PAIRPREF: WHEN SHOULD MEMORY GUIDE THE ANSWER? A BENCHMARK FOR CONTEXTUAL PREFERENCE USE

Anonymous authors
Paper under double-blind review

ABSTRACT

Memory-augmented assistants use retrieved preferences to guide their responses. Task circumstances, such as who a recommendation is for or what resources are available, can make a remembered preference inappropriate to apply. Applying it anyway can produce an answer that conflicts with the current task's needs. Memory benchmarks typically test whether systems store and retrieve preferences, with less attention to when those preferences should guide an answer. We introduce PairPref, a benchmark of contextual preference use. Each pair holds the stored preference, request wording, and four candidate replies fixed while varying the task circumstances. The preference remains valid but should guide the answer in only one of the two situations. In the selection track, models must choose the reply that applies the preference only where appropriate. In the free-generation track, they must decide when to apply it without seeing candidate replies. Both tracks use the same 1,227 pairs across 45 preferences and eight situation categories. We evaluate eight models, most of which achieve selection scores (Δ) of 51 to 65 points. In free generation, however, both responses are appropriate for their respective situations in only 3.6% to 18.3% of pairs retained after hedging exclusions. Models continue to apply the preference in both situations even with fewer retrieved memories, alternative presentation formats, and a stricter prompt. Models still struggle to judge when preferences apply and act accordingly.

1 INTRODUCTION

Language-model assistants increasingly support users in everyday tasks, from choosing meals to reviewing papers and selecting gifts. Across repeated interactions, they accumulate user preferences that help personalize their responses (Zhong et al., 2024; Xu et al., 2025; Lu et al., 2026). These preferences can remain valid even when the situation calls for a different response. An assistant must therefore judge when a remembered preference should guide its answer (Figure 1).



Figure 1: User preferences remain valid across situations, but their applicability changes with context. The same request may call for applying a preference in one situation and setting it aside in another. PairPref evaluates whether assistants make this distinction across paired situations.

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.

Memory and personalization benchmarks fall into three lines. (1) *Long-term memory benchmarks* evaluate recall and reasoning over extended interactions, as in LoCoMo and LongMemEval (Maharana et al., 2024; Wu et al., 2025). (2) *Preference-following benchmarks* assess preference inference, tracking, and application, including PrefEval, PersonaMem, and CUPID (Zhao et al., 2025; Jiang et al., 2025; Kim et al., 2025). (3) *Memory-use boundary benchmarks* examine inappropriate personalization and disclosure (Hu et al., 2026; Mireshghallah et al., 2025; Pulipaka et al., 2026; Xu et al., 2026). Within this line, RPEval evaluates whether a preference should be ignored, support a response, or dominate it (Feng et al., 2026), while BenchPreS tests expression preferences under formal communication norms (Yoon et al., 2026). We test a contextual contrast (Table 1):

When the preference and request remain unchanged, can an assistant apply the preference in one situation and set it aside in another, responding appropriately in both cases?

To address this question, we introduce PairPref, a benchmark of contextual preference use comprising 1,227 pairs across 45 preferences and eight situation categories. Each pair holds a still-valid preference and a request fixed while varying the situation so that the preference should guide the answer in one case but not the other. The selection track presents the same four candidate replies in both situations, testing whether the model changes its choice appropriately. The free-generation track removes the candidates and tests whether the model produces appropriate answers for both situations. Together, the two tracks assess whether models adjust preference use to context, both when selecting a reply and when composing their own.

We evaluate eight models and find that contextual preference use remains challenging. Seven models achieve selection scores (Δ) between 51.3 and 64.9 points, yet both generated responses are appropriate for their respective situations in only 3.6%–18.3% of pairs across the eight models. Models apply the preference in both situations in 71.0%–92.0% of generated pairs. This pattern persists with fewer retrieved memories, alternative presentation formats, and a stricter prompt. These results highlight the difficulty of translating situational differences into appropriate preference use.

Our contributions are summarized as follows:

❶ **Contextual preference benchmark.** We introduce PairPref to test contextual applicability while keeping a still-valid preference and the request fixed.

❷ **Paired evaluation protocol.** We combine reply selection and free generation on the same situation pairs to measure whether both responses are appropriate.

❸ **Empirical findings.** We evaluate eight models and identify persistent preference application across situations, even when the context calls for setting the preference aside. This behavior persists across memory-pool sizes, presentation formats, and prompting conditions.

2 Related Work

Preference and memory benchmarks increasingly assess appropriate use as well as recall. Table 1 compares their coverage of preference following, suppression, selection, and generation. Appendix D gives evaluation protocols, response formats, dataset sizes, and validation procedures.

Remembering and applying user information. Multi-session dialogue tests continuity across chats (Xu et al., 2022). LoCoMo and LongMemEval evaluate long-term conversational memory; the latter also tests preference questions and abstention when information is unavailable (Maharana et al., 2024; Wu et al., 2025). HaluMem localizes failures to extraction, update, and question answering (Chen et al., 2025). LaMP and LongLaMP test profile-conditioned classification and long-form generation (Salemi et al., 2024; Kumar et al., 2024). PrefEval evaluates explicit and implicit preference following through generation and multiple-choice tasks, including response helpfulness (Zhao et al., 2025). PersonaMem tracks evolving user profiles, while CUPID tests inference and application of context-dependent preferences (Jiang et al., 2025; Kim et al., 2025). PairPref holds a preference's validity fixed and tests how its applicability changes with the situation.

Deciding when and how to use memory. Benchmarks assess both preference strategies and information-use boundaries. For preference strategies, RPEval evaluates whether preferences should

Table 1: Capabilities assessed by preference and memory benchmarks.

Benchmark	Main evaluation target	Preference following	Preference suppression	Selection	Generation
Long-term memory					
LoCoMo	Conversational recall	—	—	—	—
LongMemEval	Long-term memory	◐	—	—	—
Preference following					
PrefEval	Preference following	✓	—	—	—
PersonaMem	Evolving user profiles	✓	◐	◐	◐
CUPID	Contextual preferences	✓	◐	—	◐
Memory-use boundaries					
RPEval	Preference-use strategies	✓	✓	◐	◐
OP-Bench	Over-personalization	—	✓	—	◐
BenchPreS	Communication norms	✓	✓	—	◐
CIMemories	Contextual disclosure	—	—	—	—
PersistBench	Leakage and sycophancy	◐	◐	—	◐
HUSH-Bench	Sensitive-history boundaries	◐	◐	—	◐
PairPref	Preference applicability	✓	✓	✓	✓

✓ Direct evaluation; ◐ related coverage; — outside the evaluation target. Selection and generation assess when to apply or withhold the same still-valid preference across situations with a fixed request. Related coverage uses different controls or task definitions; see Appendix D.1.

be ignored, support an answer, or dominate it (Feng et al., 2026); BenchPreS measures appropriate application and suppression under formal communication norms (Yoon et al., 2026). For over-personalization and information disclosure, OP-Bench tests irrelevance, sycophancy, and repetition (Hu et al., 2026); PersistBench tests leakage and sycophancy alongside beneficial memory use (Pulipaka et al., 2026). CIMemories measures inappropriate disclosure and necessary information coverage (Mireshghallah et al., 2025), while HUSH-Bench tracks sensitive-history use under different memory-access and invitation conditions (Xu et al., 2026). PairPref isolates a situational reversal within these contextual boundaries: the same still-valid preference and request warrant application in one situation and non-application in another. It evaluates both situations jointly, using fixed candidate replies for selection and the same pairs for free generation. See Appendix D.1 for details.

Organizing and retrieving memories. Memory systems study how records are formed, stored, and retrieved (Hu et al., 2025). Generative agents store experiences and reflect on them (Park et al., 2023). Tiered stores separate short- and long-term records (Packer et al., 2023; Kang et al., 2025); linked notes and temporal graphs preserve relations among entries (Xu et al., 2025; Rasmussen et al., 2025). Retrieval-augmented generation supplies external text (Lewis et al., 2020), while HippoRAG and its successor use structured retrieval as non-parametric memory (Gutiérrez et al., 2024; 2025). Other systems reduce memory-processing costs (Chhikara et al., 2025; Fang et al., 2026) or learn what to write, retrieve, and fuse (Yan et al., 2026; Du et al., 2025; Lu et al., 2026; Zhang et al., 2026). PairPref tests whether models adjust preference use to the situation given fixed memories.

3 Benchmark Construction

PairPref tests whether assistants adjust preference use when the situation changes but the preference remains valid. Each pair fixes a preference and a request, then contrasts a situation where the preference should guide the answer with one where it should be set aside. Selection presents the same four candidate replies in both situations; free generation removes the candidates. Figure 2 summarizes construction, validation, and both evaluation tracks.

3.1 TASK FORMULATION

Let M_i denote a persistent preference, q_i a shared request, and C_i^+ and C_i^- situations that warrant applying and setting aside the preference, respectively. Each pair is

$$x_i = (M_i, q_i, C_i^+, C_i^-, \mathcal{R}_i), \qquad \mathcal{R}_i = \{r_{\mathrm{M}}, r_{\mathrm{sup}}, r_{\mathrm{obl}}, r_{\mathrm{irr}}\}. \tag{1}$$

The four replies distinguish preference use from acknowledgment and task completion. The r_{M} reply fully follows the preference; r_{sup} departs from it and acknowledges doing so; r_{obl} departs without acknowledgment; and r_{irr} answers an unrelated task. The two departing replies may recommend the same action. This contrast separates acknowledgment from the action itself, while the off-task reply distinguishes withholding a preference from completing the request.

