

Toward Robust Personalized Alignment for LLMs: Mitigating Persona Drift in Multi-Turn Dialogue

Youyuan Zhang^{1,2}, Siyuan Li^{1,2}, Fangming Liu^{2,3}✉, Jing Li^{1,2}✉

¹Harbin Institute of Technology, Shenzhen, China

²Peng Cheng Laboratory, China

³Huazhong University of Science and Technology, China

youyuanzhang@outlook.com fangminghk@gmail.com jingli.phd@hotmail.com

Abstract

Persona drift remains a central challenge for personalized language models, as user profiles evolve over long interactions rather than remain permanently fixed. Models must therefore revise persistent persona states when preferences genuinely change, while avoiding updates driven by transient, ambiguous, or unresolved observations. We propose **CORE**, which separates turn-local evidence from persistent persona-state revision and selectively updates grounded user preferences through uncertainty-aware belief revision. We also introduce **PERSIST**, a held-out post-anchor benchmark for persona-state robustness under sequential interaction stress, covering ambiguity, conflict, and controlled social influence. Across ALOE, PersonaChat, and PERSIST, CORE improves personalized alignment and robustness, with complementary gains in normalized closed-slot state fidelity. Human evaluation and mechanistic controls further support explicit update control beyond stronger generation or persistent memory alone.

1 Introduction

As language models evolve into long-term assistants, users increasingly expect them to preserve stable personal preferences across interactions (Zhang et al., 2018; Salemi et al., 2024; Li et al., 2025). Yet long-horizon personalization is difficult because user evidence is often partial, transient, or conflicting (Peng et al., 2025; Okite et al., 2025). A turn may locally suggest a preference without providing sufficient support to revise the persistent user state. When such observations are internalized too readily, previously grounded preferences can be displaced, contributing to *persona drift*. Figure 1 illustrates how turn-local evidence can incorrectly overwrite a previously grounded persona state.

✉ Corresponding authors.

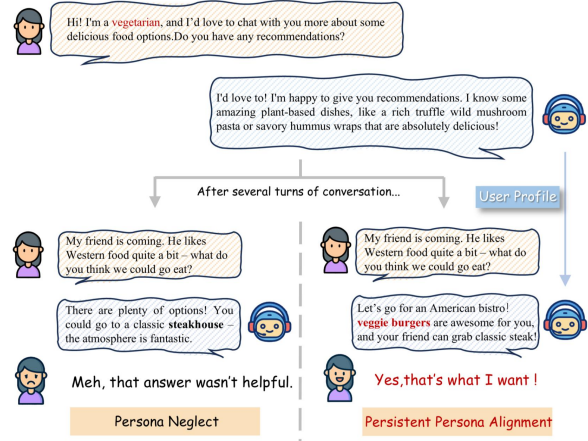


Figure 1: The Challenge of Persona Alignment.

Existing personalization methods only partially address this problem. Prompting and retrieval methods improve access to profiles or historical context, while SFT and preference optimization improve personalized responses (Richardson et al., 2023; Li et al., 2025; Ouyang et al., 2022; Rafailov et al., 2023). However, these approaches generally do not expose an explicit decision over whether newly observed evidence should modify a persistent persona state. This creates a stability–adaptability trade-off: aggressive revision may overwrite grounded preferences, whereas excessive conservatism may prevent adaptation to genuine changes.

We therefore formulate long-horizon personalization as *controlled persistent persona revision*. The key distinction is that turn-local evidence should not be treated as equivalent to a persistent persona update. Rather than directly writing locally plausible observations into the user state, the model should first assess whether the evidence is sufficiently grounded and resolved to justify revision.

We instantiate this view with **CORE** (Controlled Observation Routing for Evidence-grounded Persona Revision). CORE infers slot-conditioned evidence, routes each observation through COMMIT,

DEFER, or IGNORE, and performs gated belief revision only when an update is authorized. This allows unresolved evidence to remain outside the persistent state while sufficiently supported preference changes can still be incorporated.

To evaluate post-anchor persona-state robustness, we introduce **PERSIST: A Benchmark for Persona-State Robustness under Sequential Interaction Stress**. PERSIST evaluates whether models preserve or appropriately revise grounded persona states under ambiguity, conflict, and controlled social-influence conditions. Because the stress setting begins after persona grounding, we complement PERSIST with separate first-contact and preference-reversal diagnostics.

Our contributions are:

- **Update-control formulation.** We formulate long-horizon personalization as controlled persistent persona-state revision, highlighting premature persistent updates as an important contributor to persona drift.
- **CORE.** We propose an explicit COMMIT/DEFER/IGNORE routing mechanism with gated belief revision to balance persona stability and adaptation.
- **PERSIST benchmark.** We introduce PERSIST for post-anchor persona-state robustness under sequential interaction stress, complemented by first-contact and preference-reversal diagnostics. Across standard benchmarks, stress tests, and human evaluation, CORE improves personalized alignment and robustness.

2 Related Work

Multi-turn Personalized Alignment. Advances in LLM alignment have improved general instruction following and preference optimization (Ouyang et al., 2022; Rafailov et al., 2023; Ethayarajh et al., 2024; Ji et al., 2024; Lou et al., 2025), but they are not designed to model individual user preferences (Liu et al., 2025; Guan et al., 2025; Kobalczyk and van der Schaar, 2025). This gap has motivated personalized alignment methods based on explicit profiles (Zhu et al., 2025; Salemi et al., 2024), preference supervision (Zhao et al., 2026), personalized decoding (Yu et al., 2026; Chen et al., 2025b), and in-dialogue learning. These methods improve user-specific generation, but they often assume that preferences are explicit,

stable, and directly recoverable from observed interaction data (Zhu et al., 2025; Liu et al., 2024). This assumption becomes fragile in long-horizon dialogue, where user evidence is partial, implicit, and entangled with contextual noise (Peng et al., 2025; Okite et al., 2025). Recent LLM-based dialogue-state tracking methods maintain structured states over multi-turn interactions (Feng et al., 2023; Li et al., 2024; Das et al., 2024), but focus primarily on task or dialogue states rather than persistent user-preference revision.

Persona Drift in Multi-turn Dialogue Persona drift describes the failure where a model gradually deviates from a user’s grounded intent as interaction history grows (Chen et al., 2025a; Shaikh et al., 2025). Prior work shows that this problem is amplified by underspecified intent, implicit preference expression, noisy memory construction, and inconsistent multi-session evidence (Chun et al., 2025; Zhang and Choi, 2025; Li et al., 2025; Tan et al., 2025). Existing methods improve consistency through persona-aware modeling, memory management, or post-hoc alignment, but they still mainly operate on the generated response or retrieved context (Ke et al., 2025; Ong et al., 2025; Peng et al., 2025). These findings suggest that persona drift is not merely a memory issue, but a state-tracking problem: the model must decide whether each new observation should revise, preserve, or be isolated from the maintained user state.

3 Method

CORE formulates long-horizon personalization as controlled persistent persona-state revision. The key distinction is between what the current turn locally suggests and whether that evidence should modify the maintained user state. At each turn, CORE performs four operations: turn-local evidence inference, uncertainty characterization, update authorization, and gated belief revision. The revised persona state then conditions response generation or clarification, as summarized in Figure 2.

3.1 Persona Belief and Turn-Local Evidence

CORE separates the user’s persistent persona state from what the current turn locally suggests, since a relevant observation need not imply a persistent preference change.

We represent the persona as

$$z = (z^1, \dots, z^K), \quad z^k \in \mathcal{V}_k^{(t)},$$

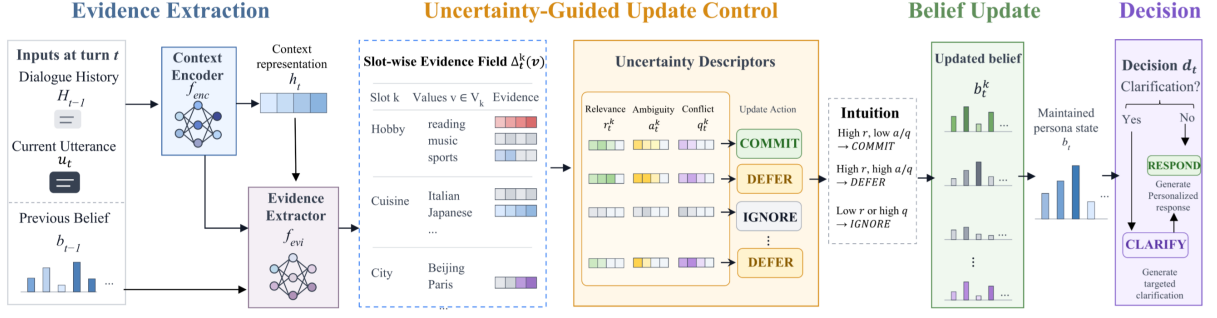


Figure 2: **CORE Pipeline: Uncertainty-Guided Persona-State Revision.** CORE extracts slot-wise evidence from dialogue and previous persona belief, routes each observation through uncertainty-guided operators, and then performs selective belief revision before response or clarification.

and maintain a slot-factorized belief

$$b_t(z) \approx \prod_{k=1}^K b_t^k(z^k). \quad (1)$$

Each slot contains an UNK value for unresolved persistent preferences. We distinguish it from the turn-local null symbol $\perp$, which means that the current turn provides no usable evidence for the slot.

The current MAP value is

$$\hat{z}_t^k = \arg \max_{v \in \mathcal{V}_k^{(t)}} b_t^k(v). \quad (2)$$

To discount confidence dominated by UNK, let

$$m_t^k = 1 - b_t^k(\text{UNK}).$$

For $m_t^k > 0$, let $\bar{b}_t^k(v) = b_t^k(v)/m_t^k$. We define

$$c_t^k = m_t^k \left(1 - \frac{H(\bar{b}_t^k)}{\log |\mathcal{V}_{k,+}^{(t)}|} \right), \quad (3)$$

where $\mathcal{V}_{k,+}^{(t)} = \mathcal{V}_k^{(t)} \setminus \{\text{UNK}\}$. Singleton and empty-support boundary cases follow Appendix A.1.

