricated-evidence failures. Reliable process-preference transfer therefore remains an open challenge for personalized memory systems, which must preserve cross-task preference gains without increasing fabricated evidence in target responses.

1 Introduction

Current personalized AI assistants can remember facts about a user but still struggle to remember how that user works. Figure 1 follows an April correction that redirects product analysis from market-share breakdowns toward who uses a product and how, then a June request about a different product that does not restate that standard. We call such recurring, user-specific standards *individual process preferences*. They shape what evidence supports a claim, how an analysis is organized, how deeply a source is explained, and whether a requested artifact is delivered without further confirmation.

Process preferences also appear through positive requests and descriptions of goals or working habits. Transfer requires a memory system to infer the reusable standard, determine whether it applies to another task, and realize it without treating the remembered interaction as evidence for unsupported claims. These are distinct failure points: the system may preserve the wrong standard, suppress a relevant one, or introduce unsupported content while trying to apply it.

Existing long-term-memory benchmarks emphasize retaining and reasoning over conversational information (Maharana et al., 2024; Wu et al., 2025), while personalization benchmarks often provide a profile or preference record directly (Jiang et al., 2025). These settings do not jointly test preference induction and evidence-bounded execution from natural interaction. It therefore remains unclear whether interaction signals can be converted into auditable behavioral tests.

We introduce ProcessPrefBench using three months of natural interactions in Margin, a reading, annotation, and AI-discussion product used primarily by knowledge workers. The benchmark is an idiographic diagnostic instrument focused on research, sensemaking, and decision tasks from two users. Each case links a preference-bearing source to a different task from the same user. Construction separately establishes that the source supports a reusable standard and that the standard should affect a target that does not restate it.

ProcessPrefBench contributes a source-to-target construction protocol, target-specific rubrics that distinguish surface mention from functional integration, and an evaluation that separates preference realization from severe task failure. Across three models, relevant history improves preference realization, but distillation offers no reliable advan-

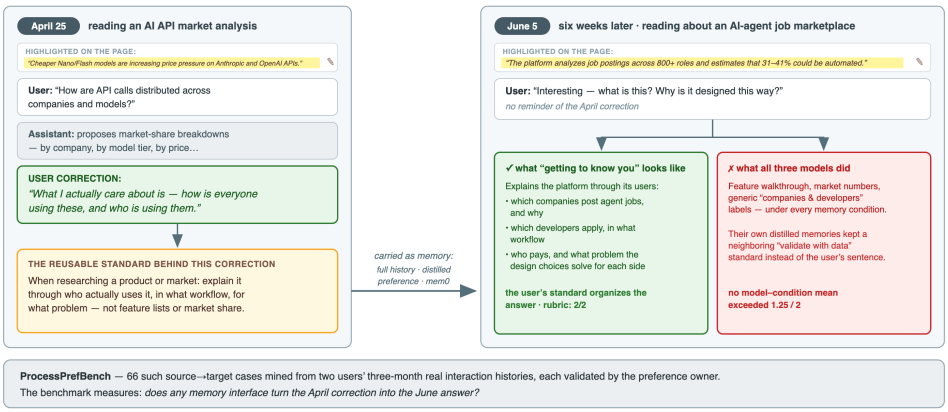

Figure 1: A ProcessPrefBench case. An April correction establishes a reusable product-analysis standard; six weeks later, the same user asks about a different product without restating it. The green panel shows full functional realization, while the red panel summarizes the observed case-level failure.

tage over full history and both memory conditions increase fabricated-evidence failures. The benchmark therefore tests whether assistants can personalize how they work while preserving the factual grounding of the target response.

2 Related Work

Long-term-memory benchmarks test whether assistants retain and reason over information distributed across extended interactions. LoCoMo and LongMemEval cover extraction, temporal reasoning, and knowledge updates across extended conversations, while MemoryAgentBench uses incremental multi-turn interactions (Maharana et al., 2024; Wu et al., 2025; Hu et al., 2025). These benchmarks establish whether information remains available over time. ProcessPrefBench asks whether an interaction also supports a reusable working standard.