Selection tests whether the model changes its reply appropriately across the pair. Free generation tests whether it produces appropriate responses without candidates. Both tracks assess success on the two situations jointly. The preference, request, shared task facts, and companion memories remain fixed, isolating the effect of the situation on preference use.

3.2 PAIRED INSTANCE CONSTRUCTION

Generating situations and replies. We construct pairs around 45 preferences across six domains and eight situation categories. GPT-5.5 generates a shared request and two situations in which the same preference differs in applicability, then produces four candidate replies representing preference use, acknowledged departure, unacknowledged departure, and an unrelated task. The request and replies remain fixed across the pair. Situations convey circumstances that affect the decision rather than directly instructing the assistant to follow or ignore the preference.

Supplying companion memories. Each target preference is accompanied by four near, five middle, and six far memories. The resulting 16-memory pool contains the same entries on both sides and is supplied directly to the model. The target preference is therefore available in both situations; the task is to judge whether to apply it. Context-to-request distances vary across one, three, and six rounds, with filler exchanges separating context from the request.

3.3 MULTI-STAGE VALIDATION

We validate candidate pairs through three complementary stages: rule-based checks, model-based semantic verification, and context-free choice screening (Figure 2).

Rule-based checks. Programmatic checks screen candidate text and option form, including reply lengths, repeated wording, and explicit cues that reveal the intended preference decision. These checks filter surface violations before semantic verification.

Semantic verification. DeepSeek V4 Flash checks whether the situations support the intended change in applicability, whether the shared request is natural in both situations, and whether the replies meet their intended roles. The checks also assess whether the constraint requires reasonable inference, whether preference use remains feasible on the positive side, and whether both sides have usable answers. Appendix B details the checks and model configurations.

Context-free choice screening. To screen for candidate cues that bypass the contextual judgment, Qwen receives only the request and four replies, without the preference or either situation, and selects the most appropriate reply. Each candidate is tested three times and excluded if the preference-enacting reply is selected more than once. The 1,227 retained pairs pass all construction filters. Separately, two annotators independently assessed 200 sampled pairs after situation revisions, reaching 99.25% applicability agreement ($\kappa = 0.985$). Appendix H reports this revised-sample assessment separately from the collection used for model evaluation.

3.4 DATASET COVERAGE

PairPref contains 1,227 pairs covering 45 preferences across food, film, activities, goods, routines, and social interaction. Eight situation categories vary companions, occasion, time constraints, ben-

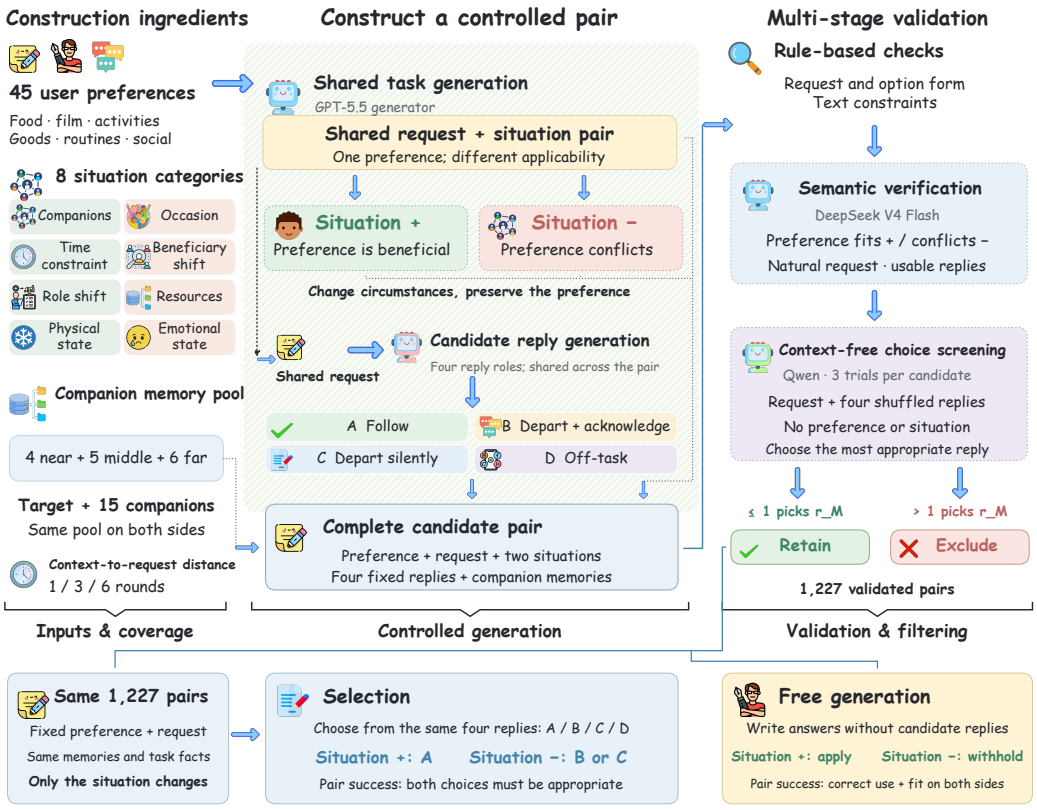


Figure 2: PairPref construction and evaluation. Controlled pairs pass three validation stages and support selection and free generation on the same situations.

eficiary, role, resources, physical state, and emotional state. Positive situations comprise 620 clean and 607 distractor cases: the latter introduce an additional situational cue while retaining the preference's applicability. Both evaluation tracks use the same pairs. Appendix C reports the category and domain distributions.

4 Experiments

4.1 Experimental setup

Models and inputs. We evaluate eight models on both tracks of PairPref: Claude Opus 5 (Anthropic, 2026), Kimi K3 (Moonshot AI, 2026), GPT-6 Astra (OpenAI, 2026), GLM-5.3 (Z.ai, 2026), Gemini 3.5 Flash (Google DeepMind, 2026), DeepSeek V4 Pro (DeepSeek-AI, 2026), Qwen 3.8 Max (Qwen Team, 2026), and Grok 4.6 (SpaceXAI, 2026). Each input contains the target preference and 15 companion memories as a system list, followed by a situation, intervening dialogue turns, and the shared request. The default instruction asks models to use memories when relevant without forcing them into the answer. The two situations are evaluated in separate calls, with intended reply roles and side labels hidden. Selection provides the four candidate replies, whereas free generation omits them. See Table 2 for the models and Appendix F.1 for implementation details.

Metrics. The positive situation calls for using the preference; the negative situation calls for setting it aside. Let u^+ and u^- denote preference use on these two sides. Selection assigns $u = 1$ when the model chooses r_M and 0 otherwise. Generation assigns $u \in \{0, 0.5, 1\}$ for no reliance, implicit reliance, and explicit or central reliance on the preference, respectively.

We report three measures of preference use. *Use*$^+$ ($\mathrm{MU}^+ = \mathbb{E}[u^+]$) measures whether models use the preference when it fits; higher is better. *Use*$^-$ ($\mathrm{MI}^- = \mathbb{E}[u^-]$) measures whether they continue using it when it should be set aside; lower is better. *Selectivity* ($\Delta = \mathrm{MU}^+ - \mathrm{MI}^-$) measures the separation between these behaviors. Always using or never using the preference both yield $\Delta = 0$; perfect separation yields 100 points. In selection, these averages are selection rates; in generation, they average graded uptake. Selection Δ counts off-task replies as non-use, not errors.

Selection *Pair success* is the percentage of pairs with r_{M} selected in the positive situation and either r_{sup} or r_{obl} in the negative situation. Selecting the off-task reply never counts as success.

For free generation, we evaluate the two answers in each pair jointly. *Use both* (Always-use) is the percentage of pairs in which the model uses the preference in both situations. Since the negative situation calls for setting it aside, this measures persistent preference use despite the contextual change; lower is better. *Pair success* (Gen-OK) is the percentage of pairs satisfying all three conditions: the positive answer uses the preference, the negative answer does not, and both answers complete the request under their respective situations. A failure on either side makes the pair unsuccessful; higher is better. These metrics are not complements: avoiding the preference in both answers lowers Use both but does not constitute Pair success, nor does withholding it in an off-task answer.

Formally, Use both requires $u^+ \geq 0.5$ and $u^- \geq 0.5$. Pair success requires $u^+ \geq 0.5$, $u^- = 0$, and a situational-fit score of 1 on both sides. Each rate uses retained pairs with all required judgments.

Selection is scored directly from the chosen reply, while generated answers are evaluated by DeepSeek V4 Flash using separate judgments of preference use and situational fit.

Included pairs and uncertainty. We compute paired results only when both sides have valid outputs. We further exclude a generated pair if either answer offers both a preference-based and a generic alternative without committing to one, such as "Choose X if alone, but Y if with company." This hedging rule distinguishes a contextual decision from covering both possibilities. Table 2 reports the remaining number of pairs for each model; generation rates are calculated over these retained pairs, with Pair success also requiring both fit judgments. For example, a Pair success of 10% means that one in ten scored pairs satisfies all success conditions. The 95% confidence intervals use 1,000 bootstrap resamples, clustering pairs by target preference to account for shared content.

4.2 Models struggle with contextual preference use in both tracks

Contextual preference use remains challenging in both tracks (Table 2). Selection Pair success ranges from 23.6% to 66.4%, while generation Pair success is 3.6%–18.3%. Even the strongest selection result leaves roughly one third of pairs unsuccessful. These rates follow each track's scoring and inclusion rules and are not directly comparable across tracks.