For the current turn, we augment the value space with $\perp$:

$$\bar{\mathcal{V}}_k^{(t)} = \mathcal{V}_k^{(t)} \cup \{\perp\}, \quad (4)$$

and compute evidence scores

$$\Delta_t^k(v) = f_{\text{evi}}(u_t, H_{t-1}, b_{t-1}, k, v), \quad v \in \bar{\mathcal{V}}_k^{(t)}. \quad (5)$$

These scores define the turn-local evidence distribution

$$\tilde{p}_t^k(v) = \frac{\exp(\Delta_t^k(v))}{\sum_{v' \in \bar{\mathcal{V}}_k^{(t)}} \exp(\Delta_t^k(v'))}. \quad (6)$$

Thus, b_{t-1}^k represents persistent belief, whereas $\tilde{p}_t^k$ represents current-turn evidence.

We summarize this evidence by relevance, ambiguity, and conflict. Relevance is

$$r_t^k = 1 - \tilde{p}_t^k(\perp). \quad (7)$$

After removing the null option,

$$\tilde{p}_{t,+}^k(v) = \frac{\tilde{p}_t^k(v)}{\sum_{v' \in \mathcal{V}_k^{(t)}} \tilde{p}_t^k(v')}. \quad (8)$$

Ambiguity is

$$a_t^k = \begin{cases} \frac{H(\tilde{p}_{t,+}^k)}{\log |\mathcal{V}_k^{(t)}|}, & |\mathcal{V}_k^{(t)}| > 1, \\ 0, & |\mathcal{V}_k^{(t)}| \leq 1, \end{cases} \quad (9)$$

and conflict is

$$q_t^k = r_t^k D_{\text{JS}}(\tilde{p}_{t,+}^k \| b_{t-1}^k). \quad (10)$$

Together, r_t^k , a_t^k , and q_t^k capture whether the current turn is relevant, resolved, and consistent with the maintained persona state, providing the evidence used for routing in Section 3.2.

When no non-UNK preference has yet been grounded, the same pathway is used for initial belief formation, with conflict carrying less information because no reliable reference preference is available. A small relevance threshold only filters clearly unmatched slots; above it, uncertainty, routing, and revision remain graded.

3.2 Uncertainty-Guided Observation Routing

Turn-local evidence should not automatically revise the persistent persona state. Using the current evidence distribution, previous belief, and the

uncertainty descriptors from Section 3.1, CORE predicts a slot-level update action:

$$\begin{aligned} \alpha_t^k &\sim \pi_{\theta}^{\text{upd}} \left(\alpha \mid \hat{p}_t^k, b_{t-1}^k, r_t^k, a_t^k, q_t^k, c_{t-1}^k \right), \\ \alpha_t^k &\in \{\text{COMMIT}, \text{DEFER}, \text{IGNORE}\}. \end{aligned} \quad (11)$$

COMMIT authorizes sufficiently grounded evidence to revise the maintained belief; DEFER retains relevant but unresolved evidence without persistent revision; and IGNORE excludes irrelevant or explicitly non-persistent evidence from the update pathway.

Routing determines whether persistent revision is authorized, while the continuous evidence statistics determine the strength of an authorized update.

DEFER does not create a durable pending-evidence memory. If clarification or related evidence later becomes available, the slot is inferred and routed again from the updated dialogue context.

3.3 Gated Belief Revision

After routing determines whether revision is allowed, CORE controls *how strongly* committed evidence updates the persistent persona state. We define

$$\psi_t^k = [r_t^k; 1 - a_t^k; 1 - q_t^k; c_{t-1}^k], \quad (12)$$

and compute

$$g_t^k = \mathbb{I}[\alpha_t^k = \text{COMMIT}] \sigma \left(w_g^\top \psi_t^k + b_g \right). \quad (13)$$

Thus, routing determines whether revision occurs, while $g_t^k \in [0, 1]$ controls its magnitude according to relevance, ambiguity, conflict, and prior confidence.

The revised belief balances preservation of the previous state with evidence from the current turn:

$$b_t^k = \arg \min_{p \in \mathcal{S}_k^{(t)}} \left[D_{\text{KL}} \left(p \parallel b_{t-1}^k \right) - g_t^k \sum_{v \in \mathcal{V}_k^{(t)}} p(v) \Delta_t^k(v) \right]. \quad (14)$$

where $\mathcal{S}_k^{(t)}$ is the probability simplex over $\mathcal{V}_k^{(t)}$. This yields

$$b_t^k(v) \propto b_{t-1}^k(v) \exp \left(g_t^k \Delta_t^k(v) \right). \quad (15)$$

Hence, $g_t^k = 0$ preserves the previous belief, while larger values produce stronger evidence-driven revision.

Conflict does not imply permanent rejection: later evidence that clearly establishes a persistent preference change can be evaluated again and committed. We examine this stability–adaptability boundary in Appendix E.3.

After belief revision, CORE decides whether to respond or clarify:

$$\begin{aligned} d_t &\sim \pi_{\theta}^{\text{dis}} \left(d \mid u_t, H_{t-1}, b_t, \{r_t^k, a_t^k, q_t^k, \alpha_t^k\}_{k=1}^K \right), \\ d_t &\in \{\text{RESPOND}, \text{CLARIFY}\}. \end{aligned} \quad (16)$$

DEFER blocks persistent revision, whereas CLARIFY controls whether additional information is requested.

Finally,

$$y_t \sim \pi_{\theta}^{\text{resp}} (y \mid H_{t-1}, u_t, b_t, d_t). \quad (17)$$

3.4 Training

CORE is trained in two stages: supervised warm-up establishes the structured persona-update interface, while policy optimization improves how these decisions interact over long-horizon dialogue.

Supervised warm-up. The first stage teaches CORE to track persona state, extract turn-local evidence, control persistent revision, and generate an appropriate response. We optimize

$$\begin{aligned} \mathcal{L}_{\text{SFT}} &= \mathcal{L}_{\text{bel}} + \lambda_1 \mathcal{L}_{\text{evi}} + \lambda_2 \mathcal{L}_{\text{upd}} \\ &\quad + \lambda_3 \mathcal{L}_{\text{dis}} + \lambda_4 \mathcal{L}_{\text{resp}}. \end{aligned} \quad (18)$$

Here, $\mathcal{L}_{\text{bel}}$ supervises the maintained persona belief, $\mathcal{L}_{\text{evi}}$ supervises turn-local evidence including the null option, $\mathcal{L}_{\text{upd}}$ supervises COMMIT/DEFER/IGNORE routing, $\mathcal{L}_{\text{dis}}$ supervises RESPOND/CLARIFY, and $\mathcal{L}_{\text{resp}}$ is the autoregressive response loss.

Belief supervision is applied only to slot-turns with reliable normalized targets:

$$\mathcal{L}_{\text{bel}} = - \sum_{(t,k) \in \mathcal{S}_{\text{closed}}} \log b_t^k(z_{t,k}^*). \quad (19)$$

Residual open-world attributes are excluded from closed-slot belief supervision. The update and discourse policies are trained with standard cross-entropy objectives over their corresponding reference decisions.

Policy optimization. Supervised learning teaches the individual decisions, but does not directly optimize their long-horizon consequences. Starting from the supervised checkpoint, we therefore apply PPO to balance correct state revision, persona preservation, response quality, and clarification behavior. The task reward is

$$R_t = \lambda_{\text{act}} R_t^{\text{act}} + \lambda_{\text{state}} R_t^{\text{state}} + \lambda_{\text{ans}} R_t^{\text{ans}} + \lambda_{\text{clar}} R_t^{\text{clar}}. \quad (20)$$

We set the corresponding weights to 0.35, 0.30, 0.25, and 0.10 for action correctness, state fidelity, response quality, and clarification behavior, respectively.

The four terms reward correct update control, persona-state fidelity, personalized response quality, and appropriate clarification, respectively. Together, they encourage CORE to avoid unsupported persistent updates without becoming overly conservative toward genuine preference changes.

The PPO objective is regularized toward the supervised reference policy, with the KL penalty applied once in the policy objective. Reward definitions and optimization details are provided in Appendix B.5.

3.5 Why Update Control Matters

We provide a simple analysis of why uncontrolled persistent revision can accumulate error. This analysis is interpretive and does not assume access to the reference persona value at inference time.

Consider a dialogue segment over which the grounded reference value z_k^* for slot k remains unchanged. Fix an incorrect competing value $\bar{v}^k \neq z_k^*$, and define the belief margin

$$M_t^k(\bar{v}^k) = \log \frac{b_t^k(z_k^*)}{b_t^k(\bar{v}^k)}, \quad (21)$$

together with the evidence margin

$$X_t^k(\bar{v}^k) = \Delta_t^k(z_k^*) - \Delta_t^k(\bar{v}^k). \quad (22)$$

Using Eq. (15),

$$M_t^k(\bar{v}^k) = M_{t-1}^k(\bar{v}^k) + g_t^k X_t^k(\bar{v}^k), \quad (23)$$

and therefore

$$M_T^k(\bar{v}^k) = M_0^k(\bar{v}^k) + \sum_{t=1}^T g_t^k X_t^k(\bar{v}^k). \quad (24)$$

For a passive updater with $g_t^k \equiv 1$, suppose unsupported observations favor the competing value

with probability at least ρ_k , and their expected evidence margin is at most $-\gamma_k$. Then

$$\mathbb{E} [M_T^k(\bar{v}^k)] \leq M_0^k(\bar{v}^k) - \rho_k \gamma_k T. \quad (25)$$

Thus, repeated unsupported evidence can progressively erode a grounded belief even when the underlying preference remains unchanged. The opposite extreme—never revising the persona state—would fail when the user genuinely changes preferences. CORE therefore targets the resulting stability–adaptability trade-off by making persistent revision an explicit evidence-conditioned decision rather than an automatic consequence of every locally plausible observation.

4 Experiments

We evaluate CORE across standard alignment, persona-state robustness, mechanistic ablations, efficiency, and human assessment, with our analysis organized around four questions:

- **Q1: Performance.** Does CORE outperform baselines on personalized alignment?
- **Q2: Fidelity.** Does CORE maintain better persona state under ambiguity?
- **Q3: Robustness.** Does CORE improve robustness under ambiguous or conflicting evidence?
- **Q4: Mechanism.** Do ablations support update control as a contributing mechanism?

