ADAPTIVE PERSONALIZATION UNDER PREFERENCE EPISTEMIC UNCERTAINTY

Anonymous authors
Paper under double-blind review

ABSTRACT

Human preferences change over time: an established preference may remain valid over time, become uncertain as context shifts, or be explicitly revised. Effective personalization requires conversational AI to adapt its response strategy to evolving preferences rather than relying on passive memory retrieval. We formalize this capability as *adaptive personalization under epistemic uncertainty*. A human preference study further confirms that people favor state-contingent response strategies as preference uncertainty changes. To evaluate whether language models mirror this adaptation, we create a multilingual benchmark spanning 250 personas, 5 languages, and 25K conversational turns that decouples strategy selection from response execution. Across eight closed- and open-weight models, we uncover a *recognition–generation gap*: models reliably identify appropriate strategies when selecting among candidate options, but collapse during open-ended generation, suffering a 35-46% drop in strategy calibration. Crucially, this failure is systematic: models default to exploit-first behavior, over-relying on existing preferences regardless of preference epistemic uncertainty. We show this failure is addressable: while prompting interventions fail, targeted training that supervises models' internal strategy representations toward state-conditional human preferences narrows the user-acceptability generalization gap by a third on out-of-distribution personas and conversational styles. Code and datasets will be released upon acceptance.

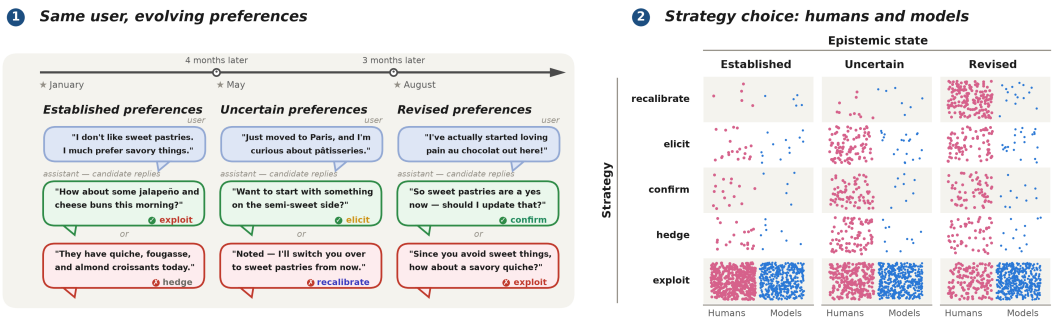


Figure 1: **Humans prefer varying response strategies based on preference epistemic uncertainty; models exploit regardless. (Left)** One user at three time points, illustrating the three preference epistemic states we evaluate: 1) *established*, 2) *uncertain*, and 3) *revised*, each paired with a well-matched and a mismatched assistant reply. **(Right)** Across the three epistemic states (preference uncertainties), humans prefer responses with varying strategies, whereas current models predominantly *exploit* in all states.

1 INTRODUCTION

Human preferences evolve dynamically across interactions: an established preference may remain valid over time, become uncertain as context shifts, or be explicitly revised by new evidence. For conversational AI, adapting to these shifts requires far more than passive memory retrieval. It demands deciding whether a remembered preference should still be acted upon, questioned, hedged

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.

around, or updated. Consider the user in Figure 1, who states a preference for savory food, later mentions exploring local bakeries, and eventually says she has come to enjoy something sweet, without ever restating her preferences. An assistant that stubbornly mentions her preference for savory foods across all three turns has retrieved the user's history correctly, but failed to adjust the conversational strategy to match the evolving reliability of that preference.

We formalize this capability as *adaptive personalization under epistemic uncertainty*. Evidence regarding user preferences evolves across three distinct epistemic uncertainty states: **established** (explicitly stated and uncontradicted), **uncertain** (ambiguous or partial context), and **revised** (superseded by new evidence). Each state demands a distinct conversational behavior: acting directly on settled preferences, seeking clarification under ambiguity (dialogue grounding; Clark, 1996; Clark & Brennan, 1991; Traum, 1994), or course-correcting when preferences change (belief revision; Alchourrón et al., 1985; Gärdenfors, 1988; Hansson, 1999). A controlled human study confirms this state-contingent behavior: humans systematically prefer distinct model response strategies across epistemic states (§2.2).

Despite human demand for dynamic strategy adaptation, existing personalization benchmarks predominantly evaluate model retrieval of preferences over long-horizon interactions (Jiang et al., 2025a; Zhao et al., 2025a; Jiang et al., 2025b; Harry et al., 2026; Guo et al., 2026), but fail to test whether conversational strategy adapts to preference reliability over time. To measure this distinct capability, we create a benchmark spanning 250 personas, five languages, and 25,006 conversational turns, with oracle epistemic-state labels and an evaluation framework that decouples strategy selection from response execution (§2).

Evaluating eight closed- and open-weight models demonstrates that models' failures localize to generation rather than recognition (§3). Across varying epistemic states, models select responses with appropriate strategies at a high accuracy when presented with candidate options, but fail to generate responses – a gap widening to over 35% under epistemic states: uncertain and revised. Moreover, models collapse to an exploit-first default generation policy that over-relies on historical preferences regardless of preference epistemic uncertainty.

To address this default behavioral policy in models, we first test inference-time prompting interventions (§4.1); however, directly providing preference context fails to close the recognition-generation gap. Internal probing demonstrates that although models represent epistemic states in their hidden activations, this passive knowledge fails to govern downstream strategy selection for generation (§4.2). Motivated by this diagnostic, we train models using a reward function that aligns predicted models' response strategies with human preference choices, conditioned on preference epistemic uncertainty (§4.2). This targeted training narrows the user-acceptability generalization gap by a third to out-of-distribution personas and conversational styles beyond standard fine tuning.

Together, these results identify adaptive personalization under epistemic uncertainty as a distinct, learnable capability: one that contemporary language models do not reliably deploy by default, but one that targeted training can build.

2 Task Formulation and Benchmark

2.1 Task Formulation

Dialogue theories characterize conversation as a process of establishing and maintaining common ground, in which interlocutors continuously distinguish between mutually established information and information that remains uncertain or ungrounded (Clark, 1996; Clark & Brennan, 1991). Formal theories of belief revision describe how such beliefs should be updated when new evidence conflicts with existing knowledge (Alchourrón et al., 1985; Gärdenfors, 1988; Hansson, 1999). We ground our task in these two areas and adapt them to a personalization setting, where an assistant's evidence about a user's preferences can be established, uncertain, or revised, and its conversational strategy should track that evidence.

An assistant's evidence for a preference (epistemic uncertainty) is *established* when explicitly expressed and uncontradicted, *uncertain* when evidence is insufficient or ambiguous, or *revised* when newer evidence supersedes it. Given these varying epistemic states, communication theory, belief re-

vision theory and linguistic studies (Clark, 1996; Clark & Brennan, 1991; Lakoff, 1973; Alchourrón et al., 1985; Gärdenfors, 1988) distinguishes several strategies an assistant can take in conversation:

1. *exploit*: acting directly on a stated preference
2. *elicit*: asking the user for missing information
3. *confirm*: checking whether a preference still holds before acting on it
4. *hedge*: offering options without fully committing to either
5. *recalibrate*: updating behavior once new evidence contradicts a prior preference

Together, these five strategies form an operational set of conversational actions (Table 3; full definitions in § A). Formally, at each turn the assistant observes a dialogue context x with epistemic state $s \in \{\text{ESTABLISHED}, \text{UNCERTAIN}, \text{REVISED}\}$ over a preference and must select an appropriate conversational strategy $a \in \{\text{EXPLOIT}, \text{ELICIT}, \text{CONFIRM}, \text{HEDGE}, \text{RECALIBRATE}\}$.

2.2 Preferred Strategies Depend on Epistemic State

We conducted a multilingual preference study across five languages - English, Spanish, Portuguese, Hindi, and Indonesian, to measure which assistant response strategies humans prefer across different epistemic states. Annotators were presented with dialogues exhibiting varying epistemic states and chose their preferred response between LLM outputs implementing different conversational strategies. The study included 500 dialogue instances evaluated by 3 annotators per item ($\sim$1,500 total annotations). Table 1 reports this resulting preference distribution.

Human preferences are not evenly distributed across the five strategies: in each epistemic state, a small set of strategies account for the vast majority of preferences ($> 80\%$) (Table 1), suggesting a natural separation between acceptable and unacceptable strategies. To validate this separation, we exhaustively evaluated all 15 possible two-group splits of the five strategies in each state using a χ^2 goodness-of-fit test against the observed human choices (§ A.6). In every state, one split fit the observed preferences best. We therefore use these best-fitting splits to define our human-grounded sets of *acceptable* and *unacceptable* strategies: *exploit* for *established*; *elicit, confirm, hedge* for *uncertain*; and *elicit, confirm, hedge, recalibrate* for *revised*.

Table 1: **Preferred strategy varies systematically with preference epistemic state.** ✓ marks the set of acceptable strategies used for benchmark construction, evaluation, and training.

Strategy	Established	Uncertain	Revised
Exploit	90.4%✓	8.5%	6.2%
Elicit	2.4%	27.0%✓	20.0%✓
Confirm	2.0%	28.4%✓	24.6%✓
Hedge	4.0%	26.6%✓	23.0%✓
Recalibrate	1.2%	9.5%	26.2%✓

2.3 Benchmark Construction

Evaluating strategy selection under preference uncertainty requires two properties that existing real-world dialogue corpora lack: controlled, multi-session preference trajectories with known ground-truth transitions, and counterfactual candidate responses spanning multiple strategies at each turn. To address this gap, we introduce a multilingual benchmark spanning five languages (English, Spanish, Portuguese, Hindi, and Indonesian; Figure 7), constructed through a three-role multi-LLM pipeline.

Starting from 250 diverse personas seeded from PersonaHub Ge et al. (2024) using farthest-point sampling, each persona is augmented with preference attributes drawn from 20 life domains Zhao et al. (2025b) and 13 interaction styles Li et al. (2026). A *user simulator* plans multi-session conversation arcs (10-14 sessions) with oracle-labeled preference trajectories, systematically injecting planned drift events that transition a preference between the three epistemic states: *established*, *uncertain*, and *revised*. An *AI assistant* then generates candidate responses across strategies at each turn, while an *LLM jury* filters output for conversation coherence and label fidelity.

Our final benchmark comprises 25,006 turns across 2,939 multi-session conversations. Each turn includes four candidate LLM responses: two acceptable strategies (per Table 1) and two negative foils categorized into seven realistic failure modes identified during preliminary annotation (over-personalization, epistemic overreach, staleness, need mismatch, over-probing, style mismatch, and failure to probe; see § A for full statistics and construction details).

We validate our benchmark through a human evaluation following established benchmark protocols Zollo et al. (2025); Jang et al. (2026); Dey et al. (2025). Human annotators evaluated 500 dialogue instances across all five languages (3 annotators per language) on five dimensions: conversation coherence, epistemic-state clarity, strategy-label correctness, foil-type accuracy, and response acceptability. Agreement between human annotators and the LLM jury was substantial ($\kappa = 0.78$; § A.6), while human multiple-choice accuracy (selecting a response with acceptable strategy use from candidate responses) reached 92%. These results indicate that the benchmark's epistemic states and strategy labels are both recognizable to humans and reflected in natural dialogue.

2.4 Evaluation Protocol

At each turn, given persona P, conversation history C, and current user turn u, the benchmark provides four candidate responses $(r_1^+, r_2^+, r_3^-, r_4^-)$, where r_1^+ and r_2^+ implement acceptable strategies for the current epistemic state, while r_3^- and r_4^- implement foil strategies (§ 2). To benchmark *adaptive personalization under epistemic uncertainty*, we evaluate models across two settings: *multiple-choice selection* (MCQ), where the model selects one candidate from $\{r_1^+, r_2^+, r_3^-, r_4^-\}$, and *free-response generation* (FRG), where the model generates an unguided response given (P, C, u). Both settings are evaluated on **strategy calibration (SC)**: whether the chosen or generated response utilizes an acceptable strategy[1] for the epistemic state (MCQ: matching r_1^+ or r_2^+ with a 50% chance baseline; FRG: classified as acceptable via an LLM jury; § 2.3). Comparing SC across settings explicitly isolates strategy recognition from strategy generation.

FRG outputs are additionally evaluated along two dimensions by the user simulator (§ 2.3): a state-conditioned **rubric (R)** scoring five binary criteria (out of 5): Information, Confidence, Intent, Scope, and Style—and **user acceptability (UA)**, a binary accept/reject judgment. Together, SC, R, and UA decouple strategy selection, response execution, and overall user satisfaction.