The predominant generation pattern is to use the preference in both situations: Use both reaches 71.0%–92.0% across the eight models on retained generation pairs (Figure 3). Even when we restrict the analysis to pairs on which the same model succeeds on both selection sides, this proportion remains 71.8%–92.2%. Selecting the intended replies does not ensure appropriate preference use in answers generated for the same situations. See Appendix F.8 for matched counts and model results.

An illustrative case. A user who drinks only hand-poured single-origin coffee asks what to order at a market counter. With sufficient credit, Claude Opus 5 selects the pour-over reply; when the wristband has only enough credit for the self-serve urn, it selects the house-drip reply. Its generated positive answer also recommends pour-over. In the negative situation, however, it begins "Skip the urn," rejects the available coffee because it does not match the user's preference, and recommends postponing coffee to visit another shop. The answer preserves the preference by moving the purchase elsewhere rather than choosing within the stated resource constraint. The judge records preference use on both sides and failed situational fit on the negative side (full example in Appendix F.9).

Table 2: Main results on PairPref.

	Selection				Free generation		
Model	Use$^+$ ↑	Use$^-$ ↓	Selectivity ↑	**Pair success ↑**	Retained pairs	Use both ↓	**Pair success ↑**
Claude Opus 5	76.4	**11.5**	**64.9** [58.2–70.6]	**66.4**	1,067	88.7	9.3
Kimi K3	76.0	18.3	57.6 [50.8–63.6]	59.5	1,091	83.6	10.6
GPT-6 Astra	71.2	14.7	56.6 [49.8–62.6]	57.9	1,068	**71.0**	17.9
GLM-5.3	75.1	18.4	56.6 [50.3–61.9]	58.4	1,056	90.8	5.7
Gemini 3.5 Flash	**78.9**	22.9	56.0 [49.1–62.5]	57.9	1,065	75.1	**18.3**
DeepSeek V4 Pro	76.4	23.0	53.5 [47.4–58.6]	55.2	1,107	81.3	13.5
Qwen 3.8 Max	71.6	20.3	51.3 [45.0–56.9]	54.1	1,011	81.2	12.7
Grok 4.6	75.1	57.4	17.7 [12.2–22.3]	23.6	1,153	92.0	3.6

Rates are percentages; selectivity (Δ) is in percentage points, with 95% CIs below. Generation rates use retained pairs. Bold marks the best value per metric. Metric definitions, scoring, and exclusions are in Section 4.1.

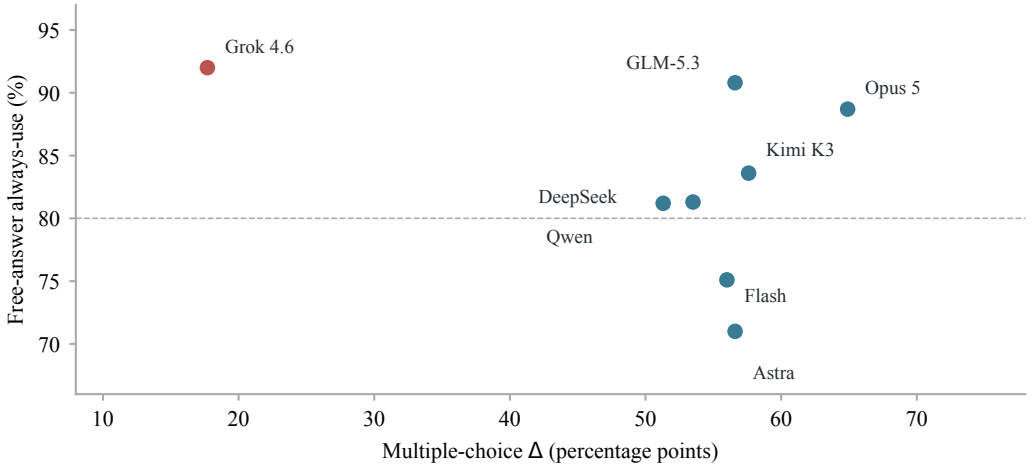


Figure 3: Multiple-choice Δ versus free-answer always-use on the same pairs (Table 2). Contextual separation in selection coexists with frequent preference use in both generated answers.

4.3 Effects of situation and memory

Removing the situation reduces selection selectivity to near zero; removing the memory block markedly reduces alignment between generated answers and the target preference (Table 3). These controls show that contextual separation in selection depends on the situation, while frequent preference-related content in generation is associated with the supplied memory.

4.4 Pool size does not explain preference misuse

A shorter pool preserves negative-side selection of the preference. Position in a long context can change what a model uses (Liu et al., 2024; Hsieh et al., 2024). Top-5 retains the target preference and four near companions. Negative-side use increases on four of the five evaluated models; GPT-6 Astra decreases from 14.7% to 11.7% (Table 4). Pair success changes by -2.0 to $+4.6$ percentage points across these models, with no consistent improvement (selectivity is reported in Appendix F.3).

Table 3: Situation and memory removal.

A. Selection: selectivity

	Full input	No situation
Model	Δ	Δ
Gemini 3.5	56.0	0.0
Kimi K3	57.6	-0.2
DeepSeek V4	53.5	-0.5
GLM-5.3	56.6	-0.7
GPT-6 Astra	56.6	1.4
Qwen 3.8	51.3	—

B. Generation: preference uptake

	Full input		No memory	
Model	Use$^+$	Use$^-$	Use$^+$	Use$^-$
Gemini 3.5	78.8	64.7	11.1	7.5
Kimi K3	84.4	77.6	15.1	13.1
DeepSeek V4	84.1	73.4	11.7	9.3
GLM-5.3	87.5	84.2	16.3	15.8
GPT-6 Astra	78.3	67.4	11.4	11.3
Qwen 3.8	83.9	75.6	11.7	11.4

A: Δ in percentage points; — means not evaluated. B: mean preference uptake (0–100). No-memory uptake measures alignment with the withheld preference. Blue marks removals. Details: Appendix F.11.

Replacing companions with 15 far items leaves Δ in the same range (Appendix F.4). Misuse persists from a 16-item list to five related items.

4.5 Stricter prompting changes selection more than generation

The default relevance instruction does not account for the selection pattern: removing the "use when relevant" clause changes multiple-choice negative use by at most a few points (Appendix F.6). Explicitly distinguishing truth from applicability in the mitigate instruction lowers multiple-choice misuse—Opus 5 from 11.5% to 3.6%, GLM-5.3 from 18.4% to 8.1%—and also lowers positive-side use for seven of eight models. Flash is the exception, moving from 78.9% to 80.3%. On free answers, always-use is 54.0%–85.5% for the five completed models (Figure 5). Astra moves from 71.0% to 54.0% always-use and from 17.9% to 26.8% Gen-OK. A stricter withhold prompt changes letter selection more than written recommendations.

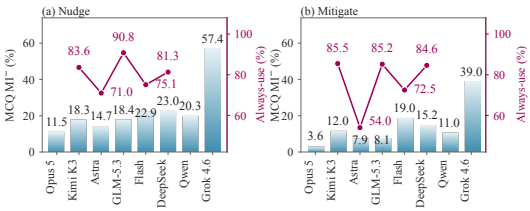


Figure 5: Prompt wording. Bars are negative-side multiple choice (left axis). The line shows free-answer always-use (right).

4.6 Preference misuse persists across memory presentations

Multiple-choice misuse persists beyond the system-list format. Moving the same 16 memories into a user block or prior dialogue turns leaves negative-side use in the same band (Figure 4b). Several models increase slightly (Qwen 3.8 Max 20.3% to 27.1% under dialog). These presentations rule out dependence on the system-list format.

4.7 Topical relevance does not resolve situational applicability

The negative situation changes whether an on-topic preference should be applied. A classifier given the preference and the request marks both sides for use on 99.4% of pairs. A relevance gate marks both sides relevant on 98.6% (Table 12). These judgments distinguish topical relevance from situational applicability. An apply head distinguishing relevance from use reaches $\Delta = 62.7$.

5 Discussion

Recognizing when a preference applies does not ensure that generated answers respect that boundary. Seven models distinguish the situations in multiple choice, while all eight apply the preference on both sides in most free-answer pairs. The dual-track protocol exposes this distinction between selecting and producing an appropriate reply.

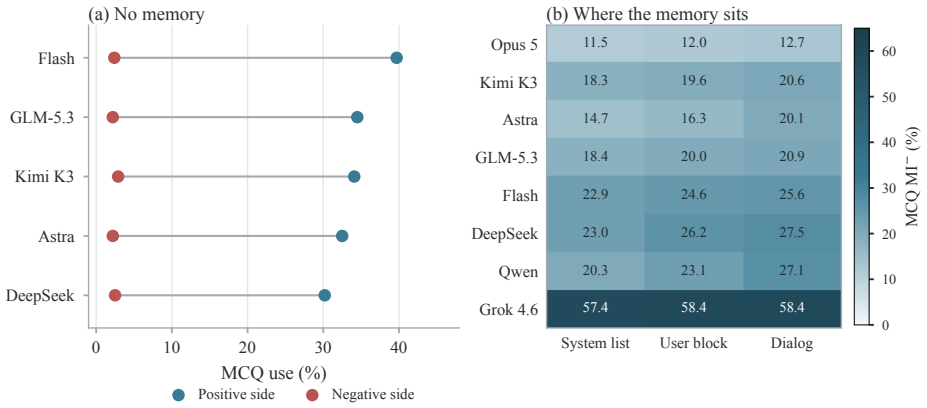


Figure 4: Two controls on multiple choice. (a) Removing memory preserves positive-side but reduces negative-side selection. (b) Negative-side use remains similar across memory presentations.