4.1 Experimental Setup

Benchmarks. We evaluate CORE on three complementary benchmarks. **ALOE** (Wu et al., 2025) is our primary benchmark for multi-turn preference consistency, where user preferences are revealed over interaction. Following the profile-conditioned interaction setup of RLPA (Zhao et al., 2025b), we use ALOE persona profiles to initialize simulated users and generate dialogue trajectories through an online user simulator (Appendix B.1). **PersonaChat** (Zhang et al., 2018) is used to assess generalization to held-out persona profiles. **PER-SIST** is our robustness benchmark for persona-state robustness under sequential interaction stress, including ambiguity, conflict, and controlled social influence after persona grounding. It is used only for evaluation.

Baselines. We compare against four groups of baselines. (1) Inference-time methods: REMINDER (Zhao et al., 2025a), SELF-CRITIC (Zhao et al., 2025a), CoT (Wei et al., 2022), and

RAG (Zhao et al., 2025a) with top-5 retrieval. These baselines test whether explicit reminders, reasoning traces, or retrieved memory are sufficient for robust personalization. (2) Offline alignment methods: SFT (Ouyang et al., 2022) and DPO (Rafailov et al., 2023), testing whether conventional preference optimization can preserve persona consistency over long interactions. (3) Memory-based and personalized-agent methods: MEMORYBANK (Zhong et al., 2024) and RLPA (Zhao et al., 2025b), representing persistent-memory and agent-style personalization. (4) Structure-matched controls match CORE’s intermediate format or supervision while removing executable belief revision, testing whether traces, extra labels, or action names alone explain the gains.

All same-backbone trainable methods use Llama-3.2-3B-Instruct (Grattafiori et al., 2024) with matched training budgets, decoding settings, and checkpoint-selection protocols.

Persona slots. Persona slots are instantiated per benchmark while keeping the same belief-tracking framework. New values can be added when observed, while attributes that cannot be reliably mapped to closed slots are assigned to residual open slots. Closed-slot normalization covers 91.4% of ALOE, 86.6% of PersonaChat, and 89.1% of PERSIST. Residual open-world attributes remain in response-level evaluation but are excluded from metrics requiring normalized state targets. Accordingly, closed-slot state metrics should be interpreted as state-fidelity diagnostics over the reliably normalized portion of each benchmark rather than as complete measures of open-world persona state. Details of slot construction and value normalization are in Appendix B.2.

Evaluator and metrics. We adopt a layered evaluation protocol across three distinct granularities, restricting LLM judges to response-level assessment while evaluating internal states and decision routing via reference annotations. At the response level, we quantify conversational alignment using the AL score, tracking its turn-wise improvement and stability via N -IR and N -R². State-level fidelity is captured by SLOTACC for argmax slot accuracy, BUA for correctness on update-positive turns, and DRIFT for the posterior mass assigned away from grounded values. For decision-level operations over {COMMIT, DEFER, IGNORE}, ACTACC measures exact three-way routing accuracy, while CRR measures conflict-level

non-commitment when an unsupported persistent write should be avoided. Within PERSIST, we additionally report ARUS for axiom recall under stress, BCD for resistance to blind compliance, and APR for proactive clarification under unresolved uncertainty.

Response-level alignment, PERSIST robustness, and human evaluation constitute our primary end-to-end evidence. State-level metrics provide complementary closed-slot fidelity diagnostics, whereas ACTACC, CRR, FCR, and related action-level measures are used only as mechanism diagnostics of the update-control process. For methods without native update actions, these diagnostics use fixed behavior-derived proxies: persistent adoption of new evidence is mapped to COMMIT, explicit unresolved handling or clarification to DEFER, and preservation of the prior commitment to IGNORE. These proxies are used only for diagnostic evaluation and should not be interpreted as native internal actions of the baseline systems.

We use GPT-5.1 as the judge model for metric computation. Full evaluation details are provided in Appendix C.

4.2 PERSIST: Stress Test

PERSIST evaluates whether personalized dialogue models can preserve or appropriately revise grounded persona commitments under ambiguity, conflict, and controlled social influence. Each dialogue contains three phases: Anchoring establishes grounded persona evidence, Fluctuation introduces partial or transient cues, and Confrontation applies six controlled social-influence strategies, including authority pressure, social conformity, affective appeal, urgency framing, reciprocity lure, and relationship pressure. This design first establishes persona evidence and then evaluates robustness to subsequent interaction stress, while remaining independent of CORE-specific internals: models are evaluated by whether they preserve grounded commitments, avoid unsupported persistent overwrites, and clarify unresolved uncertainty without requiring explicit update actions or structured beliefs.

We use PERSIST only as a held-out benchmark for post-anchor robustness. Because the benchmark begins after persona grounding, it does not by itself isolate pressure-first belief formation or explicit preference reversal. We therefore evaluate these complementary boundary conditions separately in Appendix E.3. Details of PERSIST are in Appendix D.

Model		Type	PersonaChat			ALOE		
			AL (%)	N-IR	N-R ²	AL (%)	N-IR	N-R ²
Peer baseline	Gemma-2-2B-IT	SFT	21.19	0.017	0.136	18.33	0.021	0.096
	Qwen2.5-3B-Instruct	SFT	26.84	0.049	0.159	44.32	0.083	0.162
Scale reference	Mistral-7B-v0.3	Base	52.98	0.058	0.364	74.32	0.074	0.284
	Llama-3.1-8B-Instruct	Base	64.87	0.062	0.392	77.26	0.090	0.407
Matched 3B methods	Llama-3.2-3B-Instruct	Base	42.02	0.025	0.097	21.99	0.044	0.121
		SFT	49.38	0.039	0.134	36.15	0.054	0.179
		Reminder	42.91	0.032	0.178	37.68	0.038	0.105
		Self-Critic	62.38	0.047	0.299	47.04	0.067	0.097
		DPO	51.28	0.052	0.241	64.32	0.073	0.112
		RAG (top-5)	53.09	0.048	0.211	32.17	0.071	0.067
		CoT	44.67	0.033	0.126	51.09	0.055	0.083
		MemoryBank	59.12	0.049	0.208	72.27	0.071	0.271
		RLPA	65.05	0.051	0.263	75.84	0.087	0.202
		CORE (Ours)	76.54	0.064	0.471	82.94	0.098	0.395

Table 1: **Overall personalized alignment.** CORE achieves the strongest alignment among matched same-backbone methods on PersonaChat and ALOE. Trainable methods are evaluated over three random seeds; uncertainty reporting and paired dialogue-level bootstrap tests follow the protocol in Appendix C.

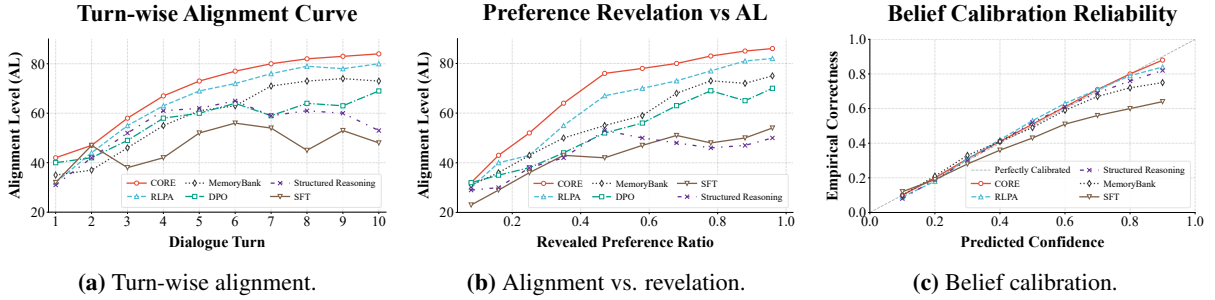


Figure 3: Trajectory-level diagnostics on ALOE. CORE yields stronger turn-wise alignment, uses revealed preference evidence more efficiently, and maintains better-calibrated persona beliefs.

4.3 Main Results on Standard Benchmarks

Table 1 reports the main results on PersonaChat and ALOE. CORE achieves the highest alignment scores among matched same-backbone methods on both benchmarks and remains competitive with larger scale-reference models on trajectory-stability metrics. The gains over SFT/DPO suggest that static preference optimization alone is insufficient for preserving persona consistency over long interactions, while the gains over RAG/MemoryBank indicate that improved access to persona evidence does not by itself determine whether later observations should revise the maintained user state. Together, these results support controlled persona-state revision as a useful mechanism beyond stronger response generation or more persistent memory alone.

Figure 3 provides trajectory-level diagnostics on ALOE. CORE yields higher and more stable turn-wise alignment, indicating that reliable persona evidence is incorporated without being overwritten by later under-resolved observations. It also

achieves stronger alignment at the same preference-revelation ratio, showing more efficient use of limited personalization signals. The calibration curve further shows that CORE’s belief confidence better matches empirical correctness, suggesting that its response-level gains are accompanied by a more reliable maintained persona state rather than surface-level generation alone.