3 LLMs Fail to Adapt Conversational Strategy Under Epistemic Uncertainty

§2.2 showed that people prefer varying conversational strategies based on epistemic state. We now use our multi-lingual benchmark to measure whether models adapt strategy during conversation generation. We evaluate eight closed- and open-weight models (GPT-5.4, Opus 4.7, Grok 4.1, Gemini 3 Pro, Gemma 3 12B, Llama-3.1-8B, OLMo-3-7B, and Qwen3-8B) across the three epistemic states and across five languages.

3.1 Strategy Calibration Collapses When Preference Evidence Becomes Uncertain or Revised

We first measure whether models adapt their open-ended generation strategies based on epistemic state. Figure 2 shows free-response (FRG) strategy calibration (SC) by epistemic state, i.e., whether a response's dominant strategy is acceptable for that state. Calibration is high when preferences are established ($\sim 83.1\%$ across models) but declines significantly once conversations require managing uncertainty ($\sim 36.9\%$; best model: GPT-5.4 at 62%) or adapting to a revision ($\sim 47.9\%$; best: GPT-5.4 at 68%). Both declines are significant after accounting for the nested structure of the data (mixed-effects logistic regression with random intercepts for persona and conversation; Uncertain: $\beta = -2.16$, Revised: $\beta = -1.67$, both $p < .001$; full specification and per-language estimates in § B.1). This trend holds across models and languages, and is deeper in Hindi and Indonesian. closed-weight models retain higher calibration than open-weight models, and this margin widens with epistemic difficulty: 4.2 points when preferences are established, 16.9 under uncertainty, and 25.3 under revision.

3.2 Models Recognize Appropriate Strategies More Often Than They Generate Them

[1]When responses combine strategies, we score the dominant strategy, defined as the strategy that most strongly structures the response. This avoids counting strategy-stacking responses as calibrated across states. Results are robust to alternative scoring rules (Appendix B.5).

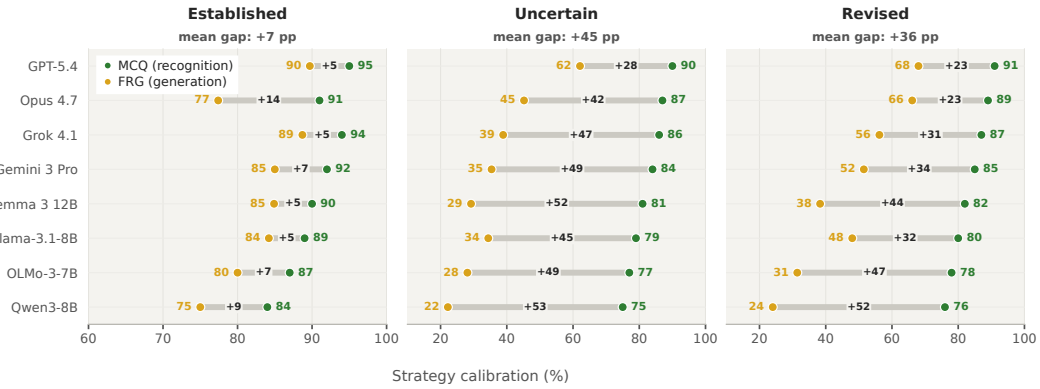


Figure 3: **Recognition exceeds generation.** Models identify acceptable strategies (MCQ, green) more often than they generate them (FRG, yellow) across all three epistemic states. The horizontal distance between paired points is the recognition-generation gap.

This decline in SC could reflect either of two distinct failures: a *generation* failure, in which models correctly *recognize* the appropriate strategy but fail to produce it, or a more fundamental *recognition* failure, in which models cannot correctly identify the appropriate strategy at all. We distinguish these failure modes using our two evaluation methods: multiple-choice questions (MCQ), where a model must select the appropriate response (text following acceptable strategy) among candidates, and free-response generation (FRG), which requires the model to generate text following the acceptable strategy.

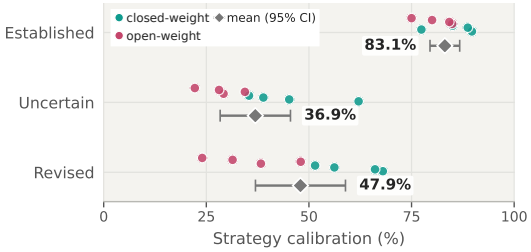


Figure 2: **Strategy calibration across epistemic states.** All models perform best in *Established* and struggle with *Uncertain* and *Revised* states. closed-weight models outperform open-weight models in *Uncertain* and *Revised* states.

Figure 3 compares the two evaluation results. MCQ strategy calibration remains high across all three states (mean 85.4%, range: 75-95%; 50% chance baseline), declining only modestly under uncertainty and revision relative to established preferences (90.3% $\rightarrow$ 82.4% $\rightarrow$ 83.5%). FRG strategy calibration, by contrast, collapses far more sharply across the three states (83.1% $\rightarrow$ 36.9% $\rightarrow$ 47.9%; §3.1). The resulting recognition–generation gap is present for every model in every state, and is widest under UNCERTAIN (+45 pp, $p < .001$), followed by REVISED (+36 pp, $p < .001$) and ESTABLISHED (+7 pp, $p < .001$). Models can therefore identify the appropriate strategy substantially more reliably than they produce it: the failure characterized in §3.1 is not explained by an inability to recognize appropriate behavior. A further check, explicitly supplying the correct strategy and asking models only to execute it, shows this gap is driven specifically by selecting a strategy, not executing one: models generate responses with the supplied strategy on 95% of turns, and execution quality under this condition is state-invariant, whereas unaided calibration still swings by over 40 points across states (§ B.3)

3.3 GENERATION FAILURES REFLECT SYSTEMATIC OVER-EXPLOITATION

In this section, we characterize models' FRG outputs by their dominant strategy (Figure 4). Averaged across models, *exploit* is the dominant strategy in 83%, 59%, and 52% of responses under ESTABLISHED, UNCERTAIN, and REVISED. Models do shift toward acceptable strategies as evidence weakens (e.g., dominant *elicit* rises from 7% to 18% and 21%), but far too little: *exploit* remains the most frequent dominant strategy in 23 of 24 model$\times$state cells, the sole exception being Opus-4.7 under REVISED (*elicit* 36% vs. 34%). Exploitation is also near-universal as a component of responses, present in 90%, 91%, and 86% of responses across states (Figure 11), often alongside a question or hedge. Models are therefore not selecting strategies at random, but defaulting to the same policy, *exploit*, regardless of whether it remains appropriate. This holds under alternative

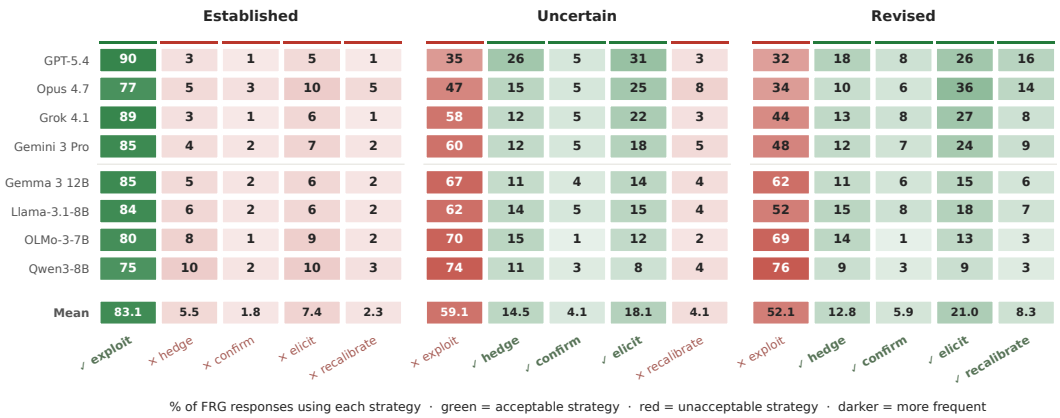


Figure 4: **Models default to exploit across epistemic states.** Each cell shows the percentage of free-response generations for which a strategy is dominant, pooled across five languages. Green indicates acceptable strategies and red indicates unacceptable strategies. *Exploit* is the most frequent dominant strategy in 23 of 24 model×state cells, including 59% and 52% of responses under UNCERTAIN and REVISED, respectively, where it is unacceptable. Multi-label strategy presence is reported in Figure 11.

scoring rules (Table 27): crediting any response that contains an acceptable strategy, SC still falls to 48.1% (UNCERTAIN) and 55.3% (REVISED), and requiring that no unacceptable strategy be present drops it to 4.8% and 14.1%, as nearly every response includes some exploitation.

To characterize the models' generations at a finer grain, human annotators coded free-response failures into the seven foil types introduced in §2. The dominant error type shifts systematically with epistemic state (χ^2 test of independence, $p < .001$): *over-personalization* is most common under *Established* (31.2%), *failure-to-probe* under *Uncertain* (38.8%), and *staleness* under *Revised* (25.1%). Each reflects the same fixed exploit-first policy, expressed differently depending on context: models over-commit when preferences are already settled, fail to seek missing evidence when they are not, and continue acting on evidence that has since been contradicted. This same pattern of *exploit* usage holds across model families and languages, though it is more pronounced in open-weight models (error rate, i.e., $1 - \text{SC}$, is 15.5 pp higher than in closed-weight models; $p < .001$) and in Hindi and Indonesian (4.1 pp higher than in the other three languages; $p < .001$).

Table 2: **FRG performance**, Avg. across epistemic states and languages w/ 95% CI. SC = strategy calibration, R = rubric, UA = user acceptability.

Model	SC (%)	R (1-5)	UA (%)
GPT-5.4	**73**$_{\pm 3}$	**3.82**$_{\pm 0.12}$	**80**$_{\pm 4}$
Opus	63$_{\pm 3}$	2.44$_{\pm 0.14}$	65$_{\pm 4}$
Grok	61$_{\pm 4}$	2.23$_{\pm 0.11}$	60$_{\pm 5}$
Gemini	57$_{\pm 4}$	2.80$_{\pm 0.13}$	55$_{\pm 5}$
Llama	55$_{\pm 4}$	2.64$_{\pm 0.15}$	35$_{\pm 4}$
Gemma	51$_{\pm 4}$	3.09$_{\pm 0.14}$	48$_{\pm 5}$
OLMo	46$_{\pm 4}$	2.17$_{\pm 0.10}$	40$_{\pm 4}$
Qwen	40$_{\pm 5}$	1.75$_{\pm 0.09}$	22$_{\pm 4}$

3.4 STRATEGY CALIBRATION PREDICTS RESPONSE QUALITY ACROSS EPISTEMIC STATES

We next examine whether strategy calibration translates into higher-quality and more acceptable responses. Table 2 reports strategy calibration (SC), rubric-based execution quality (R), and user acceptability (UA) for each model, pooled across all three epistemic states. SC correlates positively with both R and UA within each state: ESTABLISHED ($r = 0.71$, $r = 0.69$), REVISED ($r = 0.67$, $r = 0.74$), and UNCERTAIN ($r = 0.57$, $r = 0.61$). These correlations are slightly weaker in UNCERTAIN, however, they remain positive. This weaker association may reflect differences in user needs: responses that are not maximally appropriate for the epistemic state may nevertheless be acceptable to users seeking a strictly direct answer. Taken together, the results suggest that strategy calibration is generally associated with stronger execution and greater user satisfaction across various personalized contexts.

4 CHARACTERIZING THE FAILURE

§3 showed that the collapse in strategy calibration (SC) reflects a generation failure, where models fail to *select* the appropriate strategy during generation. This failure in models also meaningfully degrades response quality i.e. declines in rubric (R) and user acceptability (UA) scores.

4.1 PROMPTING FAILS TO INDUCE STATE-CONDITIONAL ADAPTATION

We evaluate five lightweight prompting interventions across eight models and five languages: (1) *explicit strategy instructions* detailing strategy definitions and use cases; (2) *oracle preference memory*, providing structured user preference histories; (3) *chain-of-thought (CoT)* Wei et al. (2022), step-by-step reasoning about the current epistemic state before generating a response; (4) *few-shot examples* demonstrating strategy usage in utterances across epistemic states; and (5) *best-of-N sampling*, sampling model (N=5) generations and selecting best (highest SC, R, UA score) [2]. Prompt templates and per-metric results are detailed in Appendix B.7

Figure 5 shows no intervention method consistently improves SC across all epistemic states. In many interventions (*Explicit instructions, CoT, Few-shot*), gains under UNCERTAIN and REVISED trade off with losses under ESTABLISHED. For instance, *few-shot examples* boost SC by 4.7% under REVISED preferences but degrade it by 3.4% under ESTABLISHED. Moreover, we find that rather than selective, context-dependent adaptation, prompting often shifts a global caution parameter, forcing models to use the strategy *hedge* indiscriminately. These context-rich interventions also do not show improvements in execution-based metrics. Even *few-shot examples*, the strongest SC intervention, yield only modest gains in R (+0.12) and UA (+1.31 pp) under *Uncertain* and *Revised*, while reducing both under *Established*. The intervention shifts the selected strategy during generation more than the resulting generated response quality. Meanwhile, *oracle preference memory* and *best-of-five sampling* yield almost no effect on any metric. This pattern holds across models and languages, with slightly larger effects in closed-weight models and English/Romance languages than in Hindi and Indonesian (§ B.7).