Table 4: Effect of memory-pool size on selection performance.

	Full pool (16)			Smaller pool (5)		
Model	Use$^+$ ↑	Use$^-$ ↓	**Pair success** ↑	Use$^+$ ↑	Use$^-$ ↓	**Pair success** ↑
Gemini 3.5 Flash	78.9	22.9	57.9	81.5	25.8	57.8
DeepSeek V4 Pro	76.4	23.0	55.2	78.6	27.8	53.1
Kimi K3	76.0	18.3	59.5	79.7	20.7	61.3
GLM-5.3	75.1	18.4	58.4	77.7	21.0	59.0
GPT-6 Astra	71.2	14.7	57.9	73.3	11.7	62.5

Rates in %; $n = 1{,}227$, except GLM-5.3 with the smaller pool ($n = 1{,}226$; one incomplete output).

Situation removal erases multiple-choice discrimination; memory removal reduces generated uptake. A five-item on-topic pool and two other presentations preserve letter-selection misuse. Stricter prompting changes selection while leaving free-answer always-use at 54%–85%. Situational applicability thus remains a distinct target beyond availability and topical relevance.

Possible mechanisms. A retrieved preference may be treated as a requirement to satisfy, leading reasoning to focus on how to preserve it before deciding whether it applies. The coffee example illustrates how an answer can accommodate the preference by moving the purchase outside the stated constraint. Longer chains of thought could therefore elaborate such workarounds rather than improve contextual judgment. Unequal attention to preference and situation cues is another possible contributor. Candidate replies may help selection by making the alternative of setting aside the preference explicit. Appendix G develops these hypotheses and proposes tests.

The benchmark studies content preferences supplied in context under English synthetic situations; clean and distractor strata record construction intent. The 1,227-pair collection supports both selection and free generation; reported free-answer rates are conditional on the output-validation and hedging exclusions described in Section 4.1. Dataset documentation preserves these conditions (Gebru et al., 2021; Mitchell et al., 2019).

6 Conclusion

PairPref tests when a still-valid preference should guide an answer across 1,227 situation pairs. Its paired design isolates contextual applicability, while two tracks distinguish selection from generation. Models often select appropriate replies yet apply preferences in both generated answers. Evaluating both tracks makes this distinction visible and provides a shared test bed for improving contextual judgment. Reliable personalization requires more than retrieving relevant preferences: it requires judging when they apply and carrying that decision into the answer.

AI USE STATEMENT

Generative AI produced synthetic situations, requests, replies, and companion memories, and performed candidate verification and response scoring. An AI assistant supported manuscript preparation, including language polishing.

ETHICS STATEMENT

The benchmark contains synthetic preferences and situations. Two annotators independently assessed revised sample materials (Appendix H). A model's social judgments should not become universal behavioral requirements. Situations state constraints explicitly, but interpretation can depend on cultural assumptions. We disclose operational LLM labels and boundaries. Public artifacts must respect licenses and omit identifying metadata.

REPRODUCIBILITY STATEMENT

Section 3 describes construction of the 1,227-pair collection; Section 4 describes the dual-track evaluation and exclusions. The supplementary evaluation data, saved judgments, and recomputation scripts support Table 2. Appendix B records construction checks and model configurations, and Appendix F provides additional controls.

REFERENCES

Anthropic. Model system cards: Claude Opus 5. https://www.anthropic.com/system-cards, 2026.

Ding Chen, Simin Niu, Kehang Li, Peng Liu, Xiangping Zheng, Bo Tang, Xinchi Li, Feiyu Xiong, and Zhiyu Li. HaluMem: Evaluating Hallucinations in Memory Systems of Agents. *arXiv preprint arXiv:2511.03506*, 2025. URL https://arxiv.org/abs/2511.03506.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building Production-Ready AI Agents with Scalable Long-Term Memory. *arXiv preprint arXiv:2504.19413*, 2025. URL https://arxiv.org/abs/2504.19413.

DeepSeek-AI. DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence. https://arxiv.org/abs/2606.19348, 2026.

Xingbo Du, Loka Li, Duzhen Zhang, and Le Song. MemR^3: Memory Retrieval via Reflective Reasoning for LLM Agents. *arXiv preprint arXiv:2512.20237*, 2025. URL https://arxiv.org/abs/2512.20237.

Jizhan Fang, Xinle Deng, Haoming Xu, Ziyan Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, Huajun Chen, and Ningyu Zhang. LightMem: Lightweight and Efficient Memory-Augmented Generation. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.18866.

Xueyang Feng, Weinan Gan, Xu Chen, Quanyu Dai, and Yong Liu. How Does Personalized Memory Shape LLM Behavior? Benchmarking Rational Preference Utilization in Personalized Assistants. *arXiv preprint arXiv:2601.16621*, 2026. URL https://arxiv.org/abs/2601.16621.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating Models' Local Decision Boundaries via Contrast Sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, 2020. doi: 10.18653/v1/2020.findings-emnlp.117. URL https://aclanthology.org/2020.findings-emnlp.117/.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé, III, and Kate Crawford. Datasheets for Datasets. *Communications of the ACM*, 64(12):86–92, 2021. doi: 10.1145/3458723. URL https://cacm.acm.org/research/datasheets-for-datasets/.

Google DeepMind. Gemini 3.5 Flash Model Card. https://deepmind.google/models/model-cards/gemini-3-5-flash/, 2026.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation Artifacts in Natural Language Inference Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 107–112, 2018. doi: 10.18653/v1/N18-2017. URL https://aclanthology.org/N18-2017/.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. HippoRAG: Neurobiologically Inspired Long-Term Memory for Large Language Models. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2405.14831.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From RAG to Memory: Non-Parametric Continual Learning for Large Language Models. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 21497–21515, 2025. URL https://proceedings.mlr.press/v267/gutierrez25a.html.

Hakeem Hannoon, Andrew Zhao, Mihir Narayan, Sharvin Goyal, and Ivaxi Sheth. Mitigating Over-Personalization in LLMs via Structured Memory. *arXiv preprint arXiv:2608.08300*, 2026. URL https://arxiv.org/abs/2608.08300.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the Real Context Size of Your Long-Context Language Models? In *Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2404.06654.

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*, 2026. URL https://arxiv.org/abs/2601.13722.

Yuyang Hu, Shichun Liu, Yanwei Yue, Guibin Zhang, Boyang Liu, Fangyi Zhu, Jiahang Lin, Honglin Guo, Shihan Dou, Zhiheng Xi, Senjie Jin, Jiejun Tan, Yanbin Yin, Jiongnan Liu, Zeyu Zhang, Zhongxiang Sun, Yutao Zhu, Hao Sun, Boci Peng, Zhenrong Cheng, Xuanbo Fan, Jiaxin Guo, Xinlei Yu, Zhenhong Zhou, Zewen Hu, Jiahao Huo, Junhao Wang, Yuwei Niu, Yu Wang, Zhenfei Yin, Xiaobin Hu, Yue Liao, Qiankun Li, Kun Wang, Wangchunshu Zhou, Yixin Liu, Dawei Cheng, Qi Zhang, Tao Gui, Shirui Pan, Yan Zhang, Philip Torr, Zhicheng Dou, Ji-Rong Wen, Xuanjing Huang, Yu-Gang Jiang, and Shuicheng Yan. Memory in the Age of AI Agents. *arXiv preprint arXiv:2512.13564*, 2025. URL https://arxiv.org/abs/2512.13564.

Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo J. Taylor, and Dan Roth. Know Me, Respond to Me: Benchmarking LLMs for Dynamic User Profiling and Personalized Responses at Scale. In *Conference on Language Modeling*, 2025. URL https://arxiv.org/abs/2504.14225.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. Memory OS of AI Agent. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25961–25970, 2025. doi: 10.18653/v1/2025.emnlp-main.1318. URL https://aclanthology.org/2025.emnlp-main.1318/.

Tae Soo Kim, Yoonjoo Lee, Yoonah Park, Jiho Kim, Young-Ho Kim, and Juho Kim. CUPID: Evaluating Personalized and Contextualized Alignment of LLMs from Interactions. In *Conference on Language Modeling*, 2025. URL https://arxiv.org/abs/2508.01674.

Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra, Alireza Salemi, Ryan A. Rossi, Franck Dernoncourt, Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang, Shubham Agarwal, Nedim Lipka,

Chien Van Nguyen, Thien Huu Nguyen, and Hamed Zamani. LongLaMP: A Benchmark for Personalized Long-form Text Generation. *arXiv preprint arXiv:2407.11016*, 2024. URL https://arxiv.org/abs/2407.11016.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2005.11401.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic Evaluation of Language Models. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=iO4LZibEqW.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Mingfei Lu, Mengjia Wu, Feng Liu, Jiawei Xu, Weikai Li, Haoyang Wang, Zhengdong Hu, Ying Ding, Yizhou Sun, Jie Lu, and Yi Zhang. Choosing How to Remember: Adaptive Memory Structures for LLM Agents. *arXiv preprint arXiv:2602.14038*, 2026. URL https://arxiv.org/abs/2602.14038.

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

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, 2019. doi: 10.18653/v1/P19-1334. URL https://aclanthology.org/P19-1334/.