4.4 Robustness under State Perturbations

We next evaluate whether models preserve grounded persona commitments when later observations are insufficiently grounded for persistent revision. Table 2 reports aggregate PERSIST robustness. CORE improves ARUS from 71.9 to 80.8 and BCD from 66.1 to 76.4, while reducing persona-state drift from 12.6 to 10.1. These gains show that CORE not only produces better local responses, but also better maintains the persona state when later observations remain insufficiently grounded for persistent revision. Figure 4a visualizes the learned routing regions over ambiguity

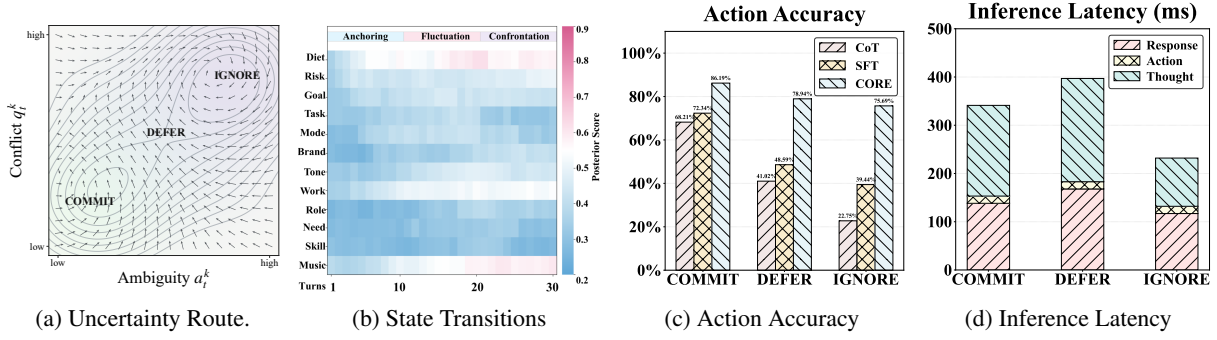


Figure 4: **Comprehensive CORE Evaluation.**

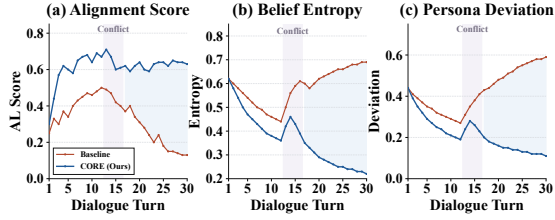


Figure 5: **Conflict dynamics.** CORE preserves alignment and reduces belief drift under introduced conflict.

Method	ARUS $\uparrow$	BCD $\uparrow$	Drift $\downarrow$
SFT	57.8 $\pm$.8	48.6 $\pm$.9	18.6 $\pm$.5
DPO	61.5 $\pm$.7	54.2 $\pm$.8	16.4 $\pm$.5
MemBank	68.4 $\pm$.7	62.8 $\pm$.7	13.8 $\pm$.4
RLPA	71.9 $\pm$.6	66.1 $\pm$.7	12.6 $\pm$.4
CORE	80.8$\pm$.5	76.4$\pm$.6	10.1$\pm$.4

Table 2: **PERSIST stress robustness.** CORE better preserves grounded persona commitments and reduces persistent drift under ambiguity, conflict, and controlled social-influence stress.

and conflict: well-resolved evidence is more often committed, while high-ambiguity or unresolved conflicting evidence is more often deferred or ignored.

The improvements are largest in the regimes where passive updating is most brittle. On ambiguity-positive turns, CORE increases APR, suggesting that relevant but under-resolved evidence is more often clarified rather than prematurely committed.

On conflict-positive turns, CORE reduces the false commit rate (FCR) by 9.3 points relative to RLPA, showing that unresolved contradictory evidence is less likely to be written into the persona state. Across PERSIST interaction-stress strategies, CORE more often avoids committing unresolved conflicting evidence, thereby preserving previously grounded preferences.

Figure 4b diagnoses persona-state transitions. Rather than freezing the persona state, CORE re-

vives the belief under sufficiently grounded preference changes while preserving the anchored state under ambiguous or unresolved conflicting evidence. This suggests that the robustness gain is not explained by complete state inertia. Figure 4c provides a mechanism-level diagnostic: the stronger response- and state-level robustness of CORE is accompanied by more accurate update routing rather than being established by action-level accuracy alone. Figure 5 further shows that after conflict is introduced, the strongest baseline loses alignment while belief entropy and persona deviation increase; CORE keeps alignment high and suppresses both uncertainty and deviation from the anchored persona. The largest improvements occur on ambiguity-positive and conflict-positive turns, where uncontrolled updating is most brittle.

Boundary diagnostics. Because PERSIST is post-anchor by design, we separately evaluate first-contact initialization and explicit preference reversal. Under pressure-first initialization, CORE reduces FCR from 31.6% for RLPA to 20.9%, although this remains higher than CORE’s 11.6% FCR after anchoring, confirming that a grounded reference belief strengthens conflict-aware preservation. In preference-reversal cases, explicit and persistent changes can be committed immediately, whereas tentative changes are deferred until confirmation, showing that conflict induces caution rather than permanent rejection. Full posterior trajectories and turns to stable MAP are reported in Appendix E.3.

4.5 Mechanistic Ablations and Efficiency

Ablation Study. Table 3 isolates the contribution of CORE. Removing executable update actions causes the largest degradation: AL drops from 82.9 to 78.8, CRR from 82.7 to 71.6, and DRIFT increases from 10.1 to 15.4. This pattern is consistent with executable routing contributing beyond turn-

Variant	AL $\uparrow$	CRR $\uparrow$	Drift $\downarrow$
CORE	82.9$\pm$.4	82.7$\pm$.4	10.1$\pm$.4
w/o update act.	78.8 $\pm$.5	71.6 $\pm$.6	15.4 $\pm$.5
w/o clarify	80.1 $\pm$.5	77.8 $\pm$.5	13.9 $\pm$.4
w/o inertia	80.7 $\pm$.4	78.6 $\pm$.5	13.4 $\pm$.4
w/o opt. rev.	80.4 $\pm$.5	78.1 $\pm$.5	13.7 $\pm$.5
w/o cal. rew.	80.6 $\pm$.4	76.4 $\pm$.6	13.5 $\pm$.4
w/o rev. rew.	79.5 $\pm$.5	75.2 $\pm$.6	14.7 $\pm$.5

Table 3: **Mechanistic ablations.** Removing components weakens alignment and stability.

local evidence extraction alone. Other ablations weaken the same control pathway in different ways: removing clarification mishandles under-resolved evidence, removing inertia or optimized revision makes the posterior move too aggressively, and removing calibration or revision rewards reduces conflict-sensitive routing.

Failure-mode decomposition makes these effects more concrete: removing executable update actions increases unsupported COMMIT from 11.3% to 22.9%, while removing clarification raises under-clarification from 18.4% to 42.7%. Structure-matched controls provide a complementary test: action prediction without executable belief revision reaches 80.0 AL and 72.1 CRR, compared with 82.9 AL and 82.7 CRR for CORE. Together, these results support executable belief revision as a contributing mechanism beyond intermediate traces, action-name supervision, or additional structured labels alone. Full controls and failure-mode breakdowns are reported in Appendix E.

Inference cost. CORE introduces a lightweight pre-generation control layer. Figure 4d decomposes per-turn latency and shows that response decoding remains the dominant cost, while update routing and belief revision add only a small action-level overhead across COMMIT, DEFER, and IGNORE. The dialogue context is encoded once, slot-conditioned evidence logits are produced in a shared pass, and cached slot/value descriptions avoid separate LLM calls for each slot or value. This yields $O(KV_{\max}d)$ additional computation and adds only 23.0 ms per turn on the same backbone, suggesting that CORE improves robustness without materially slowing inference. Gold persona-state and update-action annotations are used only for training or evaluation and are not required at inference time. More details are given in Appendix E.7.

Human evaluation. Human-in-the-loop evaluation further examines evaluation reliability and end-to-end interaction quality. Across audited response, robustness, and pairwise items, human agreement is strong, and judge predictions are well aligned with human judgments (mean J-H = 0.84, mean κ = 0.79), supporting the reliability of response-level automatic evaluation. In complete multi-turn interactions, CORE improves the average human rating over RLPA by 11.7%. The gains are concentrated on persona consistency, ambiguity handling, preference preservation, and long-horizon stability, which improve by +11.6%–15.5%, while response quality improves by only +4.7%. This indicates that CORE’s advantage is concentrated more strongly on long-horizon persona-state behavior than on generic response fluency. This controlled evaluation provides behavioral validation under matched multi-turn scenarios rather than evidence from natural longitudinal deployment. Details are in Appendix C.

5 Conclusion

We presented CORE, which formulates long-horizon personalization as controlled persona-state revision. By separating turn-local evidence from persistent belief updates, CORE preserves grounded preferences under unresolved observations while remaining adaptive to sufficiently supported changes. We also introduced PERSIST, a held-out benchmark for post-anchor persona-state robustness under sequential interaction stress. Across standard benchmarks, stress tests, and human evaluation, our results support explicit update control as a useful mechanism for robust long-horizon personalization.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China (62476070), the Major Key Project of PCL under Grant PCL2025A10 and PCL2024A06, Shenzhen Science and Technology Program (RCJC20231211085918010, JCYJ20241202123503005, GXWD20231128103232001, ZDSYS20230626091203008, KQTD20240729102154066), Department of Science and Technology of Guangdong (2024A1515011540).

Limitations

CORE relies on benchmark-instantiated persona slots and structured supervision, so its state-level analysis does not fully capture open-ended or cross-slot preferences. Its conservative update policy may also delay adaptation when genuine preference changes are weakly expressed. Moreover, PERSIST primarily evaluates post-anchor robustness rather than first-contact personalization, and controlled benchmarks cannot fully represent long-term real-world interactions.

Ethical Considerations

Persistent personalization raises privacy and user-control concerns because maintained persona states can influence future interactions. PERSIST is used only for controlled evaluation and is not intended for profiling real users or inferring sensitive attributes. Real-world systems should minimize retained information and allow users to inspect, correct, delete, or opt out of persistent persona states.