Personalization benchmarks typically provide the user signal more directly. PersonaMem studies dynamic user profiling, LaMP evaluates personalized language-model tasks, and PrefEval tests preference recognition and following (Jiang et al., 2025; Salemi et al., 2024; Zhao et al., 2025). ProcessPrefBench instead derives each standard from a natural correction, prescription, or self-description and establishes its applicability to a later target.

Learning preferences from interaction is a closer setting. PRELUDE learns latent preferences from user edits (Gao et al., 2024); our focus is the evaluation problem of tracing a source-supported standard into a different task and distinguishing functional realization from surface mention.

External memory systems such as MemoryBank, MemoryOS, and mem0 store and reintroduce information outside the base model (Zhong et al., 2024; Kang et al., 2025; Chhikara et al., 2025). A retrieved interaction can mix a reusable process standard with earlier-task content, so retrieval alone neither ensures transfer nor prevents that content from entering target-task claims.

3 Benchmark Construction

Data source. We collect three months of natural interactions from two knowledge workers using Margin, a webpage reading, annotation, and AI-discussion product. Each trace retains the available reading context, user request, assistant response, and subsequent user turns. For 13 cases with truncated reading context, we restore verified source material so that the evaluation does not force a choice between fabrication and refusal.

Human source annotation. A source is retained when the user's words support a reusable standard for future work, expressed through correction, prescription, or self-description. Each user annotates their own interactions and excludes factual repairs, current-task-only requests, and scene-bound conclusions because they do not justify transfer beyond the source task. Each retained case stores the inferred standard, its supporting words, and the full interaction as auditable evidence.

Human target pairing. Source annotation does not determine a valid target. For each source, the preference owner separately judges another same-user task. A pairing is retained when the target does not repeat the preference but applying the source-

supported standard should improve the answer. Review also records order, cue strength, material and tool availability, and evaluability. This procedure rejects topical similarity alone and prevents source validity from making target applicability automatic.

Rubric construction. Each case contains one to three source-anchored *preference atoms* with target-specific criteria for not applicable, 0, 1, or 2. A criterion is not applicable when its trigger is absent from the target. A score of 0 marks an absent or contradicted preference, 1 marks local or cosmetic realization, and 2 requires a functional change to the answer's reasoning, organization, conclusion, or delivery. This progression measures whether the remembered standard changes the answer rather than whether the answer merely mentions it.

A separate task gate marks severe omission, major factual error, fabricated evidence, hard-constraint violation, or a false claim of external execution. Task failure does not overwrite the preference score. Keeping the two judgments separate prevents a personalized but unusable answer from counting as success and prevents lack of personalization alone from becoming a general task failure.

Measurement audit. The benchmark changed in response to three measurement checks. An ablation supported removing descriptive reference answers that did not clarify the target-specific rubric. A controlled comparison exposed wording that favored injected preferences and led us to rewrite the affected criteria. A material audit identified truncated reading context that could induce fabrication, motivating the verified restoration described above.

4 Evaluation

Conditions. All conditions share the system prompt, page title, highlighted context, verified supplied material when needed, and verbatim user question. Target-only prompting adds no memory, while full-history prompting adds the complete source interaction. For both distilled-preference conditions, the tested model distills each source once per source–model pair and reuses the result. Imperative prompting always applies it, whereas conditional prompting applies it only when relevant. The mem0 external-memory system extracts from the source and retrieves against the target. Conditional prompting is our variant rather than a description of mem0 semantics.

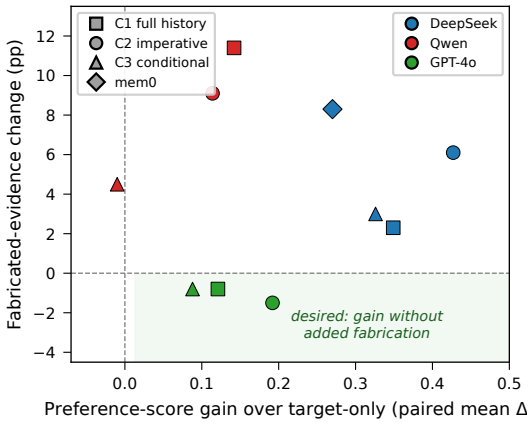

Figure 2: Preference-score and fabricated-evidence changes from target-only prompting across model–condition pairs.