Niloofar Mireshghallah, Neal Mangaokar, Narine Kokhlikyan, Arman Zharmagambetov, Manzil Zaheer, Saeed Mahloujifar, and Kamalika Chaudhuri. CIMemories: A Compositional Benchmark for Contextual Integrity of Persistent Memory in LLMs. *arXiv preprint arXiv:2511.14937*, 2025. URL https://arxiv.org/abs/2511.14937.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220–229, 2019. doi: 10.1145/3287560.3287596. URL https://doi.org/10.1145/3287560.3287596.

Moonshot AI. Kimi K3: Open Frontier Intelligence. https://arxiv.org/abs/2607.24653, 2026.

OpenAI. GPT-6 Astra: A new generation of intelligence. https://openai.com/index/gpt-6-astra/, 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. URL https://arxiv.org/abs/2310.08560.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the ACM Symposium on User Interface Software and Technology*, 2023. URL https://arxiv.org/abs/2304.03442.

Sidharth Pulipaka, Oliver Chen, Manas Sharma, Taaha S Bajwa, Vyas Raina, and Ivaxi Sheth. PersistBench: When Should Long-Term Memories Be Forgotten by LLMs? In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.01146.

Qwen Team. Qwen3.8: Official Model Repository. https://github.com/QwenLM/Qwen3.8, 2026.

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. Zep: A Temporal Knowledge Graph Architecture for Agent Memory. *arXiv preprint arXiv:2501.13956*, 2025. URL https://arxiv.org/abs/2501.13956.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, 2020. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

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

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards Understanding Sycophancy in Language Models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.13548.

SpaceXAI. Introducing Grok 4.6. https://x.ai/news/grok-4-6, 2026.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.10813.

Jing Xu, Arthur Szlam, and Jason Weston. Beyond Goldfish Memory: Long-Term Open-Domain Conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5180–5197, 2022. doi: 10.18653/v1/2022.acl-long.356. URL https://aclanthology.org/2022.acl-long.356/.

Lingxiang Xu, Jiaoyun Yang, Min Hu, and Ning An. HUSH-Bench: Measuring Memory-Use Boundaries for Sensitive History in Conversational Agents. *arXiv preprint arXiv:2606.06055*, 2026. URL https://arxiv.org/abs/2606.06055.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-MEM: Agentic Memory for LLM Agents. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2502.12110.

Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Jinhe Bi, Kristian Kersting, Jeff Z. Pan, Hinrich Schütze, Volker Tresp, and Yunpu Ma. Memory-R1: Enhancing Large Language Model Agents to Manage and Utilize Memories via Reinforcement Learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12805–12825, 2026. doi: 10.18653/v1/2026.acl-long.583. URL https://aclanthology.org/2026.acl-long.583/.

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. URL https://arxiv.org/abs/2603.16557.

Z.ai. GLM-5.3. https://z.ai/blog/glm-5.3, 2026.

Jiawen Zhang, Kejia Chen, Jiachen Ma, Yangfan Hu, Lipeng He, Yechao Zhang, Jian Liu, Xiaohu Yang, Tianwei Zhang, and Ruoxi Jia. Beyond similarity: Trustworthy memory search for personal ai agents, 2026. URL https://arxiv.org/abs/2606.06054.

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. Do LLMs Recognize Your Preferences? Evaluating Personalized Preference Following in LLMs. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2502.09597.

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*, 2024. URL https://arxiv.org/abs/2305.10250.

TECHNICAL APPENDIX

A A complete paired example 16
B Construction checks and model configurations 16
C Dataset coverage 16
D Comparison with related benchmarks 17
D.1 What the benchmarks measure 17
E Extended related work 20
F Additional evaluation results and controls 21
F.1 Implementation details 21
F.2 Multiple-choice without memory 21
F.3 Memory-pool selectivity 21
F.4 Far companions 22
F.5 Relevance-gate apply head 22
F.6 Prompt wording 22
F.7 Memory presentation 22
F.8 Matched comparison across tracks 23
F.9 A resource-constrained coffee example 23
F.10 Free-generation selectivity 25
F.11 Additional input-removal statistics 25
G Possible mechanisms of preference misuse 25
G.1 Topical relevance as a substitute for applicability 25
G.2 Attention to preference and situation cues 26
G.3 Reasoning directed toward satisfying the preference 26
G.4 Candidate replies as comparison support 26
H Human validation of revised materials 26

A A COMPLETE PAIRED EXAMPLE

The following example is drawn from the 1,227-pair evaluation collection.

Target memory. User needs food genuinely hot, and finds anything less than mouth-numbing not worth ordering.

Shared request. What should I order?

Positive situation. I'm choosing my own dinner at the noodle place after clearing the evening, and nobody else is sharing from my bowl.

Negative situation. I'm choosing our dinner at the noodle place after clearing the evening, and the person sharing from my bowl has been hiccuping after the table sauce.

Shared replies.

r_{M} Order Sichuan hot pot with extra bird's-eye chilies, request the highest heat level, and get chili oil on the side.

r_{sup} Order a lighter Sichuan hot pot at medium heat, keep chili oil separate, and let broth carry the flavor in another lane.

r_{obl} Order a mild beef noodle soup with hand-pulled noodles, ask for clear broth, and keep the chili oil on the side.

r_{irr} Check whether the restaurant is still delivering tonight. Tony's Pizza usually stops taking online orders at 9:45, even when Google says 10.

B CONSTRUCTION CHECKS AND MODEL CONFIGURATIONS

The evaluated data are stored in `gen/out/paper_fullreg/verified.jsonl`. The released `eval_1227.jsonl` preserves the same preference, request, situations, and candidate replies. Three source collections contribute 161, 549, and 517 uniquely identified pairs.

Generation uses `gpt-5.5`; verification uses `deepseek-v4-flash-0731`; the test without context uses `qwen3.6-flash`. The saved verification fields cover the applicability flip, minimal-pair structure, inference and commonsense requirements, reply roles, acknowledged departure, positive-situation tier, constraint strength, request naturalness and sufficiency, concrete recommendations, category fit, positive-side feasibility, and the availability of answers on both sides. Rule-based filters complement these judgments with checks on text and option form. All retained records have a saved verification verdict of `PASS`; this records filter acceptance, not independently measured correctness.

The test without context shuffles the four replies and asks the model to select the most appropriate reply from the request and candidates alone. It makes three calls at temperature 1 and retains candidates with at most one selection of the preference-enacting reply. Among retained pairs, 1,056 have zero such selections and 171 have one. These are selection statistics from the construction filter, not a held-out test of robustness to answer cues.

C DATASET COVERAGE

The collection comprises 1,227 pairs, 45 preferences, eight situation categories, and six domains. Positive situations include 620 clean and 607 distractor cases. These labels describe construction intent. Domain counts are food 345, film 236, activities 279, goods 171, routines 104, and social interaction 92. Context-to-request distances of one, three, and six rounds contain 335, 543, and 349 pairs, respectively. A distance of d uses $d - 1$ filler exchanges.

Table 5: Situation coverage in PairPref.

Situation category	Pairs
Companions	205
Occasion	163
Time constraint	136
Beneficiary shift	296
Role shift	54
Resource constraint	110
Physical state	191
Emotional state	72
Total	1,227

D Comparison with related benchmarks

Tables 7, 6, and 8 compare evaluation scope and construction evidence. Counts retain each benchmark's unit of evaluation; construction filtering and independent evaluator validation are distinct.

Table 6: Protocol details for selected benchmarks in Table 1.

Benchmark	Comparison or variation	Boundary assessment	Usefulness assessment
PrefEval	Explicit or implicit preferences; conversation length	Preference following	Helpfulness in generation
RPEval	Single/multiple and explicit/implicit preferences	Intent matching; strategy errors	Feasibility and verbosity errors
OP-Bench	Memory methods and memory-free BASE	Irrelevance, sycophancy, repetition	LoCoMo memory utility evaluated alongside
BenchPreS	User–recipient–task combinations; with/without preferences	Misapplication and appropriate application	Task completeness, separately scored
CIMemories	User attributes across recipient–task contexts	Inappropriate attribute disclosure	Coverage of necessary attributes
PersistBench	Leakage, sycophancy, and beneficial-use subsets	Failure rates on targeted scenarios	Separate beneficial-use control subset
HUSH-Bench	Matched memory/no-memory; fixed-target invitation pairs	Unsolicited integration, exposure, over-scope	Helpfulness and invited target uptake
This work	Same memory, request, candidates; context varies	Preference use and withholding	Situational fit and pair success in both tracks

D.1 What the benchmarks measure

PairPref evaluates changes in the applicability of a persistent preference. Its unit is a pair with the same still-valid preference and request, but different situations that warrant applying or setting aside that preference. Selection keeps the candidate replies fixed; free generation removes the candidates. Evaluating both sides measures whether appropriate application and appropriate non-application occur within the same pair. The following comparisons distinguish this target from related evaluations of preference inference, utilization strategies, communication norms, and information disclosure.

CUPID: inferring contextual preferences. CUPID constructs interaction histories in which preferences are gradually revealed through user feedback (Kim et al., 2025). Context factors, such as

Table 7: Scope of related preference and memory benchmarks.