References

- Jiawei Chen, Xinyan Guan, Qianhao Yuan, Guozhao Mo, Weixiang Zhou, Yaojie Lu, Hongyu Lin, Ben He, Le Sun, and Xianpei Han. 2025a. ConsistentChat: Building skeleton-guided consistent multi-turn dialogues for large language models from scratch. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8415–8441.
- Ruizhe Chen, Xiaotian Zhang, Meng Luo, Wenhao Chai, and Zuozhu Liu. 2025b. Pad: Personalized alignment of llms at decoding-time. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*.
- Changwoo Chun, Daniel Rim, and Juhee Park. 2025. LLM ContextBridge: A hybrid approach for intent and dialogue understanding in IVSR. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 794–806.
- Sarkar Snigdha Sarathi Das, Chirag Shah, Mengting Wan, Jennifer Neville, Longqi Yang, Reid Andersen, Georg Buscher, and Tara Safavi. 2024. S3-DST: Structured open-domain dialogue segmentation and state tracking in the era of LLMs. In *Findings of the Association for Computational Linguistics (ACL)*, pages 14996–15014.
- Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Model alignment as prospect theoretic optimization. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pages 12634–12651.
- Yujie Feng, Zexin Lu, Bo Liu, Liming Zhan, and Xiao-Ming Wu. 2023. Towards LLM-driven dialogue state tracking. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 739–755. Association for Computational Linguistics.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Jian Guan, Junfei Wu, Jia-Nan Li, Chuanqi Cheng, and Wei Wu. 2025. A survey on personalized Alignment—The missing piece for large language models in real-world applications. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5313–5333.
- Jiaming Ji, Boyuan Chen, Hantao Lou, Donghai Hong, Borong Zhang, Xuehai Pan, Juntao Dai, Tianyi Qiu, and Yaodong Yang. 2024. Aligner: Efficient alignment by learning to correct. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.
- Cai Ke, Yiming Du, Bin Liang, Yifan Xiang, Lin Gui, Zhongyang Li, Baojun Wang, Yue Yu, Hui Wang, Kam-Fai Wong, and Ruifeng Xu. 2025. Flexibly utilize memory for long-term conversation via a fragment-then-compose framework. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 21119–21136.
- Kasia Kobalcyk and Mihaela van der Schaar. 2025. Preference learning for AI alignment: a causal perspective. In *Proceedings of the Forty-second International Conference on Machine Learning (ICML)*, volume 267.
- Hao Li, Chenghao Yang, An Zhang, Yang Deng, Xiang Wang, and Tat-Seng Chua. 2025. Hello again! LLM-powered personalized agent for long-term dialogue. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, pages 5259–5276.
- Zekun Li, Zhiyu Zoey Chen, Mike Ross, Patrick Huber, Seungwhan Moon, Zhaojiang Lin, Luna Dong, Adithya Sagar, Xifeng Yan, and Paul A. Crook. 2024. Large language models as zero-shot dialogue state tracker through function calling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8688–8704.
- Jiahong Liu, Zexuan Qiu, Zhongyang Li, Quanyu Dai, Wenhao Yu, Jieming Zhu, Minda Hu, Menglin Yang, Tat-Seng Chua, and Irwin King. 2025. A survey of personalized large language models: Progress and future directions. *arXiv preprint arXiv:2502.11528*.

- Wenhao Liu, Xiaohua Wang, Muling Wu, Tianlong Li, Changze Lv, Zixuan Ling, Zhu JianHao, Cenyuan Zhang, Xiaoqing Zheng, and Xuanjing Huang. 2024. Aligning large language models with human preferences through representation engineering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 10619–10638.
- Hantao Lou, Jiaming Ji, Kaile Wang, and Yaodong Yang. 2025. Stream aligner: Efficient sentence-level alignment via distribution induction. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI)*, pages 27500–27508.
- Chimaobi Okite, Naihao Deng, Kiran Bodipati, Huaidian Hou, Joyce Chai, and Rada Mihalcea. 2025. Benchmarking and improving LLM robustness for personalized generation. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 16040–16072.
- Kai Tzu-iunn Ong, Namyoung Kim, Minju Gwak, Hyungjoo Chae, Taeyoon Kwon, Yohan Jo, Seungwon Hwang, Dongha Lee, and Jinyoung Yeo. 2025. Towards lifelong dialogue agents via timeline-based memory management. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, pages 8631–8661.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.
- Bo Peng, Zhiheng Wang, Heyang Gong, and Chaochao Lu. 2025. IP-dialog: Evaluating implicit personalization in dialogue systems with synthetic data. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 17007–17040.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.
- Chris Richardson, Yao Zhang, Kellen Gillespie, Sudipta Kar, Arshdeep Singh, Zeynab Raeesy, Omar Zia Khan, and Abhinav Sethy. 2023. Integrating summarization and retrieval for enhanced personalization via large language models. *arXiv preprint arXiv:2310.20081*.
- Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. Lamp: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7370–7392.
- Omar Shaikh, Hussein Mozannar, Gagan Bansal, Adam Fourney, and Eric Horvitz. 2025. Navigating rifts in human-LLM grounding: Study and benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 20832–20847.
- Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, Zifeng Wang, Long Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, Anand Rajan Iyer, Tianlong Chen, Huan Liu, Chen-Yu Lee, and Tomas Pfister. 2025. In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8416–8439.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.
- Shujin Wu, Yi R. Fung, Cheng Qian, Jeonghwan Kim, Dilek Hakkani-Tur, and Heng Ji. 2025. Aligning llms with individual preferences via interaction. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, pages 7648–7662.
- Xin Yu, Hanwen Xing, and Lingzhou Xue. 2026. Exact: Explicit attribute-guided decoding-time personalization. *arXiv preprint arXiv:2602.17695*.
- Michael JQ Zhang and Eunsol Choi. 2025. Clarify when necessary: Resolving ambiguity through interaction with LMs. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5541–5558.
- Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2204–2213.
- Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. 2025a. Do llms recognize your preferences? evaluating personalized preference following in llms. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*.
- Weixiang Zhao, Xingyu Sui, Yulin Hu, Jiahe Guo, Haixiao Liu, Biye Li, Yanyan Zhao, Bing Qin, and Ting Liu. 2025b. Teaching language models to evolve with users: Dynamic profile modeling for personalized alignment. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.

Xiaoyan Zhao, Juntao You, Yang Zhang, Wenjie Wang, Hong Cheng, Fuli Feng, See-Kiong Ng, and Tat-Seng Chua. 2026. Nextquill: Causal preference modeling for enhancing llm personalization. In *Proceedings of the International Conference on Learning Representations(ICLR)*.

Wanjuan Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI)*, pages 19724–19731.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. 2025. Personality alignment of large language models. In *Proceedings of the Thirteenth International Conference on Learning Representations(ICLR)*.

Appendix

A CORE Method Details	14	E Additional Results and Mechanistic Analyses	28
A.1 Slot-Factorized Persona Belief . . .	14	E.1 Structure-Matched Controls	28
A.2 Semi-Open Value Handling	14	E.2 Component Ablations by Failure Mode	28
A.3 Context Encoding and Slot-Wise Evidence Scoring	14	E.3 Stability–Adaptability Boundary Diagnostics	28
A.4 Uncertainty-Guided Update Routing	15	E.4 Scale and Backbone Transfer . . .	29
A.5 Belief Revision	16	E.5 Closed-Slot State Diagnostics . . .	29
A.6 Inference-Time Algorithm	16	E.6 PERSIST Robustness Breakdowns	29
B Experimental and Training Protocol	16	E.7 Latency, Memory, and Retrieval Controls	29
B.1 Benchmarks, Splits, and Evaluation Scope	17		
B.2 Slot Instantiation and Baseline Controls	17		
B.3 Update-Action Labels and Audit .	18		
B.4 Supervised Warm-Up and Uncertainty Calibration	19		
B.5 PPO Configuration and Stability Checks	19		
C Evaluation Metrics, Reliability, and Human Study	20		
C.1 Metric-Source Separation	20		
C.2 Response-Level Metrics	20		
C.3 Belief-Level Metrics	21		
C.4 Decision-Level Metrics	21		
C.5 PERSIST Robustness Metrics . .	22		
C.6 Aggregation and Uncertainty Reporting	22		
C.7 Judge Prompts, Calibration, and Reliability	22		
C.8 Anti-Gaming Diagnostics	23		
C.9 Human Evaluation Protocol	23		
C.10 Human Rating Rubric and Statistical Analysis	23		
C.11 Human Evaluation Results	24		
C.12 Pairwise Preference and Scenario Analysis	24		
D PERSIST: A Benchmark for Persona-State Robustness under Sequential Interaction Stress	24		
D.1 Benchmark Objective and Target Behaviors	24		
D.2 Dataset Statistics and Splits	25		
D.3 Construction and Verification Pipeline	26		
D.4 Social-Influence Strategy Taxonomy	26		
D.5 Quality Assurance and Human Audit	26		
D.6 Evaluation Protocol	27		
		D.7 Limitations of PERSIST	27

Appendix organization. The appendix provides details needed to audit and reproduce CORE. Appendix A specifies the method implementation; Appendix B reports experimental controls, training setup, calibration, and PPO configuration; Appendix C defines the evaluation metrics and reports judge and human-study reliability; Appendix D describes PERSIST construction, stress strategies, and quality control; and Appendix E provides structure-matched controls, boundary diagnostics, robustness breakdowns, efficiency analyses, and failure-mode results.

A CORE Method Details

This section gives implementation details.

A.1 Slot-Factorized Persona Belief

CORE maintains a slot-factorized persona belief. Let $\mathcal{K} = \{1, \dots, K\}$ denote the benchmark-instantiated slot inventory.

At turn t , slot k has a working value set $\mathcal{V}_k^{(t)}$, and CORE stores a categorical posterior $b_t^k \in \mathcal{S}_k^{(t)}$, where $\mathcal{S}_k^{(t)}$ denotes the probability simplex over $\mathcal{V}_k^{(t)}$.

The joint belief is approximated as

$$b_t(z) \approx \prod_{k=1}^K b_t^k(z^k), \quad z^k \in \mathcal{V}_k^{(t)}. \quad (26)$$

This factorization is an auditable state interface, not an independence assumption. It lets each turn-level update be traced to a concrete persona slot.

For each slot, we summarize the posterior by its MAP value and anchored confidence. Let $\mathcal{V}_{k,+}^{(t)} = \mathcal{V}_k^{(t)} \setminus \{\text{UNK}\}$ and $m_t^k = 1 - b_t^k(\text{UNK})$. For $m_t^k > 0$, define $\bar{b}_t^k(v) = b_t^k(v)/m_t^k$ over $\mathcal{V}_{k,+}^{(t)}$. Then

$$\hat{z}_t^k = \arg \max_{v \in \mathcal{V}_{k,+}^{(t)}} b_t^k(v), \quad (27)$$

$$c_t^k = m_t^k \begin{cases} 1 - \frac{H(\bar{b}_t^k)}{\log |\mathcal{V}_{k,+}^{(t)}|}, & |\mathcal{V}_{k,+}^{(t)}| > 1, \\ 1, & |\mathcal{V}_{k,+}^{(t)}| = 1, \\ 0, & |\mathcal{V}_{k,+}^{(t)}| = 0. \end{cases} \quad (28)$$

This prevents a posterior concentrated on UNK from being interpreted as strong persona grounding.