Models and generation. We test DeepSeek V4 Flash, Qwen-Max, and GPT-4o. Every case is generated twice for each model and memory condition. The final run contains 1980 answers.

Judging. Claude Sonnet 4.6 serves in isolated preference and task-gate roles. Both see the target, materials, and answer, but not the tested model, condition, source interaction, or alternatives. The preference judge also sees the rubric and casts three evidence-citing votes per atom. A majority of not-applicable votes excludes the atom; otherwise we use the median applicable score. The task gate sees validity checks without the preference rubric and uses two votes plus a third on ties. Structured-output parsing failure stops the run.

Aggregation. An answer score averages its applicable atoms, and model–condition means exclude answers with no applicable atoms. Task failure is reported independently. The preference score is a rubric-operationalized measure rather than user-satisfaction ground truth. Primary comparisons contrast imperative distilled-preference prompting with target-only and full-history prompting. After averaging two generations within each case–model cell, we macro-average paired differences across models and report win/tie/loss counts with exact sign tests. Conditional distilled-preference prompting is descriptive because every formal case is an intended positive match.

5 Results

Preference transfer. Access to user-specific history improves preference scores over target-only

Condition	DeepSeek V4 Flash		Qwen-Max		GPT-4o	
	Preference score	Task failure rate	Preference score	Task failure rate	Preference score	Task failure rate
Target-only prompting	0.79	8.3%	0.62	13.6%	0.47	23.5%
Full-history prompting	1.15	9.1%	0.77	23.5%	0.59	14.4%
Imperative distilled-preference prompting	1.23	14.4%	0.73	22.7%	0.66	19.7%
Conditional distilled-preference prompting	1.11	8.3%	0.61	18.9%	0.58	23.5%
mem0 external-memory system	1.06	15.2%	0.59	21.3%	0.49	22.6%

Table 1: Preference score (higher is better) and task-failure rate (lower is better).

prompting. Across models, imperative distilled-preference prompting exceeds target-only prompting by 0.244 points (48/8/9 wins/ties/losses; exact sign $p < .0001$), with positive effects for DeepSeek V4 Flash (+0.427), Qwen-Max (+0.114), and GPT-4o (+0.192). Full-history prompting gains 0.204 points over target-only prompting (44/12/9; $p < .0001$), showing that the complete source interaction already provides a strong signal.

Distillation does not yield a reliable preference-score advantage over full history. The difference is +0.026 ($p = .451$), with model-level effects of +0.080 for DeepSeek V4 Flash, -0.058 for Qwen-Max, and +0.071 for GPT-4o. This does not establish equivalence; the representations' relative preference effects remain unresolved. The mem0 cell means exceed target-only prompting for DeepSeek V4 Flash and GPT-4o but not for Qwen-Max.

Task validity and factual reliability. A preference-score gain does not guarantee answer validity (Figure 2). Imperative distilled-preference prompting raises the overall task-failure rate from 15.2% to 18.9% relative to target-only prompting, although the difference is not conventionally significant (exact McNemar $p = .086$). Fabricated-evidence failures rise from 5.8% to 10.4% ($p = .006$). Full-history prompting shows a similar 4.3-point increase over target-only prompting ($p = .010$), while the two memory conditions differ by only 0.3 points ($p = 1.0$). This aggregate risk therefore accompanies history access rather than distillation alone. The increase is largest for Qwen-Max under imperative distilled-preference prompting (+9.1 points, $p = .004$), while GPT-4o shows no same-direction increase. The plotted mem0 comparison also shows an 8.3-point increase in fabrication ($p = .019$). All failure comparisons use exact McNemar tests.

Method-level interpretation. The interfaces leave different parts of the transfer problem unresolved. Full-history prompting improves preference realization but increases fabricated-evidence failures. Imperative distilled-preference prompting preserves the transfer benefit but neither reliably improves on full history nor reduces its factual risk. Conditional distilled-preference prompting scores lower than both alternatives across all three models on intended positive matches, consistent with false suppression. The mem0 results are heterogeneous across models and do not provide a consistent preference-score gain. No evaluated interface combines reliable preference transfer with preserved factual reliability; autonomous selection from long histories remains untested.