Benchmark	Memory object	Evaluation target	Response
LoCoMo (Maharana et al., 2024)	Long dialogues	Recall, event summaries, multimodal continuity	Generation
LongMemEval (Wu et al., 2025)	Timestamped histories	Recall, reasoning, updates, abstention	QA
PrefEval (Zhao et al., 2025)	Explicit or implicit preferences	Preference following over conversations	MCQ; generation
PersonaMem (Jiang et al., 2025)	Evolving user profiles	Current-profile alignment and transfer	MCQ; generation
CUPID (Kim et al., 2025)	Context-dependent preferences	Preference inference and response alignment	Inference; generation
RPEval (Feng et al., 2026)	Preferences linked to query intent	Ignore, support, or dominate	Intent MCQ; generation
OP-Bench (Hu et al., 2026)	Dialogue-derived profiles	Irrelevance, sycophancy, repetition	Generation
BenchPreS (Yoon et al., 2026)	Response preferences	Selectivity under communication norms	Generation
CIMemories (Mireshghallah et al., 2025)	Personal information attributes	Context-appropriate disclosure	Generation
PersistBench (Pulipaka et al., 2026)	User memories and beliefs	Leakage, sycophancy, beneficial use	Generation
HUSH-Bench (Xu et al., 2026)	Marked sensitive history	Current-turn warrant for memory use	Generation
This work	Persistent content preferences	Applicability, behavior, and task completion across contexts	Selection; generation

people, activities, and tools, are associated with preferences; contrasting factors can have conflicting preferences, and preferences can also change over time. Given a new request and prior sessions, evaluation measures preference inference and response alignment. CUPID therefore already tests contextual personalization. PairPref instead supplies a single preference whose validity is unchanged and tests whether situational facts warrant acting on it. Its contrast concerns whether to apply that preference, rather than which preference the history associates with the current context.

RPEval: choosing a utilization strategy. RPEval annotates preference–query instances with *Ignore*, *Support*, or *Dominate* (Feng et al., 2026). Its discriminative task predicts these strategies, while its generative evaluation examines strategy errors and response-level problems such as low feasibility and verbosity. It includes preferences that are topically connected but inappropriate for the task, such as a preference for challenging routes in a family outing with children. Thus, neither contextual applicability nor evaluating discrimination alongside generation is unique to PairPref. PairPref selects among concrete replies held fixed across opposite-applicability situations, and evaluates generated responses on the same pairs, including whether both sides are appropriate.

BenchPreS: expression preferences under communication norms. BenchPreS combines 10 user profiles with 39 recipient–task contexts in five formal communication domains (Yoon et al., 2026). Its preference attributes concern role, style, tone, markers, and nickname. Human labels specify which attributes should be applied or suppressed; generated responses are scored using Misapplication Rate and Appropriate Application Rate, with additional task-completeness and prompt-mitigation analyses. It directly measures contextual preference selectivity. PairPref emphasizes preferences governing answer content across eight situation categories and holds the request fixed within each contrast. Its paired scoring checks application and non-application for the same preference–request combination, complementing attribute-level rates across contexts.

Table 8: Reported scale and validation in the reviewed benchmark papers.

Benchmark	Reported scale and unit	Reported validation	Source locator
LoCoMo	50 dialogues in the paper's collection	Human editing for consistency, images, and event grounding	Secs. 1, 3.4; Table 1
LongMemEval	500 questions; configurable histories	Human rewriting of questions and editing of evidence; judge meta-evaluation	Secs. 3.2–3.3; Table 6
PrefEval	1,000 base pairs in 3 forms: 3,000 preference–query pairs	Human-assisted curation; manual check of 200 sampled evaluations	Secs. 2.2, 2.5
PersonaMem	About 6,000 query–response pairs; 20 personas	90 entries from one persona, assessed by three author annotators	Sec. 2; App. B
CUPID	756 instances; 252 per instance type	Human validation and revision; separate preference-matcher meta-evaluation	Secs. 2.3, 3.2–3.3
RPEval	8,255-sample pool; 953-sample finalized test set	Manual verification and adjudication; judge–human ordinal agreement	Secs. 3.1–3.2
OP-Bench	1,700 instances; 20 users	Two independent reviews per instance; adjudication for disagreement	Secs. 3.3–3.4
BenchPreS	1,950 attribute-level instances: 10 users, 39 contexts, 5 preferences	Human applicability labels; judge comparison on 100 sampled instances	Secs. 3.2–3.3, 6
CIMemories	10 evaluated profiles; 49 seed contexts, up to 189 attributes per profile	Persona-conditioned model labels; ambiguous attribute–context pairs excluded	Secs. 4.1, 5.1
PersistBench	500 samples: 200 leakage, 200 sycophancy, 100 beneficial use	Search-model and held-out-model filtering, then human verification	Secs. 4.1–4.2
HUSH-Bench	2,400 probe cases; 10 personas	600 responses, two human annotators each; second-judge evaluation	Secs. 3.2–3.3; Apps. A.1, B.4, D.1
This work	1,227 context pairs; 45 target preferences	Construction filters; two independent annotators assess 200 revised sample pairs	Sec. 3.3; App. H

OP-Bench: over-personalization failures. OP-Bench evaluates irrelevance, sycophancy, and repetition (Hu et al., 2026). Its irrelevance subset includes both unrelated queries and deceptively related queries, so its target extends beyond simple topical filtering. Generation is assessed for unnecessary personalization, objectivity, and response diversity; useful memory performance is evaluated alongside on LoCoMo. PairPref instead constructs both an appropriate-use and an inappropriate-use situation for each fixed preference–request pair. This makes selective use within a pair the evaluation unit, rather than combining over-personalization tests with a separate memory-utility benchmark.

CIMemories: contextual information disclosure. CIMemories composes personal attributes with recipient–task contexts and labels attributes as necessary or inappropriate to disclose (Mireshghallah et al., 2025). The same attribute can receive different labels across tasks. Evaluation measures inappropriate disclosure and coverage of necessary information, while varying memory composition and examining repeated interactions. This assesses contextual boundaries on information flows. PairPref tests preference-guided content without requiring disclosure.

HUSH-Bench: warrant for sensitive-history use. HUSH-Bench tracks marked sensitive disclosures, compares matched memory and no-memory responses, and examines whether retrieved targets enter generated answers (Xu et al., 2026). Its matched invitation experiment fixes a target-only memory block and adds a generic request to use prior context, measuring target uptake, helpfulness, and over-scope. It therefore already includes controlled paired interventions. PairPref changes sit-

uational facts rather than an explicit invitation to use history, and evaluates the appropriateness of applying a persistent preference. Fixed candidate replies additionally support a response-selection test on the same contrasts used for generation.

PersistBench: leakage, sycophancy, and beneficial use. PersistBench evaluates cross-domain leakage and memory-induced sycophancy, with a separate beneficial-memory subset that checks whether models can still use helpful information (Pulipaka et al., 2026). Its memory–query cases are constructed to expose failures, and the subsets report their respective failure rates. PairPref links appropriate use and non-use within each fixed preference–request pair. This measures joint success under a situational reversal rather than across separate harmful-use and beneficial-use subsets.

Interpreting the comparison. These distinctions concern evaluation objects and protocols, not an absence of contextual reasoning in prior benchmarks. Table 1 distinguishes direct capability evaluation from related coverage. Its selection and generation columns concern one still-valid preference and a fixed request, assessed through concrete reply selection or free generation; changing preferences, changing tasks, strategy classification, or explicitly inviting memory use provide related evidence. The outside-scope symbol denotes the evaluation target, not a demonstrated inability of the evaluated models. Disclosure benchmarks remain relevant even when they do not directly evaluate preference enactment. Pair success complements marginal rates: the same per-side accuracy can yield different rates of joint correctness.

E Extended related work

Memory organization and control. Agent memory research studies how information is formed, retained, and accessed across interactions (Hu et al., 2025). Persistent storage and consolidation combine experience retrieval, reflection, and user modeling (Park et al., 2023; Zhong et al., 2024). Memory organization addresses how these records are structured: tiered storage manages information at different scales (Packer et al., 2023; Kang et al., 2025), while linked notes and temporal graphs preserve relationships among entries (Xu et al., 2025; Rasmussen et al., 2025). Retrieval-augmented generation combines language models with external knowledge (Lewis et al., 2020); HippoRAG and its successor use structured retrieval for knowledge integration (Gutiérrez et al., 2024; 2025). Recent approaches address the cost of persistent-memory pipelines (Chhikara et al., 2025; Fang et al., 2026) and adapt memory operations, retrieval actions, or structure selection to the task (Yan et al., 2026; Du et al., 2025; Lu et al., 2026). These directions improve the information available to an assistant. We examine the next decision: how an available preference should influence the answer in the current situation.

Memory and personalization benchmarks. Existing evaluations test whether an assistant maintains information across interactions and applies it to a user-specific task. Multi-session dialogue and long-term memory benchmarks assess conversational continuity, recall, reasoning, updates, and abstention (Xu et al., 2022; Maharana et al., 2024; Wu et al., 2025). Operation-level evaluation distinguishes errors introduced during extraction, updating, and question answering (Chen et al., 2025). Personalization benchmarks assess user-conditioned classification and generation, long-form writing, and preference following (Salemi et al., 2024; Kumar et al., 2024; Zhao et al., 2025). Dynamic profiles and contextual preferences extend this setting to changes in user state and interaction history (Jiang et al., 2025; Kim et al., 2025). We hold the preference and request fixed to examine how a change in situation affects the preference's applicability.