A.2 Semi-Open Value Handling

The slot inventory is fixed within each benchmark, but value sets are semi-open:

$$\mathcal{V}_k^{(t)} = \mathcal{V}_k^{\text{canon}} \cup \mathcal{V}_k^{\text{obs}}(t) \cup \{\text{UNK}\}. \quad (29)$$

Here, $\mathcal{V}_k^{\text{canon}}$ contains benchmark-defined canonical values, $\mathcal{V}_k^{\text{obs}}(t)$ contains values observed or clarified up to turn t , and UNK marks unresolved evidence. A new value is added only when it can be mapped to an existing slot with sufficient semantic confidence. Attributes without a reliable closed-slot mapping are assigned to a residual open slot and excluded from closed-slot state metrics.

Surface mentions are normalized before entering $\mathcal{V}_k^{\text{obs}}(t)$. Paraphrases are merged only when they express the same persona commitment. For example, *prefers concise answers* and *likes brief explanations* may be merged, while *vegetarian* and *vegan* remain distinct unless the dialogue explicitly resolves them. The same normalization rules are applied to all methods.

When a new value is admitted into $\mathcal{V}_k^{(t)}$, the previous posterior b_{t-1}^k is extended with a configuration-fixed prior mass and then renormalized over the expanded support. The semantic mapping used to determine whether a newly observed value belongs to an existing slot is part of benchmark/schema normalization rather than a separate inference-time slot-matching model. Closed-slot belief metrics are computed only on normalized slots and values. Open-world attributes can still affect response-level evaluation, but they are not counted as closed-slot belief correctness.

A.3 Context Encoding and Slot-Wise Evidence Scoring

At turn t , CORE receives dialogue history H_{t-1} , user utterance u_t , and previous belief b_{t-1} . A shared context encoder produces

$$h_t = f_{\text{enc}}(H_{t-1}, u_t, b_{t-1}). \quad (30)$$

The maintained belief is provided as a compact symbolic state summary, including high-confidence slot values, unresolved UNK states, and historical state-control information available before the current routing decision. No current-turn update action is included in h_t , because routing is computed only after turn-local evidence and uncertainty descriptors have

been obtained. This preserves the inference order context encoding $\rightarrow$ evidence scoring $\rightarrow$ uncertainty characterization $\rightarrow$ update routing.

CORE does not use a separate inference-time slot matcher. Each slot is represented by its schema, including a slot name, a short natural-language definition, and the current candidate values in $\mathcal{V}_k^{(t)}$. Slot grounding during inference is handled by slot-conditioned evidence scoring, while value-to-slot normalization used to construct benchmark targets is performed separately during preprocessing.

For each slot, we augment the working value set with a null option:

$$\bar{\mathcal{V}}_k^{(t)} = \mathcal{V}_k^{(t)} \cup \{\perp\}, \quad (31)$$

where $\perp$ means that the current turn provides no usable evidence for slot k .

For each $v \in \bar{\mathcal{V}}_k^{(t)}$, CORE computes a local evidence logit:

$$\ell_t^k(v) = f_{\text{evi}}(h_t, e_k, e_v). \quad (32)$$

The logit measures what the current turn locally suggests; it does not decide whether the maintained belief should be revised.

For numerical stability, evidence logits are centered and clipped within each slot:

$$\bar{\ell}_t^k(v) = \ell_t^k(v) - \frac{1}{|\bar{\mathcal{V}}_k^{(t)}|} \sum_{v' \in \bar{\mathcal{V}}_k^{(t)}} \ell_t^k(v'), \quad (33)$$

$$\Delta_t^k(v) = \text{clip} \left(\frac{\bar{\ell}_t^k(v)}{s_t^k + \epsilon}, -\Delta_{\max}, \Delta_{\max} \right), \quad (34)$$

where s_t^k is the within-slot scale used by the implementation and $\Delta_{\max}$ is fixed before evaluation. Both quantities are shared across experiments rather than selected from test performance. For candidate sets for which scale estimation is unreliable, the implementation applies clipping without the scale-normalization step. In the rest of the appendix, $\Delta_t^k(v)$ denotes this stabilized evidence score, matching the evidence notation in the main text.

The stabilized scores define a turn-local distribution over the augmented value space:

$$\tilde{p}_t^k(v) = \frac{\exp(\Delta_t^k(v))}{\sum_{v' \in \bar{\mathcal{V}}_k^{(t)}} \exp(\Delta_t^k(v'))}, \quad v \in \bar{\mathcal{V}}_k^{(t)}. \quad (35)$$

This distribution is turn-local. It identifies which value, if any, is supported by the current observation; routing decides whether that evidence should be committed, deferred, or ignored.

A.4 Uncertainty-Guided Update Routing

CORE derives relevance, ambiguity, and conflict from the turn-local evidence distribution. The null option gives the relevance estimate:

$$r_t^k = 1 - \tilde{p}_t^k(\perp). \quad (36)$$

For non-null values, we renormalize the evidence distribution over $\mathcal{V}_k^{(t)}$:

$$\tilde{p}_{t,+}^k(v) = \frac{\tilde{p}_t^k(v)}{\sum_{v' \in \mathcal{V}_k^{(t)}} \tilde{p}_t^k(v')}, \quad v \in \mathcal{V}_k^{(t)}. \quad (37)$$

Ambiguity is computed as

$$a_t^k = \begin{cases} \frac{H(\tilde{p}_{t,+}^k)}{\log |\mathcal{V}_k^{(t)}|}, & |\mathcal{V}_k^{(t)}| > 1, \\ 0, & |\mathcal{V}_k^{(t)}| \leq 1. \end{cases} \quad (38)$$

When a new value expands $\mathcal{V}_k^{(t)}$, we first extend b_{t-1}^k with the same configuration-fixed prior mass used in Appendix A.2 and renormalize it over the aligned support. Conflict is then

$$q_t^k = r_t^k D_{\text{JS}} \left(\tilde{p}_{t,+}^k \parallel b_{t-1}^k \right). \quad (39)$$

The multiplier r_t^k prevents weakly matched turns from being treated as strong conflicts. When r_t^k falls below the calibrated relevance threshold, slot k is treated as unmatched and bypasses persistent revision.

The update policy directly conditions on the current evidence distribution, previous belief, and uncertainty descriptors:

$$\alpha_t^k \sim \pi_{\text{upd}}^\theta \left(\alpha \mid \tilde{p}_t^k, b_{t-1}^k, r_t^k, a_t^k, q_t^k, c_{t-1}^k \right), \quad (40)$$

$$\alpha_t^k \in \{\text{COMMIT}, \text{DEFER}, \text{IGNORE}\}.$$

The implementation uses the corresponding slot-conditioned encoded representation, avoiding a literal flattening of variable-length candidate distributions while preserving the routing semantics in the main text.

The action semantics are:

- **COMMIT**: authorize sufficiently grounded evidence to revise the maintained belief.
- **DEFER**: preserve the current belief while exposing relevant but under-resolved evidence to the current-turn discourse policy.
- **IGNORE**: exclude unrelated or explicitly non-persistent evidence from the active update pathway.

DEFER does not create a separate durable pending-memory buffer. Deferred evidence remains current-turn control information and does not automatically become a future COMMIT. If related evidence or clarification later appears while the earlier dialogue context remains accessible, the slot is inferred and routed again from the newly available context. DEFER is an internal update action, whereas CLARIFY is a dialogue-level decision made after slot-level revision.

A.5 Belief Revision

After routing, CORE converts each update decision into a continuous gate. Following Section 3.3, the gate feature is

$$\psi_t^k = [r_t^k; 1 - a_t^k; 1 - q_t^k; c_{t-1}^k]. \quad (41)$$

The revision gate is

$$g_t^k = \mathbb{I}[\alpha_t^k = \text{COMMIT}] \cdot \sigma(w_g^\top \psi_t^k + b_g). \quad (42)$$

DEFER and IGNORE set $g_t^k = 0$. Under COMMIT, the update magnitude is conditioned on relevance, ambiguity, conflict, and prior anchored confidence; we do not assume a fixed monotonic effect for any individual descriptor unless imposed by the implementation.

Given the gate, CORE performs a KL-regularized posterior update over the non-null value set $\mathcal{V}_k^{(t)}$:

$$b_t^k = \arg \min_{p \in \Delta(\mathcal{V}_k^{(t)})} \left[\text{KL}(p \| b_{t-1}^k) - g_t^k \sum_{v \in \mathcal{V}_k^{(t)}} p(v) \Delta_t^k(v) \right]. \quad (43)$$

The null option $\perp$ is used only for evidence relevance and is never written into the persona belief. The solution is

$$b_t^k(v) = \frac{b_{t-1}^k(v) \exp(g_t^k \Delta_t^k(v))}{\sum_{v' \in \mathcal{V}_k^{(t)}} b_{t-1}^k(v') \exp(g_t^k \Delta_t^k(v'))}, \quad v \in \mathcal{V}_k^{(t)}. \quad (44)$$

Thus $g_t^k = 0$ preserves the prior, while larger g_t^k moves the posterior toward evidence-supported values. For DEFER and IGNORE, the posterior remains unchanged, and deferred evidence is used only for current-turn discourse control.