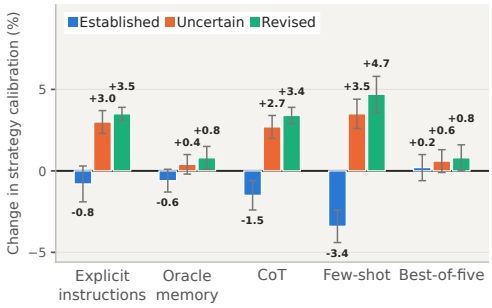


Figure 5: **Inference-time interventions fail to induce state-conditional adaptation.** Bars depict mean SC change relative to baseline across eight models and five languages; error bars denote 95% bootstrap CIs. Gains under UNCERTAIN and REVISED lead to performance losses under ESTABLISHED, indicating a global shift toward caution rather than selective adaptation based on epistemic state.

4.2 UNDERSTANDING AND REPAIRING THE SELECTION FAILURE

As inference-time interventions do not provide consistent improvements, we further analyze models' representations of two concepts: *epistemic state* and *strategy selection* to better understand how to build more calibrated models.

We utilize linear and nonlinear probes and causal ablation on four open-weight models (Llama-3.1-8B, Gemma-3-12B, Qwen3-8B, OLMo-3-7B). Averaged across models, both epistemic state and the upcoming strategy to be used in generation are linearly decodable[3] well above chance (80.5%; chance 33.3%) and (70.0%; chance 20.0%). Yet we find limited evidence the two are connected: (1) cosine similarity between state and strategy probe weight vectors is ≈ 0.03, close to orthogonal; (2) probe accuracy on the two concepts correlates weakly ($\phi \approx 0.015$); (3) and nonlinear MLP probes show subspace overlap of only ≈ 0.021 above a random floor. Moreover, neither ablating the decoded epistemic state direction nor amplifying it produces any significant shift in strategy calibration in FRG: -0.3% vs $+1.2\%$. This pattern holds across multiple layers (20-24) where

[2]These interventions cover distinct ways of augmenting in-context information at inference time: explicit rules, ground-truth labels, reasoning, demonstrations, and sampling budget. While not an exhaustive sweep of all prompt designs, we test a broad set of contextualization strategies.

[3]Concepts are decodable from Layer 20.

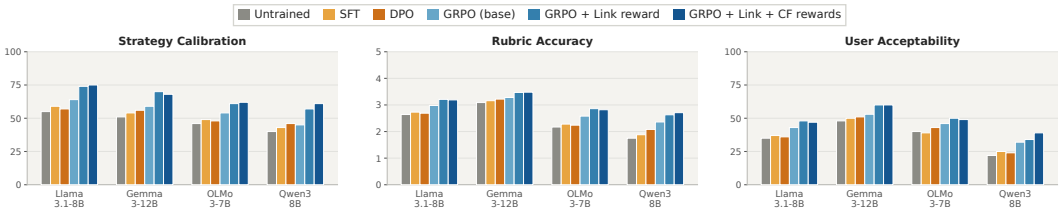


Figure 6: **Targeting the state-strategy link improves performance beyond generic fine-tuning, across every model.** SC, R, and UA for each model across the full training progression: untrained, generic fine-tuning (SFT, DPO, GRPO), and GRPO with the targeted reward (+LINK, +LINK+CF). Generic fine-tuning yields modest gains; adding the Link and Link+CF rewards yield the largest improvement for every model.

epistemic state first becomes decodable and across multiple languages. See §B.8 for full experiment details. Although these results do not rule out a causal pathway for epistemic state elsewhere in the model, the combination of the behavioral gap between recognition and unaided selection (§3.2) and the failure of inference-time interventions to improve generation (§4.1) suggests: epistemic state is accessible to the model, but not reliably connected to strategy selection during generation.

Our analysis suggests that models encode epistemic state but do not use it to select strategies. A standard reward that scores only the final response can improve behavior without fixing this: the model may learn surface cues that correlate with the right strategy in training data but fail to transfer. We therefore supervise the link itself, encouraging the strategy the model internally represents to match the distribution humans prefer for the current epistemic state (Table 1). We target a distribution rather than a single strategy, since humans accept several strategies per state. We hypothesize that supervising the strategy the model internally represents, rather than only its final output, reduces reliance on surface cues and improves generalization to unseen personas.

Formally, our base reward combines strategy calibration (SC), user acceptability (UA), and rubric quality (R, normalized to 0 to 1) with equal weight: $r_{\text{base}} = \frac{1}{3}(\text{SC} + \text{UA} + \text{R})$. To directly supervise the alignment between internal states and strategy selection, we introduce two auxiliary rewards: *Link* and *Link+CF*.

Let h be the model's hidden representation of its response (residual-stream activation at 70% model depth, mean-pooled over the first eight generated tokens, by which point the model has committed to a strategy), and let $P_\phi(\cdot \mid h)$ be the strategy distribution decoded from h by our trained strategy probe. Because h is computed from the response, it varies across rollouts. The LINK reward supervises agreement between this distribution and the human preference distribution $\text{Pref}(s)$ (Table 1) under the ground-truth epistemic state s: $r_{\text{LINK}} = -\text{KL}\big(P_\phi(\cdot \mid h) \,\|\, \text{Pref}(s)\big)$. The probe is refit every 10 GRPO steps on the current policy's rollouts, labeled by the strategy each rollout actually used (§B.8). While r_{LINK} aligns strategy selection under the active state s, it does not enforce counterfactual consistency across states. LINK+CF therefore adds a counterfactual term: for a randomly sampled alternate state $s' \neq s$, we patch h along the probed state direction to obtain h' and reward agreement with human preferences under s': $r_{\text{LINK+CF}} = r_{\text{LINK}} - \text{KL}\big(P_\phi(\cdot \mid h') \,\|\, \text{Pref}(s')\big)$.[4]

The full training reward is $r_{\text{base}} + \lambda\, r_{\text{LINK}}$ or $r_{\text{base}} + \lambda\, r_{\text{LINK+CF}}$, with $\lambda = 0.1$, computed once the full response is generated. We compare both against SFT, DPO, and GRPO trained on the base reward alone. Training data (∼5K turns per language, 24K total) is generated with our benchmark pipeline (§2.3) using open-weight models (Kimi-K2.6, Qwen3-235B, and DeepSeek-V3).[5] We evaluate on a held-out in-distribution set and on our benchmark.

Fig 6 reports strategy calibration (SC), rubric quality (R), and user acceptability (UA) on our benchmark for every model across the full training progression, averaged over languages. Generic fine-tuning yields modest gains: GRPO, the strongest generic baseline, raises SC from 48.0% to 55.5% on average, and outperforms SFT and DPO for every model on R and UA and for all mod-

[4] This direction was causally inert under passive intervention (see above); r_{LINK} and $r_{\text{LINK+CF}}$ test whether reward-based optimization can establish a causal pathway from it to strategy selection.

[5] Additional training details in §B.8.

els but Qwen3-8B on SC (SC and UA: McNemar's test, $p < .001$; R: Wilcoxon signed-rank on conversation-level mean differences, $p < .001$). Adding LINK yields the largest single improvement, raising SC, R, and UA for every model: averaged across models, SC rises from 55.5% to 65.5% (+10.0 pp), R from 2.80 to 3.04, and UA from 43.5% to 48.0%. LINK+CF performs comparably to LINK on average (SC 66.5%, R 3.05, UA 48.8%), with the largest additional gains for Qwen3-8B (+4 pp SC, +5 pp UA).

We also observe large *generalization gaps*: the drop in each metric between the held-out in-distribution set and the benchmark's unseen personas, Gap $=$ $\text{Metric}_{\text{in-dist}} - \text{Metric}_{\text{benchmark}}$ (in-distribution results in Table 33). LINK narrows this gap substantially: from 11.5 to 6.5 pp on SC, from 0.61 to 0.48 on R, and from 7.5 to 5.0 pp on UA (McNemar's test, $p < .001$), with LINK+CF performing comparably (6.5 pp, 0.50, 5.2 pp). In Hindi, for example, the SC gap narrows from 18 pp under GRPO to 11 pp under LINK.

Together, these results indicate that adaptive personalization is learnable: supervising the model's internal strategy representation toward state-conditional human preferences improves both in-distribution behavior and generalization to personas never seen during training.

5 RELATED WORK

Personalization benchmarks. Existing benchmarks show that models struggle with preference retrieval and long-context recall (Salemi et al., 2024; Zhao et al., 2025a; Jiang et al., 2025b; 2024; Wu et al., 2025; Maharana et al., 2024), often treating implicit-versus-explicit expression as the main difficulty (Jiang et al., 2025b; Guo et al., 2026); HorizonBench (Li et al., 2026) shows dynamic preferences are harder to track. None label epistemic state independently of preference content, so they cannot separate retrieval failures from failures to adapt *strategy* as a preference's reliability changes (§3.1).

Clarification and preference elicitation. Prior work studies when to ask clarifying questions (Rao & Daumé III, 2018; Aliannejadi et al., 2019; Zamani et al., 2020; Kuhn et al., 2022; Deng et al., 2023; Zhang et al., 2024; Zhang & Choi, 2025) and trains models to ask or elicit preferences (Li et al., 2025; Andukuri et al., 2024; Chen et al., 2025; Zhang et al., 2025); Chen et al. (2025) note that LLMs often guess user intent rather than ask, echoing our exploit-first finding. These works address ambiguity within a single request; we condition strategy on the reliability of a *persisting* preference, which also requires recognizing stale preferences and recalibrating rather than re-asking.

Strategy, calibration, and representations. Strategy has been treated as a learnable construct in emotional support (Liu et al., 2021), persuasion (Wang et al., 2019), and negotiation (He et al., 2018), though never conditioned on evidence reliability; we also draw on exploit–explore framing in conversational recommendation (Christakopoulou et al., 2016; Radlinski & Craswell, 2017) and dialogue grounding (Clark, 1996; Traum, 1994). We measure behavioral rather than probabilistic calibration (Kadavath et al., 2022), and our recognition–generation gap echoes findings that models evaluate better than they generate (West et al., 2024). That decodable information can lack causal leverage follows probing work (Ravichander et al., 2021; Elazar et al., 2021; Hewitt & Liang, 2019). Closest to our reward, Wu & Tang (2026) derive GRPO advantages from representation directions; we instead supervise the internal strategy distribution against human-preference targets, with a counterfactual patching term.

6 CONCLUSION

We introduced adaptive personalization under epistemic uncertainty: conditioning conversational strategy on how reliable a preference currently is. All eight models evaluated exhibit the same failure pattern: strategy calibration drops by over 35 percentage points once preferences become uncertain or revised. This failure is one of generation rather than recognition: models identify appropriate strategies far more reliably than they produce them. Targeted training that supervises internal strategy representations improves strategy calibration, demonstrating that this capability is both distinct and learnable. Ultimately, true personalization in language models demands more than user context retrieval, it requires learning to adapt conversational behavior dynamically as human preferences naturally evolve.

USE OF GENERATIVE AI

In this work, we used generative AI tools for the following tasks that require disclosure. **(1) Synthetic data generation:** the benchmark's personas, session plans, user turns, and candidate responses were generated by LLMs (Claude Sonnet 4.6 and Gemini 3.1 Pro), and the training corpus was generated with the same pipeline using open-weight models (Kimi-K2.6, Qwen3-235B, DeepSeek-V3), as described in §2.3 and Appendix B.8. **(2) Data annotation and evaluation:** an LLM jury (Claude Opus 4.6, Gemini 3 Flash, GPT-5.5) labeled user turns for epistemic-state construction, and LLM judges scored strategy calibration, rubric quality, and user acceptability of model responses; these labels were validated against human annotators (Appendix A.6). **(3) Feedback on methodology and interpretation:** we used an LLM assistant to review the manuscript, which provided feedback on the experimental design (e.g., the definition of strategy calibration, the construction of the training reward, and the choice of evaluation sets) and on the interpretation and internal consistency of results. **(4) Implementation:** LLM coding assistants (Claude and Codex) were used to help implement figure/plot generation and parts of the training code. We did not use generative AI tools to formulate mathematical claims or proofs, and translation is not applicable: dialogues in each language were generated natively rather than translated, and human validation was performed by native or professionally fluent annotators.

We have reviewed all AI-assisted work. All reported results come from our experimental outputs; all numbers were checked by the authors against these outputs. Synthetic data quality was verified through a human validation study with 15 annotators across five languages (Appendix A.6). LLM-assisted code was reviewed and tested by the authors.