Selective personalization and contextual boundaries. Appropriate memory use depends on more than topical relevance. Recent benchmarks examine irrational preference integration, over-personalization, and preference suppression under communication norms (Feng et al., 2026; Hu et al., 2026; Yoon et al., 2026). Related work studies cross-domain leakage and memory-induced sycophancy, including mitigation through structured memory presentation (Pulipaka et al., 2026; Hannoon et al., 2026). Sycophancy concerns agreement with user beliefs at the expense of truthfulness (Sharma et al., 2024); preference applicability concerns whether a recommendation suits the present task. CIMemories evaluates contextual disclosure violations alongside coverage of necessary attributes (Mireshghallah et al., 2025). HUSH-Bench examines unsolicited use of sensitive history, with matched memory/no-memory and explicit-invitation controls (Xu et al., 2026). These

studies establish contextual boundaries as an evaluation target. We test contextual contrasts within a shared preference–request pair and examine what candidate texts reveal about model choices.

Behavioral testing and evaluator reliability. Behavioral tests and contrast sets evaluate expected responses to controlled changes (Ribeiro et al., 2020; Gardner et al., 2020). A high score need not reflect the intended capability: annotation artifacts and superficial heuristics can provide another route to the answer (McCoy et al., 2019; Gururangan et al., 2018). Multi-dimensional evaluation makes such distinctions explicit under standardized conditions (Liang et al., 2023). These principles motivate assessing contextual applicability, candidate cues, and answer completion separately. Language-model judges make this assessment scalable, and the construction checks and response-scoring protocol serve different purposes (Sections 3.3 and 4.1).

F Additional evaluation results and controls

These analyses use the 1,227-pair collection from Section 4. Each table reports the models with available results for that condition. Use$^+$ and Use$^-$ denote MU$^+$ and MI$^-$, respectively; rates are percentages and Δ is in percentage points. Generation metrics use the retained pairs defined in Section 4.1. Dashes indicate unavailable results, not zero scores.

F.1 Implementation details

The two sides use separate calls. Side-dependent seeds determine memory order, candidate order, and filler exchanges; these presentation differences remain in the no-situation control. Temperature is 0 where supported and omitted at endpoints that reject it. DeepSeek V4 Flash scores every free-answer model, including DeepSeek V4 Pro. Uptake judgments use the preference, request, and answer; separate situational-fit judgments use the current situation, request, and answer.

The controls vary memory availability, situation availability, pool composition, or presentation. No-situation removes situation turns; no-memory removes the memory block. Top-5 retains the target and four near companions; Far-15 retains the target with 15 far companions. Bare removes the relevance clause, while Mitigate explicitly distinguishes a preference's truth from its applicability. User and dialog place the same memories in user messages or prior dialogue. The preference-only classifier receives the target preference and request. The relevance gate also receives the situation and predicts relevance, sufficiency, and applicability separately. Neither receives a generated answer.

F.2 Multiple-choice without memory

Table 9: Multiple-choice results with the memory block removed.

Model	MU$^+$	MI$^-$	Δ
Gemini 3.5 Flash	39.7	2.4	37.3
GPT-6 Astra	32.5	2.2	30.3
Kimi K3	34.1	2.9	31.1
GLM-5.3	34.5	2.2	32.3
DeepSeek V4 Pro	30.2	2.5	27.7

Figure 4a shows the same comparison. Without a memory block, negative-side selection of r_{M} is 2.2%–2.9%. Positive-side selection remains 30.2%–39.7%. The remaining gap shows that situation and candidate content can favor the preference-enacting reply even without an explicit memory block; it does not demonstrate retrieval of a withheld preference.

F.3 Memory-pool selectivity

Table 10 supplements Table 4 with selectivity (Δ) for the same pool-size controls.

Table 10: Memory-pool controls: negative-side use and selectivity.

	Use$^-$		Δ	
Model	Top-16	Top-5	Top-16	Top-5
Gemini 3.5 Flash	22.9	25.8	56.0	55.7
DeepSeek V4 Pro	23.0	27.8	53.5	50.8
Kimi K3	18.3	20.7	57.6	59.0
GLM-5.3	18.4	21.0	56.6	56.6
GPT-6 Astra	14.7	11.7	56.6	61.5

Table 11: Multiple-choice results with the target and 15 far companions.

	MI$^-$		Δ	
Model	Far-15	Top-16	Far-15	Top-16
Gemini 3.5 Flash	29.2	22.9	53.5	56.0
DeepSeek V4 Pro	23.6	23.0	52.8	53.5
Kimi K3	22.9	18.3	56.4	57.6
GLM-5.3	21.1	18.4	55.5	56.6
GPT-6 Astra	12.8	14.7	62.4	56.6

F.4 Far companions

Relative to Top-16, Far-15 changes Δ by -2.5 to $+5.8$ points. Negative-side use does not collapse when the extra memories are unrelated. Both pool edits preserve the main selection pattern (Table 4).

F.5 Relevance-gate apply head

Table 12: Answer-free classifiers on the same 1,227 pairs.

Condition	MU$^+$	MI$^-$	Δ
Preference-only	99.4	99.4	0.0
Gate apply	87.9	25.2	62.7
Both sides relevant		98.6	
Both sides sufficient		61.3	

The preference-only classifier makes one decision per preference–request pair without seeing either situation; its 99.4% application rate therefore appears on both sides. The relevance gate sees each situation separately. Both-sides relevance and sufficiency require the corresponding decision to be positive on both sides. The apply output explicitly distinguishes relevance from applicability; its Δ measures contextual application decisions, not retrieval performance.

F.6 Prompt wording

Figure 5 summarizes these comparisons. The mitigate instruction reduces positive-side multiple-choice use for seven of eight models. Flash is the exception (78.9% to 80.3%). Most models thus reduce both appropriate and inappropriate use, creating a trade-off.

F.7 Memory presentation

Figure 4b reports these rates; dialog raises Qwen 3.8 Max from 20.3% to 27.1%.

Table 13: Effect of prompt wording on multiple-choice and free-answer performance.

	MCQ MI$^-$			Always-use		
Model	Nudge	Bare	Mitigate	Nudge	Mitigate	Gen-OK mitigate
Claude Opus 5	11.5	11.5	3.6	88.7	—	—
GPT-6 Astra	14.7	16.3	7.9	71.0	54.0	26.8
Kimi K3	18.3	20.0	12.0	83.6	85.5	9.5
GLM-5.3	18.4	20.0	8.1	90.8	85.2	10.3
Qwen 3.8 Max	20.3	22.2	11.0	81.2	—	—
Gemini 3.5 Flash	22.9	25.3	19.0	75.1	72.5	20.8
DeepSeek V4 Pro	23.0	25.7	15.2	81.3	84.6	11.9
Grok 4.6	57.4	57.0	39.0	92.0	—	—

Table 14: Positive-side multiple-choice use under nudge versus mitigate.

Model	Nudge	Mitigate
Claude Opus 5	76.4	67.6
GPT-6 Astra	71.2	66.0
Kimi K3	76.0	68.9
GLM-5.3	75.1	66.1
Qwen 3.8 Max	71.6	67.6
Gemini 3.5 Flash	78.9	80.3
DeepSeek V4 Pro	76.4	72.4
Grok 4.6	75.1	69.6

Table 15: Effect of memory presentation on negative-side multiple-choice use.

Model	System list	User block	Dialog turns
Claude Opus 5	11.5	12.0	12.7
GPT-6 Astra	14.7	16.3	20.1
Kimi K3	18.3	19.6	20.6
GLM-5.3	18.4	20.0	20.9
Qwen 3.8 Max	20.3	23.1	27.1
Gemini 3.5 Flash	22.9	24.6	25.6
DeepSeek V4 Pro	23.0	26.2	27.5
Grok 4.6	57.4	58.4	58.4

F.8 Matched comparison across tracks

We join default-condition selection and generation records by sample identifier within each model. Common pairs have valid uptake judgments on both generated answers and neither answer flagged as hedging. Selection success requires r_{M} on the positive side and r_{sup} or r_{obl} on the negative side. The final column divides pairs with $u^+ \geq 0.5$ and $u^- \geq 0.5$ by selection-successful common pairs. Situational-fit judgments are not required for this uptake comparison.

F.9 A resource-constrained coffee example

The example is `M07-C6-c26634`, evaluated with Claude Opus 5. It illustrates a scored failure and is not used to estimate its frequency.

Preference. User only drinks hand-poured single-origin, and rules out instant and chain coffee.

Table 16: Preference use in generation after successful paired selection.

Model	Common pairs	Selection success	Use both	Rate (%)
Claude Opus 5	1067	719	641	89.2
Kimi K3	1091	654	561	85.8
GPT-6 Astra	1068	622	455	73.2
GLM-5.3	1056	615	567	92.2
Gemini 3.5 Flash	1065	613	440	71.8
DeepSeek V4 Pro	1107	624	517	82.9
Qwen 3.8 Max	1011	549	446	81.2
Grok 4.6	1153	272	245	90.1

Shared request. What should I order?

Positive situation. I'm settled at the market hall counter and they asked me to pick, with the afternoon clear and my wristband still loaded for whatever they can make.

Negative situation. I'm settled at the market hall counter and they asked me to pick, with the afternoon clear and my wristband showing only enough credit for the self-serve urn.

Selected positive reply. Order the Ethiopian pour-over, preferably a washed Yirgacheffe if they have it. Ask for it black so the florals stay clear.

Selected negative reply. Order the house drip coffee, medium roast if labeled. Take it black first, then add a small splash if it tastes harsh.

Generated positive answer. Ask whoever's on bar which single-origin they've got set up for pour-over right now, and let them pick the one they're most excited about. That usually lands better than reading the board, and it's exactly the kind of thing a counter person enjoys being asked. Worth asking how it's prepped too — whether it's a V60, a Kalita, whatever they're dialed in on that day.