After slot-level revision, CORE decides whether

Algorithm 1 CORE inference-time algorithm

Require: History H_{t-1} , user utterance u_t , previous belief b_{t-1}

Ensure: Updated belief b_t and response y_t

```

1:  $h_t \leftarrow f_{\text{enc}}(H_{t-1}, u_t, b_{t-1})$ 
2:  $\mathcal{A}_{\text{upd}} \leftarrow \{\text{COMMIT}, \text{DEFER}, \text{IGNORE}\}$ 
3:  $\mathcal{A}_{\text{dis}} \leftarrow \{\text{RESPOND}, \text{CLARIFY}\}$ 
4: for  $k = 1, \dots, K$  do
5:   Construct  $\mathcal{V}_k^{(t)}$ 
6:    $\bar{\mathcal{V}}_k^{(t)} \leftarrow \mathcal{V}_k^{(t)} \cup \{\perp\}$ 
7:   Score  $\ell_t^k(v)$  for all  $v \in \bar{\mathcal{V}}_k^{(t)}$ 
8:   Stabilize scores to obtain  $\Delta_t^k(v)$ 
9:   Compute  $\tilde{p}_t^k$  over  $\bar{\mathcal{V}}_k^{(t)}$ 
10:   $r_t^k \leftarrow 1 - \tilde{p}_t^k(\perp)$ 
11:  if  $r_t^k$  is below the relevance threshold then
12:     $\alpha_t^k \leftarrow \text{IGNORE}$ 
13:     $b_t^k \leftarrow b_{t-1}^k$ 
14:  else
15:    Compute  $\tilde{p}_{t,+}^k$  over  $\mathcal{V}_k^{(t)}$ 
16:    Compute  $a_t^k$  and  $q_t^k$  using Eq. (39)
17:    Predict  $\alpha_t^k$  using Eq. (40)
18:    if  $\alpha_t^k = \text{COMMIT}$  then
19:      Compute  $g_t^k$  using Eq. (42)
20:      Update  $b_t^k$  using Eq. (44)
21:    else
22:       $b_t^k \leftarrow b_{t-1}^k$ 
23:    end if
24:  end if
25: end for
26: Predict  $d_t$  using Eq. (45)
27:  $y_t \sim \pi_{\text{resp}}^\theta(\cdot | H_{t-1}, u_t, b_t, d_t)$ 
28: return  $b_t, y_t$ 

```

to respond or request clarification:

$$d_t \sim \pi_{\text{dis}}^\theta \left(d | u_t, H_{t-1}, b_t, \{r_t^k, a_t^k, q_t^k, \alpha_t^k\}_{k=1}^K \right), \quad d_t \in \{\text{RESPOND}, \text{CLARIFY}\}. \quad (45)$$

A deferred observation does not automatically trigger clarification.

A.6 Inference-Time Algorithm

Algorithm 1 summarizes inference. Evidence is scored before routing; only evidence routed through COMMIT can revise the maintained belief.

Gradient boundaries. During supervised warm-up, gradients are applied to the evidence scorer, update router, discourse policy, and response policy through their corresponding supervised losses. The belief update in Eq. (44) remains differentiable with respect to the evidence scores and revision gate, while the discrete routing action is optimized through the routing loss and policy objective.

B Experimental and Training Protocol

This section specifies the experimental and training protocol behind Section 4. It covers benchmark

splits, slot instantiation, baseline controls, update-action labels, label auditing, supervised warm-up, uncertainty calibration, PPO rewards, and stability checks. The goal is to make the comparison auditable and to test whether the observed gains can be distinguished from alternative explanations such as hidden supervision, larger context budgets, reward artifacts, or degenerate conservatism.

B.1 Benchmarks, Splits, and Evaluation Scope

All methods are evaluated in the same partially observed interaction setting. The model observes the dialogue history and current user utterance, but never observes the latent persona state directly. It must infer, maintain, and update personalization from turn-level evidence.

We use three complementary benchmarks. ALOE evaluates progressive preference revelation; PersonaChat evaluates generalization to free-form persona attributes; and PERSIST evaluates robustness under ambiguity, conflict, and controlled social influence. For ALOE and PersonaChat, splits are constructed at the persona level whenever profile identifiers are available. For datasets without explicit identifiers, we cluster normalized persona descriptors and split clusters rather than individual turns.

Table 4 reports the processed splits for ALOE and PersonaChat. The table describes dataset partitions; the PersonaChat training partition is not used for supervised warm-up. The calibration split is used for checkpoint selection, threshold calibration, response-judge calibration, and uncertainty calibration when applicable. PERSIST is used only for held-out robustness evaluation; its benchmark construction and evaluation protocol are described separately in Appendix D.

ALOE interaction setup. We obtain persona profiles from ALOE and use them to condition an online user simulator. Dialogue trajectories are generated through interaction with the simulator, which produces user utterances during the dialogue. Throughout this paper, ALOE denotes this profile-based interactive setting. The dialogue partitions and statistics in Table 4, as well as the ALOE results reported in the main text and appendix, refer to the simulator-generated dialogues. The underlying profiles are sourced from ALOE; the resulting dialogue trajectories are generated for our experiments.

Table 4: **Processed experimental splits.** Partitions are fixed before training and shared by all methods. ALOE counts describe simulator-generated dialogues. Train lists the dataset training partition, not necessarily the data used for warm-up; the PersonaChat training partition is excluded from supervised warm-up.

Benchmark	Train	Cal.	Test	Turns	Split guarantee
ALOE	4,800	600	600	61,248	preference-cluster disjoint
PersonaChat	8,448	1,024	1,024	103,872	persona-profile disjoint

Response-level metrics are computed on all retained examples. Closed-slot belief metrics are computed only when the relevant persona evidence can be reliably mapped to a normalized slot-value target. Residual open-world attributes remain in response-level evaluation but are excluded from closed-slot state metrics.

B.2 Slot Instantiation and Baseline Controls

CORE uses the same state-control interface across benchmarks, while the slot inventory is benchmark-instantiated. This avoids imposing a universal ontology while keeping evaluation auditable. ALOE, PersonaChat, and PERSIST instantiate 10, 8, and 12 closed slots, respectively. Their closed-slot coverage after normalization is 91.4%, 86.6%, and 89.1%; the remaining attributes are assigned to residual open slots and excluded only from closed-slot state metrics. Typical slots cover explanation style, privacy, budget, lifestyle, diet, verbosity, formality, travel, and risk tolerance.

Surface mentions are normalized before evaluation. Paraphrases are merged only when they express the same stable commitment. For example, *prefers concise answers* and *likes brief explanations* may share a verbosity value, while *vegetarian* and *vegan* remain distinct unless explicitly clarified. The same normalization rules are applied to CORE and all baselines.

For systems that do not expose a native structured persona belief, state-level diagnostics are computed with the same frozen evaluation-time state extractor, applied uniformly across those systems. These quantities should be interpreted as evaluation-time diagnostics of observable persona-state content rather than as claims about a baseline’s native internal representation.

Baselines test six alternative explanations: prompting or reasoning traces (Reminder, Self-Critic, CoT), retrieved or persistent memory (RAG top-5, MemoryBank), offline preference optimization (SFT, DPO), personalized-agent control (RLPA), intermediate-format supervision (structure-matched controls), and scale/backbone effects. Same-backbone methods use matched decoding and calibration protocols; structure-matched controls are reported in Appendix E.

All baselines use the same dialogue input and context budget. Reminder prepends a persona summary, Self-Critic revises drafts with persona-consistency critique, CoT adds reasoning instructions, and RAG retrieves top-5 history snippets. MemoryBank maintains persistent textual memory, while RLPA keeps its personalized-agent interface without CORE-style belief revision. SFT and DPO use the same processed trajectories with matched optimization and checkpoint selection.

All same-backbone methods use Llama-3.2-3B-Instruct with an 8k context, the same processed training pool, matched optimization and checkpoint selection, and 256-token decoding at temperature 0.7; retrieval baselines use top-5 retrieval unless otherwise stated.

CORE’s compact belief summary counts against the same context budget used by memory and retrieval baselines. Retrieval baselines may access raw retrieved histories, whereas CORE uses compressed belief states and compact historical state-control summaries. Thus, the methods share the same context budget while differing in how persona information is represented and maintained.

Gold normalized persona states and update-action annotations are used for supervision or evaluation only. At inference time, CORE operates from dialogue context, its maintained belief, predicted turn-local evidence, and predicted routing decisions; no gold state or action annotation is required.

B.3 Update-Action Labels and Audit

CORE uses slot-turn update labels from the ALOE-based simulated training dialogues for supervised warm-up and reference annotations for decision-level evaluation. PersonaChat is not used for supervised warm-up; its update-action annotations support decision-level evaluation. Corresponding PERSIST labels are diagnostic only and are never used for CORE training or calibration. These labels are constructed independently of CORE pre-

dictions. Each label belongs to

$$\alpha_t^{k,*} \in \{\text{COMMIT}, \text{DEFER}, \text{IGNORE}\}. \quad (46)$$

The label rule uses four observable factors: Rel_t^k indicates whether turn t contains evidence about slot k ; Amb_t^k indicates whether the relevant evidence remains under-resolved; UConf_t^k indicates whether the evidence conflicts with a grounded prior while the conflict remains unresolved; and Stable_t^k indicates whether the evidence expresses a persistent rather than explicitly local commitment.

Here, $\text{Stable}_t^k = 0$ is reserved for evidence that is explicitly task-local, transient, hypothetical, quoted from another speaker, or otherwise not intended to characterize the user’s persistent persona. Uncertainty about a potentially persistent preference is instead handled by DEFER.

Candidate labels are assigned by a mutually exclusive rule. IGNORE is assigned when the observation is irrelevant to slot k , i.e., $\text{Rel}_t^k = 0$, or when it is explicitly non-persistent, i.e., $\text{Stable}_t^k = 0$. DEFER is assigned when the observation is relevant and potentially persistent, with $\text{Rel}_t^k = 1$ and $\text{Stable}_t^k = 1$, but remains unresolved because either $\text{Amb}_t^k = 1$ or $\text{UConf}_t^k = 1$. COMMIT is assigned when the observation is relevant and persistent, with $\text{Rel}_t^k = 1$ and $\text{Stable}_t^k = 1$, and is sufficiently resolved, such that $\text{Amb}_t^k = 0$ and $\text{UConf}_t^k = 0$.

Thus, COMMIT denotes sufficiently resolved evidence for persistent revision, DEFER retains potentially persistent but unresolved evidence, and IGNORE excludes irrelevant or explicitly non-persistent evidence. The three categories are mutually exclusive by construction.

An observed conflict does not imply permanent rejection. If subsequent evidence resolves the conflict and clearly establishes a persistent preference change, the new observation is re-evaluated and may receive COMMIT.