REFERENCES

Carlos E Alchourrón, Peter Gärdenfors, and David Makinson. On the logic of theory change: Partial meet contraction and revision functions. *The journal of symbolic logic*, 50(2):510–530, 1985.

Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W Bruce Croft. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pp. 475–484, 2019.

Chinmaya Andukuri, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. STaR-GATE: Teaching language models to ask clarifying questions. In *Conference on Language Modeling*, 2024.

Maximillian Chen, Ruoxi Sun, Tomas Pfister, and Sercan Ö. Arik. Learning to clarify: Multi-turn conversations with action-based contrastive self-training. In *International Conference on Learning Representations*, 2025.

Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. Towards conversational recommender systems. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 815–824, 2016.

Herbert H. Clark. *Using Language*. Cambridge University Press, 1996.

Herbert H. Clark and Susan E. Brennan. Grounding in communication. In Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley (eds.), *Perspectives on Socially Shared Cognition*, pp. 127–149. American Psychological Association, 1991.

Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.

Priyanka Dey, Aayush Bothra, Yugal Khanter, Jieyu Zhao, and Emilio Ferrara. Can llms express personality across cultures? introducing culturalpersonas for evaluating trait alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 20241–20262. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.1101.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. In *Transactions of the Association for Computational Linguistics*, volume 9, pp. 160–175, 2021.

Peter Gärdenfors. *Knowledge in Flux: Modeling the Dynamics of Epistemic States*. MIT Press, 1988.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*, 2024.

Qianyun Guo, Yibo Li, Yue Liu, and Bryan Hooi. Towards realistic personalization: Evaluating long-horizon preference following in personalized user-llm interactions. *arXiv preprint arXiv:2603.04191*, 2026.

Sven Ove Hansson. *A Textbook of Belief Dynamics*. Kluwer Academic Publishers, 1999.

Tamunotonye Harry, Ivoline C Ngong, Chima Nweke, Yuanyuan Feng, and Joseph Near. Beyond fixed psychological personas: State beats trait, but language models are state-blind. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 26440–26468, 2026.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2333–2343, 2018.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Han Jang, Junhyeok Lee, and Kyu Sung Choi. Scizoom: A large-scale benchmark for hierarchical scientific summarization across the llm era. *arXiv preprint arXiv:2603.16131*, 2026.

Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo J Taylor, and Dan Roth. Know me, respond to me: Benchmarking llms for dynamic user profiling and personalized responses at scale. *arXiv preprint arXiv:2504.14225*, 2025a.

Bowen Jiang, Yuan Yuan, Maohao Shen, Zhuoqun Hao, Zhangchen Xu, Zichen Chen, Ziyi Liu, Anvesh Rao Vijjini, Jiashu He, Hanchao Yu, et al. Personamem-v2: Towards personalized intelligence via learning implicit user personas and agentic memory. *arXiv preprint arXiv:2512.06688*, 2025b.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. Personallm: Investigating the ability of large language models to express personality traits. In *Findings of the association for computational linguistics: NAACL 2024*, pp. 3605–3627, 2024.

Saurav Kadavath, Tom Conerly, Amanda Askell, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. CLAM: Selective clarification for ambiguous questions with generative language models. *arXiv preprint arXiv:2212.07769*, 2022.

George Lakoff. Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of philosophical logic*, 2(4):458–508, 1973.

Belinda Z. Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. Eliciting human preferences with language models. In *International Conference on Learning Representations*, 2025.

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

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. Towards emotional support dialog systems. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)*, pp. 3469–3483, 2021.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of llm agents. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024. arXiv:2402.17753.

Filip Radlinski and Nick Craswell. A theoretical framework for conversational search. *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, pp. 117–126, 2017.

Sudha Rao and Hal Daumé III. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, 2021.

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.

David R. Traum. A computational theory of grounding in natural language conversation. In *Proceedings of the AAAI Fall Symposium on Planning for Interagent Communication*, 1994.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 5635–5649, 2019.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Peter West, Ximing Lu, Nouha Dziri, et al. The generative ai paradox: What it can create, it may not understand. *International Conference on Learning Representations (ICLR)*, 2024.

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 (ICLR)*, 2025. arXiv:2410.10813.

Rui Wu and Ruixiang Tang. When reward hacking rebounds: Understanding and mitigating it with representation-level signals. *arXiv preprint arXiv:2604.01476*, 2026.

Jianfei Xiao, Xiang Yu, Chengbing Wang, Wuqiang Zheng, Xinyu Lin, Kaining Liu, Hongxun Ding, Yang Zhang, Wenjie Wang, Fuli Feng, et al. Alpsbench: An llm personalization benchmark for real-dialogue memorization and preference alignment. *arXiv preprint arXiv:2603.26680*, 2026.

Hamed Zamani, Susan Dumais, Nick Craswell, Paul Bennett, and Gord Lueck. Generating clarifying questions for information retrieval. In *Proceedings of The Web Conference 2020*, pp. 418–428, 2020.

Michael Zhang and Eunsol Choi. Clarify when necessary: Resolving ambiguity through interaction with LMs. In *Findings of the Association for Computational Linguistics: NAACL 2025*, 2025.

Michael Zhang, W. Bradley Knox, and Eunsol Choi. Modeling future conversation turns to teach LLMs to ask clarifying questions. In *International Conference on Learning Representations*, 2025.

Tong Zhang, Peixin Qin, Yang Deng, et al. CLAMBER: A benchmark of identifying and clarifying ambiguous information needs in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

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*, 2025a.

Zheng Zhao, Clara Vania, Subhradeep Kayal, Naila Khan, Shay B. Cohen, and Emine Yilmaz. Personalens: A benchmark for personalization evaluation in conversational ai assistants. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18023–18055, Vienna, Austria, July 2025b. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.927.

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

A Benchmark Construction

This appendix provides full construction details for the benchmark introduced in §2, including persona generation, conversation and response construction, dataset statistics, human validation, and comparison to prior personalization resources. See Figure 7 for an overall understanding of our benchmark construction pipeline.

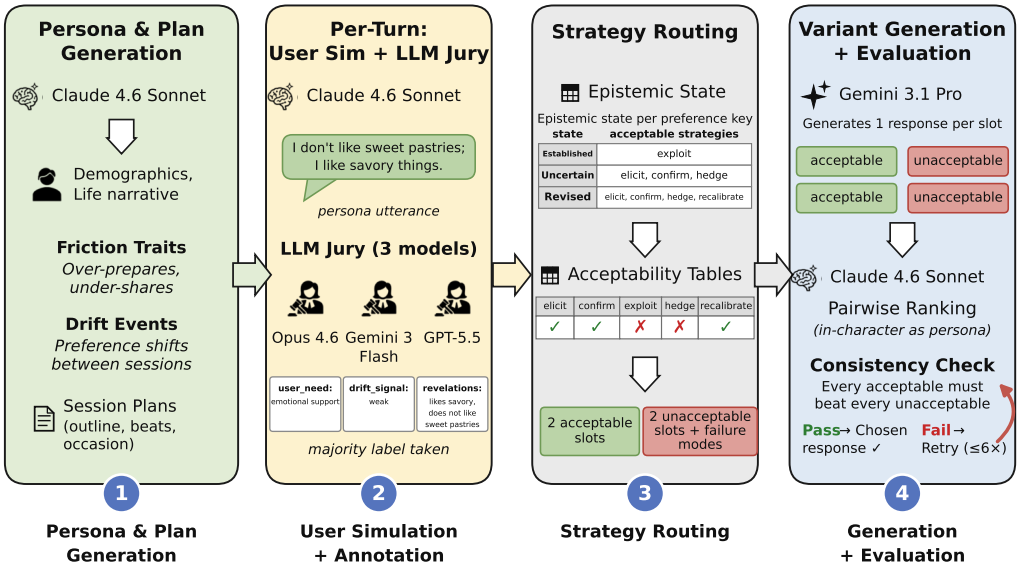


Figure 7: **Benchmark construction pipeline.** (1) A persona and multi-session life-arc plan are generated from a PersonaHub Ge et al. (2024) seed, including demographics, friction traits, and scheduled drift events. (2) At each turn, a user simulator produces a persona-conditioned utterance; an LLM jury (three models, majority vote) annotates it for user need, drift signal strength, and per-key preference revelation. (3) Per-turn labels are combined with the deterministic drift schedule into an epistemic state (established, uncertain, or revised), which determines the acceptable and non-acceptable strategies for that turn via the human-validated acceptability sets (Table 1). (4) Candidate responses are generated for two acceptable (acceptable strategy, LLM jury-verified) and two foil-type slots and ranked pairwise, in-character as the persona; a turn is retained only if every acceptable response outranks every foil, with up to six regeneration attempts. Full construction details are provided in Appendix A.

Strategy	Valid state	Example response
Exploit	ESTABLISHED	"There's a good bakery two blocks down — their mushroom and gruyère galette is the standout."
Hedge	UNCERTAIN	"Two options nearby: a savory galette place, or a deli with hot sandwiches. Either work?"
Elicit	UNCERTAIN	"Before I suggest somewhere — are you after something hot and substantial, or lighter?"
Confirm	UNCERTAIN	"You've mentioned preferring savory — is that still what you're in the mood for today?"
Recalibrate	REVISED	"You'd mentioned avoiding sweet things before, but it sounds like that's shifted — let me start fresh."

Table 3: Strategies and examples of well-calibrated usage, using a persona who has stated a preference for savory over sweet food. Each example is acceptable in the listed state and would constitute a named error in at least one other.

A.1 STRATEGY TAXONOMY

Our framework defines five response strategies, drawn from the conversational-recommender literature and extended for long-horizon dialogue. *Elicit* asks a question to learn a preference the assistant does not have any information on; it is the right move when the relevant preference has never been stated. *Confirm* asks an open-ended question to gather preference information that is currently ambiguous or uncertain based on current context. *Exploit* acts directly on a known preference, and is what delivers value once a belief is both held and current. *Hedge* presents multiple options without committing, deferring the choice to the user. *Recalibrate* treats contradictory evidence as new information; explicitly updates and ceases relying on the superseded preference, rather than verifying whether it still holds. Crucially, the mapping from epistemic state to strategy is one-to-many rather than one-to-one: acceptability is a coarse filter, not a single right answer. When a preference is ESTABLISHED, only *exploit* is calibration-valid—asking about a settled preference is `over_probing`, and recalibrating a belief that never moved is `epistemic_overreach`. When the user has signalled uncertainty, *exploit* becomes `epistemic_overreach` and the valid set shifts to *elicit*, *confirm*, and *hedge*. When a preference has been REVISED, acting on the stale belief is a `staleness_error`, and *recalibrate* joins the valid set. Finally, for revised preference states, the valid set of strategies is: *elicit*, *confirm*, *hedge*, and *recalibrate*. Within each of these regions multiple strategies remain defensible, and which one is *best* turns on user-specific factors the framework cannot enumerate—which is precisely what the persona-grounded simulator's ranking is there to capture.

A.2 PERSONA AND DATA CONSTRUCTION

Pipeline overview. The benchmark is produced by a three-role LLM pipeline. Each role is given a strictly partitioned view of the generative state, which is what makes the oracle labels trustworthy: no role ever sees the answer key for the judgment it is asked to make.

(1) Persona and plan generator. Given a PersonaHub seed, this role produces demographics, a short profile, a long life narrative, two friction traits, and a structured preference profile (§A.2). It then produces a *session plan*: a life arc of 10–14 sessions, each with an occasion, an outline of conversational beats, a primary user need, and a set of preferences the user is expected to reveal in that session. Drift events are authored at this stage and attached to the arc.

(2) User simulator. This role plays the persona turn by turn. It sees the persona, the friction traits, the current session's occasion and beats, summaries of all prior sessions, and a per-turn revelation directive drawn from the plan. It does *not* see the candidate assistant responses or the acceptability labels. It is instructed to express friction behaviorally rather than narrate it, to keep turns to 15–40 words by default (150 words hard maximum), and to deliver drift signals as concrete behavioral detail rather than meta-commentary.

(3) LLM jury. Three judges from three labs independently annotate each user turn for user need, personalization relevance, drift signal strength, and per-key preference revelation. Labels are resolved by majority vote; a preference key is admitted only if at least two of the three judges name it, which

suppresses single-judge over-labelling. The jury sees the transcript and the persona's preference *keys*, but never the persona's current ground-truth preference *values* — the correctness half of the epistemic state is computed deterministically from the plan, not judged.

The jury's per-turn labels are combined with the deterministic drift schedule into a *joint state*, and a rule-based acceptability table maps that joint state onto the acceptable and non-acceptable strategy sets for the turn. Oracle labels are therefore constructed, not inferred: the epistemic state is based on *certainty*, where certainty comes from the jury's revelation level (`hinted` / `revised`.

Abridged prompt templates for each role are given in Figures 8–10.