For food, see if there's a regional pastry from a small baker in the hall rather than the standard case stuff. And since you're in a market hall with a clear afternoon, the fermented and preserved stalls are worth a slow walk after the coffee — pickles, krauts, anything cured. That's a good browse even if you only end up with one jar.

Keeps the whole thing modest, and you'll leave knowing something you didn't walk in with.

Generated negative answer. Skip the urn. It's the one thing on offer that doesn't fit how you drink coffee — batch-held drip isn't going to resemble a hand-poured single origin, and paying for it just to have a cup in hand is a poor trade.

What the wristband is actually good for at a market hall counter is food. Ask whoever's behind it what's made in-house and what came in from a producer they know — that's usually where the interesting answer lives. A regional pastry or something from the ferment/preserve side of the counter tends to be the better use of limited credit than a mediocre coffee, and it's the kind of thing you can ask a real question about: who bakes it, what the starter or brine is, whether it's a local tradition.

Then let the coffee wait. Your afternoon is clear and three kilometres is only 35–40 minutes on foot, so there's easy room to walk to a place that pours to order. Worth asking the counter staff directly where they go on their own break — that'll get you somewhere better than searching ratings.

The saved scores are $u^+ = u^- = 1$, with situational fit 1 on the positive side and 0 on the negative side; neither generated answer is flagged as hedging.

F.10 FREE-GENERATION SELECTIVITY

Table 17 reports the generation uptake gap omitted from the main table. These values use the generation scoring rule and the same retained pairs as the main results.

Table 17: Selectivity in free generation.

Model	Δ (points)
Claude Opus 5	1.5
Kimi K3	6.8
GPT-6 Astra	10.9
GLM-5.3	3.3
Gemini 3.5 Flash	14.0
DeepSeek V4 Pro	10.7
Qwen 3.8 Max	8.3
Grok 4.6	2.8

F.11 ADDITIONAL INPUT-REMOVAL STATISTICS

Table 18: Additional statistics for input-removal controls.

	No situation				No memory, free answer		
Model	MCQ Δ	Gen Δ	MCQ Use$^+$	MCQ Use$^-$	Δ	MU$^+$	MI$^-$
Gemini 3.5 Flash	0.0	-1.8	90.0	90.0	3.6	11.1	7.5
Kimi K3	-0.2	-1.0	89.6	89.9	1.9	15.1	13.1
DeepSeek V4 Pro	-0.5	0.4	89.1	89.6	2.4	11.7	9.3
GLM-5.3	-0.7	-0.3	87.2	87.9	0.5	16.3	15.8
GPT-6 Astra	1.4	2.1	83.9	82.6	0.1	11.4	11.3
Qwen 3.8 Max	—	-0.4	—	—	0.3	11.7	11.4

The no-situation Use columns concern selection. Dashes denote unreported results, not zero scores. Generation applies paired output-validation and hedging exclusions per condition.

G POSSIBLE MECHANISMS OF PREFERENCE MISUSE

PairPref reveals a gap between selecting an appropriate reply and producing one under the same situations. Several mechanisms could contribute to this pattern, and they may operate together. The following hypotheses connect the behavioral findings to predictions that future experiments can test; they are not measurements of the evaluated models' internal computation.

G.1 TOPICAL RELEVANCE AS A SUBSTITUTE FOR APPLICABILITY

A model may treat a true preference that matches the request as sufficient grounds for using it. In PairPref, both properties remain intact even when the situation makes preference use inappropriate. The situation-blind classifier recommends application for 99.4% of pairs, while the gate marks both sides relevant for 98.6%. These results illustrate why relevance alone cannot resolve applicability. One hypothesis is that generation relies on a similar shortcut. A direct test would compare relevance-only instructions with a separate applicability decision conditioned on the full situation, followed by answer generation. Scoring the intermediate decision and the final answer separately would distinguish incorrect judgments from failures to follow a correct judgment. Evaluation should measure appropriate use and misuse with inputs fixed.

G.2 Attention to preference and situation cues

The retrieved preference may exert a more persistent influence during decoding than the situational constraint. A preference directly specifies desired answer content, whereas recognizing a constraint can require connecting several details. Unequal attention to these cues is therefore a plausible contributor. Misuse with shorter pools and alternative presentations suggests that a simple account based only on context length or one memory location is insufficient. Follow-up experiments could vary cue position and explicitness while preserving meaning. For models with accessible internals, attention and activation analyses could identify candidate components, followed by targeted interventions to test whether they change contextual preference use. Attention weights alone cannot establish causality; interventions should reduce misuse while preserving appropriate use.

G.3 Reasoning directed toward satisfying the preference

Once a model adopts preference satisfaction as its objective, additional reasoning may produce ways to preserve the preference despite the situation. The coffee example in Appendix F.9 is consistent with this possibility: the answer proposes another shop or a later purchase rather than selecting within the stated budget. This behavior motivates a distinction between reasoning about how to satisfy a preference and reasoning about whether to apply it. A controlled comparison could evaluate direct answering, generic step-by-step reasoning, and an explicit applicability-first procedure under comparable generation budgets. The hypothesis predicts that generic reasoning need not improve pair success, whereas checking applicability first may reduce workarounds. Written rationales may reveal justification patterns without establishing the underlying computation.

G.4 Candidate replies as comparison support

Selection supplies explicit alternatives, including replies that set the preference aside while meeting the request. Free generation requires the model to construct those alternatives itself. This difference may help explain why a model selects the intended replies yet uses the preference in both generated answers for the same pair. A follow-up could compare direct generation with a procedure that first drafts both a preference-enacting answer and a preference-withholding answer, then chooses between them using the situation. Comparing model-generated alternatives with the benchmark candidates would help separate difficulty producing an appropriate alternative from difficulty choosing it. Evaluation should retain the generation track's situational-fit criteria: withholding a preference alone does not ensure an appropriate answer.

H Human validation of revised situations and candidate replies

Sample and protocol. Two annotators independently assessed 200 pairs spanning 45 preferences, eight situation categories, and six domains. Sampling assigned one pair per preference, then allocated 155 proportionally by largest remainder and sampled without replacement. Before annotation, 112 situations in 107 pairs were revised; preferences, requests, and replies were unchanged. This assessment concerns revised materials, separately from the 1,227-pair model evaluation.

Sheets randomized situations and options and hid original labels and automated judgments. Annotators first judged 400 situations for applicability, continued preference validity, and request naturalness and answerability, then checked pair control and selected all acceptable replies. Labels allowed uncertainty, distinguished from an empty acceptable set. Results retain independent, pre-adjudication judgments.

Inter-annotator agreement. Table 19 reports exact agreement and unweighted Cohen's κ. Applicability agreement is 99.25%, with $\kappa = 0.985$. Resampling complete pairs 5,000 times gives descriptive percentile 95% intervals of [98.25%, 100.00%] for agreement and [0.965, 1.000] for κ, preserving dependence between the two situations in a pair.

Both annotators marked all 400 preferences as still valid. For request answerability, they marked 397 and 394 situations as acceptable, respectively; their three and six negative judgments did not overlap. The observed agreement of 97.75% is slightly below the 97.77% expected from these marginal distributions, yielding $\kappa = -0.010$. For pair control, one annotator accepted all 200 pairs

Table 19: Inter-annotator agreement on revised materials.

Dimension	Judgments	Agreement (%)	Cohen's κ
Preference applicability	400	99.25	0.985
Continued preference validity	400	100.00	—
Request naturalness and answerability	400	97.75	−0.010
Controlled pair comparison	200	99.50	0.000
Candidate acceptability	1,580	87.72	0.728

Candidate judgments exclude five situations with an uncertain label from either annotator. A dash denotes undefined κ because both annotators used a single category.

and the other accepted 199, yielding $\kappa = 0$. These distributions explain why high raw agreement can coexist with low or undefined κ: agreement on the dominant category does not establish agreement on the exceptions.

For candidate acceptability, each selected option is treated as accepted and each unselected option as rejected. The five uncertain situations are excluded, leaving 395 situations and 1,580 binary option judgments. An explicitly empty set contributes four rejections. Exact agreement on the complete acceptable set is 62.0% across all 400 situations. Mean Jaccard similarity is 0.771 across the 395 jointly determinate situations, with two empty sets assigned similarity one.

Table 20: Joint human confirmation of labels and candidate replies.

Criterion	Count	Rate (%)
Situations and labels		
Original positive side is applicable	183/200	91.5
Original negative side is inapplicable	180/200	90.0
Both sides match the original direction	180/200	90.0
Applicability contrast, either direction	196/200	98.0
Candidate replies		
Positive-side r_{M} is acceptable	157/200	78.5
Negative-side r_{M} is unacceptable	165/200	82.5
No candidate is acceptable	20/400	5.0

Each row requires confirmation from both annotators. Either-direction contrasts include 16 pairs opposite to the original labels. Rates are unweighted sample statistics.

Support for the original labels. Table 20 retains the original positive/negative mapping, verified against identifiers and both annotation sheets. Both annotators confirm the original direction for 180 pairs and the opposite direction for 16 revised pairs. Thus, 98.0% confirm a contrast, while 90.0% support its original direction. All 93 unchanged pairs receive joint support for their original direction.

Candidate quality and scope. Candidate judgments are less uniform than applicability judgments. Both annotators find no acceptable reply in 20 situations; at least one does so in 46. These cases and the direction reversals require review, not automatic relabeling. The reported rates describe the revised sample, not correctness of the original 1,227 pairs. Incorporating revisions requires updating affected model evaluations. No generated answers were annotated, so these results do not validate the automatic free-generation scorer.