Robustness and scope. The main comparison between imperative distilled-preference prompting and target-only prompting remains positive separately for user1 and user2, after excluding the three temporally inverted cases, under source- and target-cluster aggregation, and when restricting the analysis to normally terminated generations. Excluding the temporally inverted cases yields a gain of 0.252 points ($p < .0001$). Corresponding comparisons with full-history prompting remain non-significant. These checks support history access within the two-user benchmark but do not resolve the relative value of distillation or establish population-level effects.

6 Conclusion

Natural interactions provide evidence about how a user expects recurring work to be performed. ProcessPrefBench turns this evidence into 66 cross-task cases by separately judging source evidence and target applicability. Across three models, relevant history improves functional preference realization, yet a system may still preserve the wrong standard, fail to realize a correctly preserved one, or import remembered content as unsupported target claims. Reliable transfer requires solving all three failure points, and ProcessPrefBench makes them measurable.

Limitations

ProcessPrefBench is an idiographic benchmark built from two knowledge workers using one product, with an imbalanced 52/14 case split and a concentration in research, sensemaking, and decision tasks. Its 66 cases arise from 30 corrective sources and 59 distinct targets, so repeated sources and targets limit case-level inference even when cluster-level sensitivity analyses preserve the main direction.

References

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. 2025. Mem0: Building production-ready ai agents with scalable long-term memory. *Preprint*, arXiv:2504.19413.

Ge Gao, Alexey Taymanov, Eduardo Salinas, Paul Mineiro, and Dipendra Misra. 2024. Aligning LLM agents by learning latent preference from user edits. In *Advances in Neural Information Processing Systems 37*.

Yuanzhe Hu, Yu Wang, and Julian McAuley. 2025. Evaluating memory in LLM agents via incremental multi-turn interactions. *Preprint*, arXiv:2507.05257.

Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo J. Taylor, and Dan Roth. 2025. Know me, respond to me: Benchmarking LLMs for dynamic user profiling and personalized responses at scale. In *Proceedings of the Conference on Language Modeling (COLM)*.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. 2025. MemoryOS of ai agent. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25972–25981.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. 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)*, pages 13851–13870.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. LaMP: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7370–7392.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2025. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *The Thirteenth International Conference on Learning Representations*.

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. 2025. Do LLMs recognize your preferences? evaluating personalized preference following in LLMs. In *The Thirteenth International Conference on Learning Representations*.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. MemoryBank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19724–19731.

A Scoring Protocol

Preference score. The preference judge scores each target-specific atom independently. N/A means that the atom's trigger is absent and is excluded rather than scored as zero. A score of 0 marks an absent or contradicted preference; 1 marks nominal, local, or cosmetic realization; and 2 requires the preference to change the answer's reasoning, organization, conclusion, or delivery. Each vote must cite answer evidence. Three votes are collected per atom; a majority of N/A votes excludes the atom, and otherwise the median applicable score is used.

Task gate. The separate task gate flags only severe unusable errors: non-answer, major factual error, fabricated evidence, hard-constraint violation, or a false claim of external execution. Missing personalization alone is not a task failure. The gate uses two votes and a third on ties, without access to the preference rubric.

B Robustness Analyses

Table 2 expands the checks summarized in the Results. Differences are computed after averaging the two generations within each case–model cell. Exact sign tests exclude ties.

Comparison	Check	n	Diff.	p
Imp. − T0	user1	51	+0.194	<.0001
Imp. − Full	user1	52	+0.020	.8601
Imp. − T0	user2	14	+0.429	.0034
Imp. − Full	user2	14	+0.051	.3877
Imp. − T0	excl. inversions	62	+0.252	<.0001
Imp. − Full	excl. inversions	63	+0.030	.3604
Imp. − T0	source clusters	30	+0.268	<.0001
Imp. − Full	source clusters	30	+0.046	.4049
Imp. − T0	target clusters	58	+0.249	<.0001
Imp. − Full	target clusters	59	+0.024	.2559
Imp. − T0	normal termination	52	+0.274	<.0001
Imp. − Full	normal termination	52	+0.038	.7283

Table 2: Sensitivity analyses for the primary imperative distilled-preference comparisons. T0 denotes target-only prompting and Full denotes full-history prompting.