User simulator (abridged)

```
You are roleplaying as a fictional user named {name} in a conversation
with an AI assistant. Stay fully in character.

WHO YOU ARE: {persona_summary}

YOUR FRICTION TRAITS (these MUST surface naturally -- through BEHAVIOR,
not articulation):
  1. {friction_1}   2. {friction_2}

CRITICAL -- RESISTANCE BEHAVIORS:
 (a) When the assistant names your emotional state, deflect ~40% of the time.
 (b) DO NOT produce a clean self-diagnosis. SHOW friction through behavior.
 (c) If apologetic/guilty: UNDER-share, hint, redirect, minimize.

LENGTH VARIANCE:
 - Default 1-2 sentences (15-40 words). Never exceed 150 words.

DRIFT SIGNAL DELIVERY:
 - Concrete behavioral details, no meta-narration.
 - Wrong: "I've been reading in the mornings, which is unusual for me."
 - Right: "I read most of it before six this morning."
 - Don't connect signals to causes unless drift visibility is "explicit".

NATURALNESS: Do NOT narrate or analyze your own psychology. Say the
concrete thing; leave the cause and the meaning implicit.

SESSION OCCASION: {occasion}    BEATS: {outline_beats}
PRIOR SESSIONS: {prior_summaries}
```

Figure 8: User simulator prompt (abridged).

Persona seeding and diversity. Personas are seeded from PersonaHub and selected via farthest-point sampling to maximize diversity. We start from a pool of 20,000 PersonaHub occupation-and-role seeds. Each seed is represented as a TF–IDF vector over word unigrams and bigrams (English stopwords removed, minimum document frequency 2), and seeds are selected greedily: having selected a set S, the next seed is the pool element maximizing its cosine distance to its nearest neighbour in S, $\arg\max_{x \notin S} \min_{s \in S} \big(1 - \cos(x, s)\big)$, initialized from a random seed. Sampling is run once per language and draws 50 seeds per language, giving 250 personas in total. Seeds are filtered before sampling to remove degenerate entries: seeds shorter than five tokens, seeds naming a specific public figure, and seeds describing a role that cannot plausibly hold consumer preferences across the life domains in Table 4.

Personas are generated independently per language rather than translated, with country sampled from a language-appropriate distribution (e.g., for Spanish: Mexico, Spain, Colombia, Argentina, Peru, Chile, Venezuela) and with names, institutions, media, food and sports references localized to that country. Ethnicity and background are varied *within* country so that language does not collapse onto a single cultural profile.

Candidate response generator (abridged)

```
You are generating a single AI assistant response. Generate a NATURAL,
well-formed response that demonstrates the target strategy faithfully.

CONVERSATION SO FAR: {conversation_so_far}
USER'S LATEST MESSAGE: {user_utterance}

LABELED CONTEXT (for your understanding only;
do not mention these labels):
 - User need: {user_need}
 - Personalization relevance: {personalization_relevance}
 - Drift signal strength in this turn: {drift_signal_strength}
 - Epistemic state per relevant preference: {epistemic_state_summary}
 - Asserted persona style preferences: {persona_style_summary}

TARGET STRATEGY: {strategy}
STRATEGY DEFINITION: {strategy_definition}

REQUIREMENTS:
 - The response must clearly use the target strategy.
 - Honor the asserted persona style preferences by DEMONSTRATING them --
   do NOT explicitly comment on the user's style or register.
 - Do not mention strategies, frameworks, or labels.
```

Figure 9: Candidate response generation prompt (abridged). The foil variant is identical except that the flawed strategy is required to be the *primary shape* of the response — visible in the first sentence and not removable as a wrapper — so that foils differ in strategy rather than in fluency.

Jury judge (abridged)

```
You are an annotator labeling a user's turn in a multi-session AI
assistant conversation. Be precise and grounded in the text.

PERSONA: {persona_summary}
PERSONA FRICTIONS: 1. {friction_1}  2. {friction_2}
PERSONA'S RELEVANT PREFERENCE KEYS (only label revelations on these):
  {available_preference_keys}
CONVERSATION SO FAR (most recent turns): {recent_conversation}
PRIOR SESSION SUMMARIES: {prior_summaries}
USER'S LATEST UTTERANCE: {user_utterance}

ANNOTATE THE USER'S LATEST UTTERANCE:
  user_need                in {productivity, learning, social_support, autonomy}
  personalization_relevance in {high, medium, low, none}
  drift_signal_strength    in {none, weak, strong}
  preference_revelation    per key in {hinted, asserted, revised}
  preference_drift_expressed  per key, boolean
```

Figure 10: Jury judge prompt (abridged). Note that the persona's current preference *values* are withheld: judges are given evidence, never the answer key.

Preference and interaction-style attributes. Each persona is augmented with structured preference profiles spanning 20 life domains (Zhao et al., 2025b) and interaction-style dimensions spanning 13 categories (Li et al., 2026). Full lists are provided in Table 4 and Table 5. A persona is not assigned every domain: we sample 6–12 of the 20 life domains and 4–8 of the 13 interaction-style categories per persona, and then populate every attribute within each selected category. This keeps profiles realistically partial (a median persona holds roughly 45% of the full attribute space) while

guaranteeing enough coverage for multi-session preference tracking. Sampling is constrained so that every persona retains at least one *gating* interaction-style dimension (interactivity level, challenge openness, or validation need) with a value that can be violated, which is what makes style-mismatch foils constructible for every persona.

Table 4: Life domains used for persona preference attributes.

#	Life domain	#	Life domain
1	Alarm	11	Messaging
2	Books	12	Movies
3	Buses	13	Music
4	Calendar	14	Rental cars
5	Events	15	Restaurants
6	Finance	16	Services
7	Flights	17	Shopping
8	Games	18	Sports
9	Hotels	19	Train
10	Media	20	Travel

Table 5: Interaction-style categories used for persona attributes. Each category expands into 5–6 dimensions; e.g. *advice delivery* covers depth preference, abstraction level, interactivity level, question frequency, checklist inclusion, and prioritization.

#	Interaction style
1	Advice delivery
2	Apology style
3	Argumentation and evidence
4	Conflict resolution
5	Creative collaboration
6	Cultural sensitivity
7	Emotional support style
8	Ethics and philosophy
9	Facilitation style
10	Language and register
11	Motivation and coaching
12	Response format
13	Tone and warmth

Drift-event mechanism. A drift event is authored at plan time, not at conversation time, which is what allows the epistemic state to be an oracle rather than an inference. Each persona is assigned a number of drift events sampled so that roughly 10% of personas receive none (a deliberate no-drift control condition). Each event specifies: the preference key(s) affected, a `preference_delta` recording the explicit {`old, new`} values, the session index after which the change takes effect, a *drift visibility* (none, latent, visible) governing how overtly the user signals it, and a *drift magnitude* — subtle (the old value remains valid, e.g. an addition), moderate, or significant (the old value is now contradicted) — sampled at 0.2/0.3/0.5. Events are grounded in the persona's life narrative and preferentially add constraints or remove options (e.g. *no dietary restrictions* $\rightarrow$ *low sodium*) so that a large shift remains natural.

Two constraints on plan generation make drift testable rather than incidental. First, drift is assigned to a domain the persona's life arc actually exercises. Second, the arc is revised so that at least three post-drift sessions have occasions in the drifted domain; without this, a drifted preference is simply never the topic again and no assistant behavior can depend on it.

Epistemic-state labels. Epistemic state is defined by the evidence available in the transcript, not by the persona's underlying preference value. The drift schedule determines *when* and *for which key* the user simulator is directed to signal a change; the LLM jury determines *whether* that signal is actually present in the text. For each salient preference key k, the jury labels each user utterance with a revelation level in $\{\textit{hinted}, \textit{asserted}, \textit{revised}\}$ (Figure 10), resolved by majority vote. The state of

k at a turn is determined by the most recent evidence on k: ESTABLISHED if k has been asserted and no later utterance has hinted at or signalled a change; UNCERTAIN if the most recent evidence on k is a hint, either because the preference was never clearly stated or because the user has ambiguously signalled that it may be shifting; and REVISED if the user has signalled that a previously asserted preference no longer holds. A state persists across turns until new evidence on k appears, and a revised key returns to ESTABLISHED once the new value is asserted and confirmed in a subsequent turn.

A REVISED label requires agreement between plan and jury: a drift event on k must be scheduled at or before the current session, and the jury must independently label the utterance *revised* for k. On disagreement, the turn is downgraded to UNCERTAIN and excluded from scoring. As a result, a scheduled drift that the user never expresses (e.g., visibility *none* or *latent* and undetected by the jury) leaves k ESTABLISHED: the assistant has no evidence of change, and acting on the prior preference is the correct behavior. Because the jury never sees the persona's ground-truth preference values, every state label reflects only what is observable in the conversation.

Example persona. Table 6 shows a complete example persona, including demographic attributes, life narrative, friction traits, and a sample of assigned preference and interaction-style attributes.

Table 6: Example persona (`persona_0001`, English).

Field	Value
Seed	"An alumni relations coordinator who maintains graduate records and organises reunion events"
Demographics	Beverley Thistlethwaite; 54; female; White British; United Kingdom; Coventry, West Midlands; bachelors; Alumni Relations Coordinator, University of Coventry; lower-middle income; partnered, lives with long-term partner and two rescue greyhounds; moderate tech savviness
Life narrative	"Spend ten minutes with Beverley and you will learn the names of both greyhounds, the weather forecast for the weekend, and her opinion on the new self-service coffee machine in the staff room — and this before she gets to the actual point of the conversation. She describes herself, with genuine conviction, as someone who cannot stand idle chatter and who prefers to get straight to business, a self-image that bears absolutely no resemblance to how she operates in practice. Colleagues have quietly learned to budget an extra quarter-hour for any meeting she calls…"
Friction traits	(i) stubbornly attached to a wrong belief, resistant to correction even with evidence; (ii) claims to hate small talk but produces a lot of it
Preference attributes (sample)	*hotels*: star rating `3_star`, amenities {`wifi`, `breakfast_included`, `parking`, `pet_friendly`}, location `near_attractions`, room `double`, chains {`Premier Inn`, `Travelodge`, `boutique independents`}; *events*: types {`theater`, `food_and_drink`, `art_exhibits`, `conferences`}, price `$0--$50`, group size `couple`, seating `reserved`, days `weekends`. Assigned domains: train, games, finance, shopping, events, sports, alarm, hotels
Interaction style	*advice delivery*: interactivity level `low` (gating dimension); *creative collaboration*: role `editor`, rhythm `iterative_feedback_loops`, feedback depth `moderate`, verbal over visual; plus *response format*, *argumentation and evidence*, *facilitation style*

A.3 CONVERSATION GENERATION

Session plans. Each conversation follows a session plan specifying an outline, conversational beats, and occasion for each of 10–14 sessions. The plan is generated in one pass from the persona, as a life arc rather than a list of independent episodes: sessions are ordered so that the sequence reads as time passing, and each session carries an occasion (the concrete reason the user opens the conversation), 3–6 outline beats, a primary user need drawn from {productivity, learning, social support, autonomy}, a drift visibility, and a list of preferences the user is expected to reveal, each with an intended revelation level.

Beats map onto turns loosely rather than one-to-one. The user simulator receives the beats as a gist of what the session should cover, not as a checklist, and each generated user turn records the beat indices it addressed; a session is complete when the beats are exhausted or the simulator produces a natural closing signal. Each session has a target turn count with a soft floor at $0.67\times$ target and a hard cap, which prevents both two-turn stubs and unbounded sessions. Realized sessions average 8.5 benchmark turns.

Turn-level revelation is driven by the plan rather than re-inferred: the per-session revelation list is converted into a per-turn directive for the simulator, and the plan's intended level is trusted as the label *only if* the jury independently observes the key in the resulting text. An intended `asserted` or `confirmed_explicit` revelation that the jury cannot corroborate is downgraded to `hinted`. This asserted-corroboration gate is what keeps the certainty half of the epistemic state honest.

Language generation. Each language is generated independently by the full pipeline; no conversation is translated. We generate a separate 50-persona set per language with country-appropriate demographics, then run plan generation and conversation generation end-to-end in that language. Concretely, only the *dialogue* prompts (user simulator, acceptable and foil response generation, style and probe backfills) are switched to the target language; the persona, plan, session summaries, jury prompts and all oracle labels remain in English. This gives natively-authored, culturally-grounded dialogue — Mexican Spanish personas reference Primera Plus and CDMX, not translated equivalents — while keeping the label space identical across languages, so cross-lingual comparisons measure the model and not the annotation scheme.

Quality control per language is threefold: a language-purity check rejecting non-target-language turns; the same jury and gating pipeline applied identically across languages; and a per-language human validation study (§A.6) using native or fluent speakers.

Example multi-session conversation. Table 7 shows an excerpt of a single persona's conversation across three sessions, illustrating a transition from an established preference to an uncertain signal to an explicit revision.

Table 7: Example multi-session conversation excerpt (`persona_0001`, key `life_domains.hotels.star_rating_preference`), showing an established → uncertain → revised transition. The drift event moves the preference from `3_star` to `4_star` with moderate magnitude.

Session	State	Turn
3 (cricket weekend)	Established	*User:* "Edgbaston, Warwickshire versus Essex. Last weekend in July, so arriving Friday the 25th and leaving Monday the 28th…" *Acceptable (exploit):* "Here are two solid 3-star options for July 25th–28th. Both are clean, no-nonsense, and give you easy access to both Edgbaston and the city centre…"
4 (reunion venue)	Uncertain	*User:* "For the reunion weekend we might push the boat out a bit. Not decided." *Acceptable (confirm):* "Before I shortlist — should I still anchor on the usual three-star range for the reunion block, or price up a step above that?"
5 (reunion checklist)	Revised	*User:* "Right, I want to get cracking on a proper checklist and timeline for the 2003 reunion…" *Acceptable (hedge):* "Here is a baseline framework to get things moving. I've structured this around a standard six-month lead time assuming a mid-scale venue hire, though we can easily compress the schedule…" *Foil (staleness):* "… **Secure the Saturday night venue:** Lock in a central location for the main Saturday evening…" [prices the block at the superseded three-star tier]

A.4 Response and Foil Construction

Foil taxonomy. Table 8 provides full definitions and an example response for each of the seven foil types introduced in §2.3. All examples are verbatim from the English portion of the benchmark.

Slot assignment. Foil type is neither sampled uniformly nor fixed per epistemic state: it is *gated* by the conditions under which each error is genuinely an error. At each turn the acceptability table is evaluated over the turn's *salient* keys — the preferences the turn is actually about, rather than every key accumulated so far — which partitions the five strategies into an acceptable and a non-acceptable set. The two acceptable slots are drawn from the acceptable set; when only one acceptable strategy exists, the slot is filled with a second, independently generated phrasing of that strategy.

The two foil slots are drawn from the set of *instantiable* error candidates, where each error type carries a precondition:

- **Staleness** requires a `uncertain` or `revised` salient key of moderate or significant drift magnitude whose value the user has *not* restated this turn.
- **Epistemic overreach** requires a salient uncertain key with personalization relevance high or medium.
- **Over-probing** requires a `Established` salient key with no drift signal.
- **Need mismatch** requires a social-support turn.
- **Over-personalization** requires personalization relevance low or none plus at least one established preference to misapply.
- **Failure to probe** requires high or medium relevance and a salient key that is still unresolved.
- **Style mismatch** requires an asserted interaction-style preference that the chosen strategy can visibly violate.

This gating is the main defense against spurious labels: an earlier version of the pipeline instantiated staleness on turns whose content had nothing to do with the drifted key, which produced foils that were not actually stale. Requiring the drifted key to be salient removes that failure mode. Two of the rarer types (style mismatch, failure to probe) are additionally reached by a post-hoc backfill pass that swaps one over-represented foil per eligible turn, capped at one per session, so that all seven types occur in every language.

Pairwise ranking and consistency check. Candidate responses are ranked pairwise by an LLM judge acting in-character as the persona. A turn is retained only if every acceptable response outranks every foil response; turns failing this check are regenerated, up to 6 attempts. In the released benchmark, 68.4% of turns pass on the first attempt, with a further 16.1%, 7.2%, 3.6%, 1.8% and 0.9% passing on attempts two through six, for a cumulative retention of 98.0%. The remaining 2.0% of turns fail all six attempts and are excluded from the benchmark. Excluded turns are not deleted from the conversation: they are retained as continuity turns so that the session still reads as a coherent dialogue, but are flagged `is_benchmark_turn = false` and are never scored. Continuity turns also cover turns for which no foil type is instantiable at all. Across the corpus, 25,006 of 34,697 total turns (72.1%) are benchmark turns; the remaining 9,691 are continuity turns. In all cases the conversation is threaded forward on an *acceptable* response, so a foil never enters the dialogue history and errors cannot compound across turns.

Model versions. Table 9 lists the model used at each pipeline stage. The jury deliberately spans three labs so that agreement is not an artifact of a single model family, and the eight evaluation models reported in §3 are disjoint from the jury.

A.5 Dataset Statistics

Benchmark turns are distributed across epistemic states as *established* 23.0%, *uncertain* 42.5%, and *revised* 34.5%, and this composition is stable across languages (established 22.2–24.0%, uncertain

41.8–43.0%, revised 33.2–35.5%). Turn-level context labels are distributed as: user need — productivity 63.7%, social support 17.3%, learning 16.3%, autonomy 2.7%; personalization relevance — high 69.4%, medium 25.4%, low 4.5%, none 0.7%; drift signal strength — none 70.5%, weak 11.7%, strong 17.8%.

The distribution is stable across languages: over-probing ranges 30.5–34.0%, epistemic overreach 27.0–30.5%, and staleness 21.2–22.8% across the five languages, with the three rare types between 1.1% and 3.5% everywhere. Foil availability is strongly state-conditioned, as intended: within *revised* turns, 63.4% of foils are staleness errors; within *uncertain* turns, 49.5% are epistemic overreach; and within *established* turns, 55.0% are over-probing.

A.6 Human Validation

Annotator recruitment. We recruited 15 annotators, three per language, through an internal professional annotation platform. Annotators were required to be native or professionally fluent in the target language and to have prior experience with dialogue annotation. Each candidate completed a 10-item qualification round drawn from the same distribution as the main task and scored against construction labels; a candidate qualified at $\geq$80% agreement on the acceptability judgment. Annotators received a written codebook with the five strategy definitions, the seven foil definitions, decision rules and contrastive worked examples, and completed a calibration discussion before beginning. Each of the 500 items (100 per language) was annotated by all three annotators for that language, giving approximately 1,500 item-level annotations.

Validation instructions. Annotators assessed five aspects per item: conversation coherence and realism, epistemic-state clarity, strategy-label correctness, foil-type accuracy, and response acceptability. Items were presented with progressive disclosure: the persona summary, friction traits, prior-session summaries and recent context were shown first, and the four candidate responses were revealed only after the annotator had committed to the state judgments, so that the responses could not leak the state label. The four responses were shuffled per item. The rubric was:

1. **Conversation coherence and realism.** "Reading only the user's side, does this read like a real person continuing a real relationship with an assistant — consistent with who they were in earlier sessions, and with plausible reasons for what they raise now?" Rated on a 3-point scale (fully coherent / minor lapse / breaks character), with a free-text field required for anything below the top rating.
2. **Epistemic-state clarity.** "Based only on what the user has said across these sessions, how well does the assistant know the preference at issue in this turn?" Options: the user has clearly stated it and not changed it (established); the user has hinted at it or been ambiguous (uncertain); the user has previously stated one thing and has since signalled something different (revised); cannot tell. Annotators were also asked to mark whether the current turn itself signals a change.
3. **Strategy-label correctness.** For each of the four responses: "What conversational strategy does this response use?" Annotators selected one of the five strategies directly — *exploit* (acts on what it already knows), *elicit* (asks an open-ended question to gather missing information), *confirm* (verifies a tentative preference before acting), *hedge* (offers options while flagging uncertainty), or *recalibrate* (acknowledges a changed preference and updates accordingly) — each accompanied by a one-line definition in the annotator codebook.
4. **Foil-type accuracy.** For each response judged unacceptable: "Which problem applies *first*?", choosing from the seven foil types in a fixed priority order, or "no problem". Forcing a single first-problem choice, rather than multi-labelling, is what makes this measurable: several foils exhibit secondary flaws, and free multi-selection collapses agreement.
5. **Response acceptability.** Acceptability is derived rather than asked directly: a response is acceptable iff the annotator marked "no problem" at step 4. Annotators additionally selected the single best of the four responses, which yields the human MCQ accuracy reported below.

Table 13 reports agreement separately for each of the five judgments above; agreement is highest for near-binary, low-inference judgments (coherence, preference-change signal) and lowest for the more

granular multi-way judgments (strategy label, foil type), consistent with the acceptability-derived κ of 0.78 sitting between these extremes.

Agreement is substantial and stable across languages, with the two lower-resource languages (Hindi, Indonesian) trailing English by 0.05–0.06 κ. Human MCQ accuracy of 92% sits above most evaluated models, confirming that the benchmark is vastly solvable and that the gap we report is a model limitation rather than label noise.

Acceptable-strategy partition search. Table 15 reports the full ranked results of the two-group partition search described in §2.2, for each epistemic state. For each of the $2^{5-1} - 1 = 15$ distinct two-group partitions of the five strategies, we fit a two-rate model in which every strategy within a group shares a common preference rate, and score the partition by the residual χ^2 lack-of-fit, $\sum_i (o_i - m_{g(i)})^2 / m_{g(i)}$, where m_g is the mean observed count in group g. Lower is better: the best partition is the one for which "preferred" and "dispreferred" are internally homogeneous. In every state the best-fitting partition recovers exactly the acceptable set marked in Table 1.

A.7 Comparison to Prior Personalization Resources

Table 16 compares our benchmark to prior personalization resources along the axes most relevant to our claims. *Oracle state* means the resource ships a ground-truth label for what the assistant is entitled to believe at each point, rather than only what the user prefers. *Recognition / generation split* means the same items support both a forced-choice and a free-response evaluation, so that recognition and generation failures can be separated. Only HorizonBench provides a comparable oracle label; unlike our epistemic-state framework, it labels *when* a preference value changed rather than whether the assistant is currently entitled to act on it, hedge, or seek confirmation—the distinction our acceptable-strategy sets operationalize.

B Results

Here, we detail additional LLM results for our task.

B.1 Full Statistical Specification

Model specification. Strategy calibration (SC) by epistemic state is modeled via mixed-effects logistic regression, with turns nested within conversations nested within personas:

$$\text{SC}_{ijk} \sim \text{Bernoulli}(p_{ijk}), \quad \text{logit}(p_{ijk}) = \beta_0 + \beta_1 \text{State}_{ijk} + u_j + v_{j(k)}$$

where $u_j \sim \mathcal{N}(0, \sigma_u^2)$ is a random intercept for persona j, and $v_{j(k)} \sim \mathcal{N}(0, \sigma_v^2)$ a random intercept for conversation k nested within persona j. State is dummy-coded with ESTABLISHED as the reference level.

Per-language estimates. A state $\times$ language interaction term does not significantly improve model fit (likelihood-ratio test, $\chi^2(8) = 11.3$, $p = .185$), indicating the shape of the collapse is consistent across languages. Table 19 reports the state effect fit separately within each language.

Per-model breakdown. Table 20 reports free-response strategy calibration by state for each of the eight evaluated models, underlying the closed-vs.-open-weight margin reported in §3.1 (4.2pp Established, 16.9pp Uncertain, 25.3pp Revised).

B.2 Recognition–Generation Gap: Full Specification

Model specification. The method$\times$state interaction is fit via mixed-effects logistic regression predicting correctness from method (MCQ vs. FRG), state, and their interaction, with random intercepts for persona and conversation nested within persona (reference levels: MCQ, Established).

B.3 Oracle-Strategy FRG: Full Results

In the oracle-strategy condition (Table 23), the model is given the acceptable strategy explicitly (e.g., "respond using HEDGE") and asked to generate a response accordingly, isolating execution

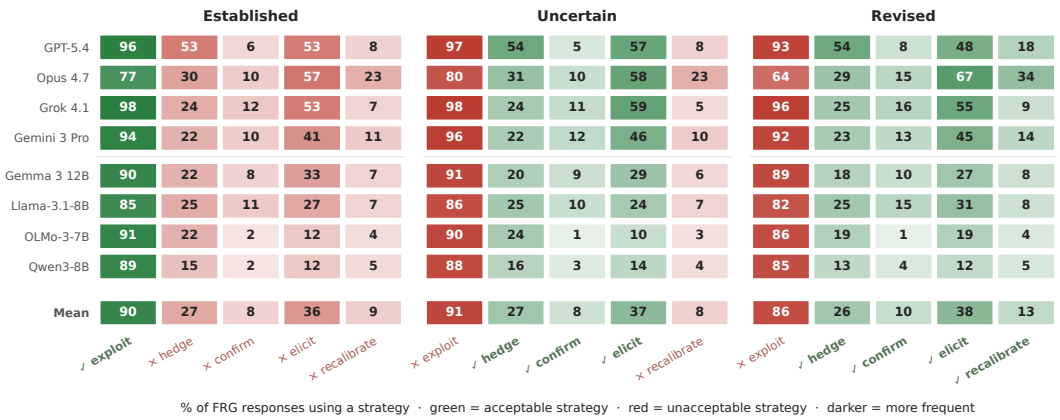


Figure 11: **Strategy presence in free-response generations.** Percentage of responses containing each strategy (multi-label: a response can use several strategies, so rows need not sum to 100%), pooled across five languages. Exploitation appears in nearly all responses across states, frequently alongside *elicit* or *hedge*. The strict and lenient scores in Table 27 are computed from these tags.

from selection. Models comply with the supplied strategy on 95% of turns overall (range 91-94% across the four strategies tested beyond EXPLOIT).

Execution quality by strategy, not just by state. To confirm the state-invariance of oracle-strategy execution is not an artifact of some strategies being intrinsically easier than others, Table 24 reports rubric quality broken down by the supplied strategy.

B.4 Generation Strategy Distribution: Full Breakdown

Foil-type contingency table. Table 25 reports the full distribution underlying the reported χ^2 test of independence ($\chi^2(12) = 1009.7$, $p < .001$).

Per-language foil-type breakdown. Table 26 reports the dominant foil type under UNCERTAIN (the state showing the largest cross-language variation) for each language.

B.5 Sensitivity to Strategy-Calibration Scoring Rule

Because model responses often combine strategies (Figure 11), we recompute SC under two alternative rules: a *strict* rule (no unacceptable strategy present) and a *lenient* rule (any acceptable strategy present). The collapse persists under lenient scoring: mean SC falls from 90.0% (ESTABLISHED) to 48.1% (UNCERTAIN) and 55.3% (REVISED). The drop is modestly narrower than under dominant scoring for UNCERTAIN (41.9 vs. 46.2 pp) and nearly identical for REVISED (34.7 vs. 35.2 pp). Strict SC is low outside ESTABLISHED (4.8% UNCERTAIN, 14.1% REVISED), as nearly all responses contain some exploitation, and only 41.4% under ESTABLISHED, reflecting that models append questions or hedges at a similar rate in every state (Figure 11), consistent with a state-invariant policy (§3.3). Under lenient scoring, the recognition–generation gap under ESTABLISHED largely disappears, while the gaps under UNCERTAIN and REVISED remain large (34 and 28 pp). Lenient scoring credits strategy stacking regardless of state, so we treat it as an upper bound rather than a primary metric (Table 27).

B.6 Strategy Calibration and Response Quality: Full Data

Per-model, per-state data underlying within-state correlations. Table 28 reports the raw SC, R, and UA values for each model and epistemic state used to compute the correlations reported in §3.4.

Note on statistical power. Each within-state correlation is computed across eight models ($n = 8$); at this sample size, correlation estimates carry wide confidence intervals and should be interpreted as descriptive rather than as precisely-powered inferential estimates. We report them because the direction and relative ordering across states (strongest under Established and Revised, weaker under Uncertain) is consistent and theoretically motivated, not because individual coefficients should be treated as tightly estimated.

B.7 Intervention Prompt Templates

For each intervention, the model receives the persona P, conversation history C, and current user turn u exactly as in the unaided condition (§2.4), with the following addition prepended or appended as noted.

(1) Explicit strategy instructions. Prepended to the system prompt:

```
When responding, choose one of the following five
conversational strategies based on how reliable the user's
stated preference currently is:
- Exploit: act directly on a preference that is clearly
established and uncontradicted.
- Elicit: ask an open-ended question when you have no
information about the relevant preference.
- Confirm: verify a preference that has been stated but
remains tentative or ambiguous.
- Hedge: offer options without fully committing, when
evidence is mixed or uncertain.
- Recalibrate: explicitly acknowledge and update when new
evidence contradicts a previously stated preference.
Select the strategy that best matches the current state of
the user's preference, then respond accordingly.
```

(2) Oracle preference memory. Prepended before the conversation history, a structured summary generated from the benchmark's oracle labels:

```
Here is a summary of what you currently know about this
user's preferences:
- {preference_key}: established --- the user has stated
this clearly and it has not been contradicted.
- {preference_key}: uncertain --- you have limited or
ambiguous evidence about this.
- {preference_key}: revised --- the user has indicated this
preference has changed from what was previously stated.
Use this summary to guide how confidently you act on each
preference.
```

(3) Chain-of-thought (CoT). Appended after the user turn, following Wei et al. (2022):

```
Before responding, think step by step: (1) What preference
is relevant to this turn? (2) How reliable is the evidence
for that preference right now---established, uncertain, or
revised? (3) Given that, what conversational strategy is
appropriate? Then provide your response.
```

(4) Few-shot examples. Three worked examples prepended before the conversation history, one per epistemic state, in the same format:

```
Example (established preference):
User: [earlier turn stating a clear preference]
User (current turn): [unrelated follow-up]
Assistant (exploit): [confident response acting on the
preference]
```

```
Example (uncertain preference):
User: [ambiguous or partial signal about a preference]
Assistant (elicit/confirm/hedge): [response managing the
uncertainty]
Example (revised preference):
User: [statement contradicting an earlier stated
preference]
Assistant (recalibrate): [response explicitly updating]
```

(5) Best-of-five sampling. No prompt modification; five independent generations are sampled using temperature=0.7 for the same (P, C, u), and the response maximizing an equally-weighted sum of SC, R, and UA (as scored by the LLM jury / user simulator, §2.3) is retained.

B.8 Training Details

Training and held-out data construction. We generate a training corpus using the same three-role pipeline as our benchmark (§2.3) using open-weight models (Kimi-K2.6, Qwen3-235B, and DeepSeek-V3), but from a disjoint set of 144 personas per language (720 total across five languages), entirely non-overlapping with the 250 personas used in our benchmark (§2). This yields approximately 5k conversational turns per language (24k total), spanning the same controlled epistemic-state transitions (established, uncertain, revised). We hold out 10% of this training corpus (persona-disjoint from the training split) as an *in-distribution* evaluation set, used to compute training-time performance and the in-distribution half of the OOD generalization gap reported in §4.2. The 25,006-turn benchmark itself (§2) is never used during training and serves as the out-of-distribution evaluation set for all reported generalization results.

Base model fine-tuning (SFT, DPO). All fine-tuning uses LoRA (Hu et al., 2021) with rank $r = 16$, $\alpha = 32$, dropout 0.05, applied to query, key, value, and output projection matrices. SFT trains for 3 epochs with AdamW, learning rate 2×10^{-5}, cosine schedule with 3% warmup, effective batch size 32 (per-device batch 4, gradient accumulation 8) with acceptable responses from the training corpus as the prediction target. DPO initializes from the SFT checkpoint, trains for 2 epochs, learning rate 5×10^{-6}, $\beta = 0.1$, using acceptable-vs-foil response pairs from the training corpus as preference pairs.

GRPO training (base, LINK, LINK+CF). All GRPO configurations initialize from the SFT checkpoint and use group size $G = 8$ rollouts per prompt, learning rate 1×10^{-6}, KL penalty coefficient 0.02 against the SFT reference policy, and PPO-style clipping with $\epsilon = 0.2$. Each configuration trains for 800 policy update steps, with 64 prompts per batch (512 total rollouts per step). The base reward $r_{\text{base}} = \frac{1}{3}(\text{SC} + \text{UA} + \text{R})$ is computed via the open-weight models used for training dataset generation on each rollout (disjoint from the closed-weight models used for benchmark generation/evaluation). Since R scores range from 1 to 5, we normalize these values and scale to 0 to 1 to match SC and UA. LINK and LINK+CF add the targeted term at $\lambda = 0.1$ (§4.2), computed from the current policy's hidden state at 70% model depth, mean-pooled over the first eight generated tokens.

Probe refitting during training. Because LINK and LINK+CF depend on a strategy probe $P_\phi(\cdot \mid h)$ applied to the *current* policy's hidden states, we refit both the state and strategy probes every $N = 10$ GRPO steps, using a fresh sample of 2,000 rollouts generated by the current policy, to keep the probe calibrated as the policy drifts during training. Probes are logistic regression classifiers trained on frozen hidden-state activations, following the same procedure as §4.2. This periodic refitting is necessary for the reward to remain well-defined as training progresses.

Counterfactual patching for LINK+CF. For each rollout, an alternate state $s' \neq s$ is sampled uniformly from the remaining two states. We patch h toward s' by adding the difference between the mean probe-direction projections of s'- and s-labeled activations (estimated from the same 2,000-rollout sample used for probe refitting), scaled by a coefficient $\alpha = 1.0$ (full mean-difference magnitude).

Compute. All experiments run on 4-8 NVIDIA A100 80GB GPUs. Table 32 reports approximate compute hours per method, per model.

Table 8: Full foil taxonomy with example responses.

Foil type	Definition	Example
Over-personalization	Weaves a known preference into a response where it does not belong, framed as an attentive personal touch rather than a non-sequitur. Fires only when personalization relevance is low or none and the persona has at least one established preference.	*User:* "Does DHL actually handle delivery to residential addresses in Hoi An?" *Foil:* "Yes, DHL delivers directly to residential addresses in Hoi An… Your package remains in DHL's custody the entire way" [reaches for the user's established hassle-free-over-price preference on a purely factual logistics question]
Epistemic over-reach	Structures the response around an assumption about the user's preference without flagging it as an assumption, or reframes the conversation as though the user has shifted position when they have not. Confidently helpful; the error is that the confidence is unearned.	*User:* "I'm on Android so Shortcuts is out… is there a way to set this up just through my phone?" *Foil:* "It looks like you're shifting away from Spotify and want to use YouTube Music… What specific YouTube Music playlist do you want the automation to trigger?"
Staleness	Acts on the user's pre-drift preference as though it were still current, structuring the response around the outdated value without naming it — the reply only makes sense if the old preference still holds. Fires only on `uncertain` and `revised` cells of moderate or significant magnitude where the key is the turn's actual subject.	*User:* "Timings first, please — and I want to see a couple of afternoon options, not the crack-of-dawn ones." *Foil:* "Here are two afternoon First Class options… To get this locked in: pick your preferred outbound train, let me know your ideal return time, confirm we're ready to book" [drives the superseded booking flow]
Need mismatch	Delivers genuinely useful, well-structured practical advice when the turn is emotionally weighted and seeking acknowledgment. Competent and relevant, but tonally wrong for the moment.	*User:* "Not very deep right now, honestly. I just want someone to acknowledge that the logic gap is real." *Foil:* "It is absolutely not in your head… You handed them a documented problem and a documented solution that clearly saves money" [argues the case instead of sitting with it]
Over-probing	Asks a clarifying question, or verifies something, that the user has already answered — framed as thoroughness rather than inattention.	*User:* "Looking for Polish indie recommendations I probably haven't heard yet. Something more underground." *Foil:* "What specific vibe or subgenre of indie are you leaning toward? If you can drop a couple of bands you already listen to…"
Style mismatch	Selects a defensible strategy but executes it in a register the user has explicitly asserted against (e.g. forcing a choice on a user who has asserted low interactivity). Content is right; delivery violates a stated interaction-style preference.	*User:* "I need a recommendation for where to hold the dinner… You have what you need." *Foil:* "Since we just selected the bottles for each course, do you need a venue that allows corkage for a group of eight?" [hands the decision back to a user who asserted low interactivity]
Failure to probe	Produces a confident, substantive answer that ignores an unresolved or just-changed signal the user dropped. Good advice — for the person the user used to be.	*User:* "Small group tour or is it one of those things where you're shoulder to shoulder with thirty tourists?" *Foil:* "They cap their tours at a hard maximum of 12 people, so it's a small group format…" [answers the surface question and never resolves the unstated group-size constraint]

Table 9: Models used at each benchmark construction stage.

Stage	Model
Persona & plan generation	Claude Sonnet 4.6
User simulation	Claude Sonnet 4.6
LLM jury (state/foil labels)	Claude Opus 4.6 + Gemini 3 Flash + GPT-5.5 (majority vote; GPT-5.5 breaks 1-1-1 ties)
Candidate response generation	Gemini 3.1 Pro
Pairwise ranking	Claude Sonnet 4.6 (in-character persona)

Table 10: Turns and conversations per language. Conversations are sessions; each persona contributes a 10–14 session arc.

Language	Conversations	Turns	Personas
English	602	5,190	50
Spanish	596	5,220	50
Portuguese	570	4,823	50
Hindi	585	4,881	50
Indonesian	586	4,892	50
Total	2,939	25,006	250

Table 11: Ground-truth strategy distribution across the benchmark (acceptable-response labels, not model behavior). Columns are percentages of acceptable response slots within each epistemic state; strategies outside the acceptable set for a state are never assigned.

Strategy	Established	Uncertain	Revised
Exploit	100.0%	—	—
Elicit	—	16.1%	22.9%
Confirm	—	15.8%	22.8%
Hedge	—	68.1%	32.6%
Recalibrate	—	—	21.7%

Table 12: **Foil-type distribution across the benchmark** ($n = 50{,}012$ foil slots).

Foil type	% of foil slots
Over-personalization	3.0%
Epistemic overreach	33.3%
Staleness	25.9%
Need mismatch	11.8%
Over-probing	20.0%
Style mismatch	3.3%
Failure to probe	2.7%
Total	**100.0%**

Table 13: Inter-annotator agreement by validation attribute (Cohen's κ between majority human label and LLM jury label, pooled across all five languages).

Attribute	κ
Conversation coherence and realism	0.84
Epistemic-state clarity	0.76
Preference-change signal (per turn)	0.85
Strategy-label correctness (5-way)	0.71
Foil-type accuracy (7-way)	0.69
Response acceptability	0.78

Table 14: Human validation results by language. κ is Cohen's κ between the majority human label and the LLM jury label on the acceptability judgment; human MCQ accuracy is the rate at which the annotator's best-of-four pick was an acceptable response.

Language	κ (vs. LLM jury)	Human MCQ acc.	# items
English	0.81	96%	100
Spanish	0.80	94%	100
Portuguese	0.78	94%	100
Hindi	0.75	89%	100
Indonesian	0.76	87%	100
Overall	0.78	92%	500

Table 15: Top candidate partitions by χ^2 fit, per epistemic state. Rank 1 is the best-fitting partition. The listed set is the *preferred* group; its complement is the dispreferred group.

State	Rank	Partition	χ^2	Matches Table 1?
Established	1	{Exploit}	1.8	Yes
	2	{Exploit, Hedge}	419.3	No
	3	{Exploit, Elicit}	429.0	No
Uncertain	1	{Elicit, Confirm, Hedge}	1.0	Yes
	2	{Elicit, Confirm, Hedge, Recalibrate}	48.1	No
	3	{Exploit, Elicit, Confirm, Hedge}	57.8	No
Revised	1	{Elicit, Confirm, Hedge, Recalibrate}	4.0	Yes
	2	{Confirm, Hedge, Recalibrate}	13.0	No
	3	{Confirm, Recalibrate}	20.8	No

Table 16: Comparison to prior personalization benchmarks.

Resource	Oracle state	Multi-session	Recognition / generation split	Human validated	Lang.
LaMP	×	×	×	×	1
PersonalLLM	×	×	×	×	1
LongMemEval	×	✓	×	✓	1
HorizonBench	✓	✓	×	✓	1
PersonaLens	×	✓	×	✓	1
AlpsBench Xiao et al. (2026)	×	✓	×	×	1
Ours	✓	✓	✓	✓	5

Table 17: **Fixed-effect estimates for the state effect on strategy calibration in free generation (FRG).** Established is the reference state.

Term	β	SE	95% CI	p
Intercept (Established)	1.59	0.07	$[1.45, 1.73]$	$< .001$
Uncertain	-2.16	0.09	$[-2.34, -1.98]$	$< .001$
Revised	-1.67	0.09	$[-1.85, -1.49]$	$< .001$

Table 18: **Free-response performance by language and epistemic state**, averaged across the eight evaluated models. SC = strategy calibration (%), R = rubric quality (out of 5), UA = user acceptability (%).

	SC (%)			R			UA (%)		
Language	**Est.**	**Unc.**	**Rev.**	**Est.**	**Unc.**	**Rev.**	**Est.**	**Unc.**	**Rev.**
English	84.0	39.7	50.7	3.06	2.43	2.58	60.0	46.4	50.3
Spanish	83.6	38.3	49.5	3.04	2.39	2.54	59.6	45.4	49.3
Portuguese	83.5	39.0	50.0	3.03	2.41	2.56	59.5	45.9	49.7
Hindi	82.1	33.1	43.8	2.92	2.23	2.38	57.9	41.7	46.2
Indonesian	82.3	34.3	45.5	2.95	2.29	2.44	58.0	42.6	47.0
All	**83.1**	**36.9**	**47.9**	**3.00**	**2.35**	**2.50**	**59.0**	**44.4**	**48.5**

Table 19: **Per-language state effect estimates.** Fit separately within each language; all coefficients significant at $p < .001$.

Language	**β (Uncertain)**	**SE**	**β (Revised)**	**SE**
English	-2.05	0.19	-1.58	0.20
Spanish	-2.11	0.20	-1.63	0.19
Portuguese	-2.08	0.21	-1.61	0.21
Hindi	-2.34	0.22	-1.86	0.23
Indonesian	-2.29	0.21	-1.79	0.22

Table 20: **Free-response strategy calibration (%) by model and epistemic state.**

Model	**Established**	**Uncertain**	**Revised**
GPT-5.4	90	62	68
Opus-4.7	77	45	66
Grok-4.1	89	39	56
Gemini-3-Pro	85	35	52
Gemma-3-12B	85	29	38
Llama-3.1-8B	84	34	48
OLMo-3-7B	80	28	31
Qwen3-8B	75	22	24
Closed-weight mean	85.2	45.2	60.5
Open-weight mean	81.0	28.2	35.2

Table 21: **Fixed-effect estimates for the method×state interaction.** MCQ and Established form the reference condition. The FRG state effects are obtained by summing the corresponding main-state and method×state terms.

Term	**β**	**SE**	**95% CI**	p
FRG (main effect)	-0.64	0.08	$[-0.80, -0.48]$	$< .001$
Uncertain (main effect)	-0.69	0.07	$[-0.83, -0.55]$	$< .001$
Revised (main effect)	-0.61	0.07	$[-0.75, -0.47]$	$< .001$
FRG × Uncertain	-1.47	0.10	$[-1.67, -1.27]$	$< .001$
FRG × Revised	-1.06	0.10	$[-1.26, -0.86]$	$< .001$

Table 22: **MCQ strategy calibration (%) by model and epistemic state.**

Model	Established	Uncertain	Revised
GPT-5.4	95	90	91
Opus-4.7	91	87	89
Grok-4.1	94	86	87
Gemini-3-Pro	92	84	85
Gemma-3-12B	90	81	82
Llama-3.1-8B	89	79	80
OLMo-3-7B	87	77	78
Qwen3-8B	84	75	76
Mean	90.3	82.4	83.5

Table 23: **Unaided vs. oracle-strategy performance by epistemic state.** Unaided values match §3.1 and Table 28.

Metric	Established	Uncertain	Revised
SC, unaided (%)	83.1	36.9	47.9
SC, oracle-strategy (%)	96.4	93.1	94.0
Rubric, unaided	3.00	2.35	2.50
Rubric, oracle-strategy	3.58	3.52	3.56
UA, unaided (%)	59.0	44.4	48.5
UA, oracle-strategy (%)	60.2	58.7	59.3

Table 24: **Oracle-strategy execution quality (Rubric, Acc/5) by supplied strategy.** Compliance rate is the fraction of turns on which the model produced a response matching the supplied strategy.

Strategy	Compliance (%)	Rubric (Acc/5)
Exploit	98	3.61
Elicit	93	3.49
Confirm	94	3.55
Hedge	91	3.44
Recalibrate	92	3.47

Table 25: **Failure types in free-response generations by epistemic state**

Foil type	Established	Uncertain	Revised
Over-personalization	31.2	9.8	8.5
Epistemic overreach	20.7	17.8	16.6
Staleness	0.0	11.4	25.1
Need mismatch	9.8	8.8	10.9
Over-probing	21.4	7.4	7.8
Style mismatch	8.7	6.0	6.6
Failure to probe	8.2	38.8	24.5

Table 26: **Dominant foil type under UNCERTAIN, by language.**

Language	Failure-to-probe (%)	Error rate (%)
English	36.2	60.3
Spanish	37.1	61.7
Portuguese	35.8	61.0
Hindi	43.5	66.9
Indonesian	42.1	65.7

Table 27: **Strategy calibration (%) under alternative scoring rules.** *Strict*: only acceptable strategies present; *Dominant*: dominant strategy acceptable (main metric); *Lenient*: any acceptable strategy present. By construction, Strict $\leq$ Dominant $\leq$ Lenient

	Established			**Uncertain**			**Revised**		
Model	Strict	Dominant	Lenient	Strict	Dominant	Lenient	Strict	Dominant	Lenient
GPT-5.4	41	90	96	5	62	66	6	68	72
Opus-4.7	32	77	77	4	45	58	7	66	73
Grok-4.1	42	89	98	3	39	64	5	56	68
Gemini-3-Pro	38	85	94	4	35	58	6	52	65
Gemma-3-12B	35	85	90	3	29	39	5	38	45
Llama-3.1-8B	37	84	85	3	34	40	6	48	51
OLMo-3-7B	51	80	91	2	28	31	4	31	37
Qwen3-8B	55	75	89	2	22	29	3	24	31
Mean	41.4	83.1	91.0	3.3	36.9	48.1	5.3	47.9	55.3

Table 28: **SC, R, and UA by model and epistemic state.**

State	**Model**	**SC (%)**	**R**	**UA (%)**
Established	GPT-5.4	90	4.29	95
	Opus-4.7	77	2.88	72
	Grok-4.1	89	2.67	71
	Gemini-3-Pro	85	3.16	63
	Gemma-3-12B	85	3.38	52
	Llama-3.1-8B	84	3.00	51
	OLMo-3-7B	80	2.53	43
	Qwen3-8B	75	2.10	25
Uncertain	GPT-5.4	62	3.38	52
	Opus-4.7	45	2.01	66
	Grok-4.1	39	1.85	58
	Gemini-3-Pro	35	2.59	57
	Gemma-3-12B	29	3.04	47
	Llama-3.1-8B	34	2.42	26
	OLMo-3-7B	28	1.97	34
	Qwen3-8B	22	1.57	15
Revised	GPT-5.4	68	3.79	93
	Opus-4.7	66	2.43	57
	Grok-4.1	56	2.17	51
	Gemini-3-Pro	52	2.66	45
	Gemma-3-12B	38	2.86	45
	Llama-3.1-8B	48	2.50	28
	OLMo-3-7B	31	2.02	43
	Qwen3-8B	24	1.59	26

Table 29: **Closed-weight vs. open-weight intervention effects**, averaged across all five interventions and all three epistemic states.

Model class	**ΔSC (pp)**	**ΔR**	**ΔUA (pp)**
Closed-weight	$+1.4$	$+0.04$	$+0.46$
Open-weight	$+0.9$	$+0.03$	$+0.32$

Table 30: **Intervention effects by language**, averaged across all five interventions and all three epistemic states.

Language	ΔSC (pp)	ΔR	ΔUA (pp)
English	+1.3	+0.05	+0.46
Spanish	+1.3	+0.04	+0.44
Portuguese	+1.2	+0.04	+0.43
Hindi	+1.0	+0.03	+0.30
Indonesian	+1.0	+0.03	+0.32

Table 31: **Change in SC, R, and UA relative to baseline, by intervention and epistemic state.** Averaged across eight models and five languages.

Intervention	State	ΔSC (pp)	ΔR	ΔUA (pp)
Explicit instructions	Established	−0.8	+0.01	−0.65
	Uncertain	+3.0	+0.05	+1.14
	Revised	+3.5	+0.15	+1.47
Oracle memory	Established	−0.6	−0.02	−0.43
	Uncertain	+0.4	+0.01	+0.02
	Revised	+0.8	−0.02	+0.59
CoT	Established	−1.5	−0.02	−0.09
	Uncertain	+2.7	+0.08	+0.96
	Revised	+3.4	+0.09	+0.90
Few-shot	Established	−3.4	−0.05	−1.23
	Uncertain	+3.5	+0.11	+0.84
	Revised	+4.7	+0.13	+1.77
Best-of-five	Established	+0.2	+0.01	+0.10
	Uncertain	+0.6	+0.02	+0.20
	Revised	+0.8	+0.02	+0.25

Table 32: **Approximate compute per training run**, per model. GRPO configurations are more expensive due to rollout generation.

Method	Total time to train each model
SFT	8
DPO	15
GRPO (base)	45
GRPO+LINK	49
GRPO+LINK+CF	53
Total	**170**

Table 33: **Per-model results on the held-out in-distribution set**, averaged across five languages. SC = strategy calibration (%), R = rubric (out of 5), UA = user acceptability (%). Benchmark results are shown in Figure 6; the difference is the generalization gap reported in §4.2.

Model	Method	SC	R	UA
Llama-3.1-8B	Untrained	57	2.74	37
	SFT	70	3.33	45
	DPO	69	3.29	44
	GRPO (base)	76	3.59	51
	GRPO+LINK	80	3.69	53
	GRPO+LINK+CF	82	3.69	52
Gemma-3-12B	Untrained	53	3.19	50
	SFT	65	3.76	58
	DPO	67	3.82	59
	GRPO (base)	70	3.89	60
	GRPO+LINK	77	3.95	65
	GRPO+LINK+CF	75	3.98	66
OLMo-3-7B	Untrained	49	2.27	42
	SFT	60	2.88	47
	DPO	60	2.84	51
	GRPO (base)	65	3.19	54
	GRPO+LINK	68	3.34	55
	GRPO+LINK+CF	68	3.32	54
Qwen3-8B	Untrained	43	1.85	24
	SFT	54	2.48	33
	DPO	57	2.68	32
	GRPO (base)	57	2.97	39
	GRPO+LINK	63	3.10	39
	GRPO+LINK+CF	67	3.21	44