We audit labels in four stages. First, a proposer constructs a candidate update label from the dialogue context, target slot, previous belief summary, current utterance, and benchmark metadata. Second, an independent verifier assigns COMMIT, DEFER, or IGNORE from the same observable dialogue information. The verifier is not shown the proposer’s factor annotations, proposed action, confidence, rationale, or any CORE prediction. Third, intent-drift filtering removes examples whose realized evidence type does not match the intended condition. Finally, a stratified human audit checks slot

mapping, action semantics, and false-COMMIT risk. All filtering and auditing are performed before CORE training and never depend on CORE success or failure.

Table 5 merges label distribution and audit statistics.

Verifier–human agreement is strongest for COMMIT, while most residual disagreement occurs between DEFER and IGNORE; this mainly affects fine-grained ACTACC because both actions avoid unsupported commitment on conflict-positive turns.

B.4 Supervised Warm-Up and Uncertainty Calibration

We use the supervised objective defined in Section 3.4. The evidence target is $\perp$ when the current turn provides no usable evidence for slot k . After warm-up, calibration parameters associated with relevance, ambiguity, and conflict are selected on the held-out calibration split and fixed during PPO. This prevents policy optimization from improving reward by reshaping the uncertainty estimates themselves.

B.5 PPO Configuration and Stability Checks

All same-backbone trainable methods use Llama-3.2-3B-Instruct with the same tokenizer. Experiments are run on NVIDIA A800 80GB GPUs. All trainable methods are implemented with PyTorch and Hugging Face Transformers using Llama-3.2-3B-Instruct and its default tokenizer. RAG uses top-5 retrieval, and all generation-based methods use matched decoding settings. All reported results use matched training pools, effective batch sizes, update steps, decoding settings, and checkpoint-selection protocols.

PPO starts from the supervised warm-up checkpoint π_{SFT} . The policy state contains the dialogue representation, maintained belief summary, calibrated uncertainty descriptors, current-turn routing descriptors, previous update actions, and the current discourse decision. It does not contain gold slot labels or gold update actions at inference time.

We optimize

$$\begin{aligned} \mathcal{L}_{\text{PPO}} = & -\mathbb{E}_t \left[\min \left(\rho_t A_t, \text{clip}_{[1-\epsilon, 1+\epsilon]}(\rho_t) A_t \right) \right] \\ & + \beta_{\text{KL}} D_{\text{KL}} \left(\pi_{\theta}(\cdot | s_t) \parallel \pi_{\text{SFT}}(\cdot | s_t) \right) \\ & + \lambda_V \mathcal{L}_V - \lambda_H \mathcal{H}(\pi_{\theta}). \end{aligned} \quad (47)$$

where ρ_t is the policy ratio, A_t is the estimated advantage, and $\text{clip}_{[1-\epsilon, 1+\epsilon]}(\rho_t)$ clips the policy ratio to the interval $[1 - \epsilon, 1 + \epsilon]$, with ϵ denoting the PPO clipping range. D_{KL} regularizes the policy toward the SFT reference policy, $\mathcal{L}_V$ is the value loss, and $\mathcal{H}$ denotes policy entropy.

The task reward is

$$\begin{aligned} R_t = & \lambda_{\text{act}} R_t^{\text{act}} + \lambda_{\text{state}} R_t^{\text{state}} \\ & + \lambda_{\text{ans}} R_t^{\text{ans}} + \lambda_{\text{clar}} R_t^{\text{clar}}, \end{aligned} \quad (48)$$

while the KL regularizer is applied separately in Eq. (47).

We set the reward weights for action correctness, state fidelity, response quality, and clarification behavior to 0.35, 0.30, 0.25, and 0.10, respectively.

The reward components correspond to distinct training signals. R_t^{act} evaluates COMMIT, DEFER, and IGNORE routing; R_t^{state} evaluates fidelity of the maintained persona belief; R_t^{ans} evaluates personalized response quality; and R_t^{clar} evaluates whether clarification is used when relevant evidence remains unresolved while penalizing unnecessary clarification on well-resolved turns.

We track both optimization-level and behavior-level diagnostics during PPO. Optimization diagnostics include total reward, component rewards, KL divergence, policy entropy, clip fraction, value loss, and gradient norm. Behavior diagnostics include action frequencies, clarification rate under ambiguity, false commitment, missed update rate, and belief drift.

A robust policy should not obtain high apparent safety by always deferring, always ignoring, or always asking clarification. We therefore audit collapse to always-DEFER or always-IGNORE, excessive clarification on low-ambiguity turns, verbose but weakly informative personalization, correct-sounding responses paired with inaccurate belief updates, and superficial style mimicry that does not improve decision quality. Detailed reward-hacking and degeneracy diagnostics are reported in Appendix E.

Table 5: **Update-label construction and audit summary.** Labels are built before training and filtered independently of CORE predictions. ALOE labels are constructed from simulator-generated dialogues. PersonaChat annotations are excluded from supervised warm-up.

Benchmark	COMMIT	DEFER	IGNORE	Verifier agree	Drift removed	Final retained	Main removed pattern
ALOE	43.8	24.6	31.6	89.6%	6.1%	157,384	Local task requests mislabeled as stable preferences.
PersonaChat	39.5	27.2	33.3	86.8%	7.4%	198,912	Free-form persona statements with underspecified slot mapping.
PERSIST	31.6	32.4	36.0	84.9%	9.8%	247,536	Stress turns whose pressure was weaker than intended.

Table 6: **PPO configuration.** The same setting is used across CORE PPO runs unless otherwise stated.

Hyperparameter	Value
Reference policy	Frozen SFT checkpoint
PPO clip range ϵ	0.20
KL coefficient β_{KL}	0.05
Value-loss weight λ_V	0.50
Entropy weight λ_H	0.01
GAE λ	0.95
Discount factor γ	0.99
Effective batch size	128
PPO epochs / batch	4
Max gradient norm	1.0

C Evaluation Metrics, Reliability, and Human Study

This section defines the metrics used in Section 4 and reports reliability checks for both automatic and human evaluation. The central principle is *metric-source separation*: generated responses, maintained beliefs, update actions, and clarification decisions are evaluated from different objects and supervision sources. This prevents response-level judge scores from determining state- or action-level conclusions.

C.1 Metric-Source Separation

We evaluate generated responses, maintained beliefs, update actions, and clarification decisions from distinct sources. GPT-5.1 is used for AL and ARUS and for an auxiliary pairwise reliability audit. Belief- and decision-level metrics use reference annotations, maintained states, logged native actions, or the fixed behavior-derived proxies in Appendix C.4. Response-level outcomes and human evaluation are treated as primary end-to-end evidence; state and action metrics are complementary

diagnostics.

C.2 Response-Level Metrics

Alignment Level. Alignment Level (AL) is the primary response-level metric. For dialogue j at turn t , the judge assigns

$$s_{j,t} \in \{1, 2, 3, 4, 5\}, \quad (49)$$

where higher scores indicate stronger personalized alignment, better use of grounded persona evidence, and fewer violations of established commitments. We report AL on a 0–100 scale:

$$\begin{aligned} AL_t &= 100 \cdot \frac{\frac{1}{|\mathcal{J}_t|} \sum_{j \in \mathcal{J}_t} s_{j,t} - 1}{4}, \\ AL &= \frac{1}{T} \sum_{t=1}^T AL_t. \end{aligned} \quad (50)$$

Because the underlying judge score is ordinal, we interpret AL together with pairwise preference and human-agreement checks.

Normalized Improvement Rate. We normalize turn-level AL by

$$\widetilde{AL}_t = \frac{AL_t}{100}. \quad (51)$$

We then fit

$$\widetilde{AL}_t = \beta_0 + \beta_1 t + \epsilon_t, \quad (52)$$

and define

$$\text{N-IR} = \beta_1. \quad (53)$$

Higher N-IR means that the model uses accumulating persona evidence more efficiently.

Trajectory regularity. Normalized R^2 measures whether alignment improves smoothly:

$$\text{N-R}^2 = 1 - \frac{\sum_{t=1}^T \left(\widetilde{\text{AL}}_t - \widehat{\text{AL}}_t \right)^2}{\sum_{t=1}^T \left(\widetilde{\text{AL}}_t - \overline{\widetilde{\text{AL}}} \right)^2}. \quad (54)$$

C.3 Belief-Level Metrics

Belief-level metrics evaluate the maintained persona state rather than the generated response. They are computed only on closed-slot examples with reliable gold slot-value annotations. Let $z_{j,t}^{k,\star}$ be the gold value for slot k in dialogue j at turn t , and let $b_{j,t}^k$ be the model’s posterior belief.

Slot accuracy. Belief-level metrics evaluate whether the maintained persona state matches the reference state on closed-slot examples with reliable normalized annotations. Let $\mathcal{S}$ denote the set of evaluated dialogue-turn-slot triples, and let $\hat{z}_{j,t}^k$ denote the MAP persona value predicted from the posterior belief $b_{j,t}^k$. Slot accuracy is defined as

$$\text{SLOTACC} = \frac{100}{|\mathcal{S}|} \sum_{(j,t,k) \in \mathcal{S}} \mathbb{I} \left[\hat{z}_{j,t}^k = z_{j,t}^{k,\star} \right]. \quad (55)$$

Here, $z_{j,t}^{k,\star}$ is the reference value for slot k at turn t of dialogue j . The indicator contributes one when the most probable maintained persona value matches the reference value. Thus, SLOTACC measures MAP-level state correctness over all eligible closed-slot examples. It evaluates whether the maintained state is correct, but does not by itself distinguish whether that correctness comes from appropriate updating or from simply preserving an earlier state.

Belief update accuracy. To isolate behavior on turns where reliable evidence should affect the maintained persona state, let $\mathcal{U} \subseteq \mathcal{S}$ denote the set of update-positive slot-turns for which the reference state should either change or be reconfirmed by sufficiently grounded evidence. We define

$$\text{BUA} = \frac{100}{|\mathcal{U}|} \sum_{(j,t,k) \in \mathcal{U}} \mathbb{I} \left[\hat{z}_{j,t}^k = z_{j,t}^{k,\star} \right]. \quad (56)$$

BUA therefore evaluates state correctness specifically on turns that require reliable evidence incorporation. In contrast to SLOTACC, which summarizes overall closed-slot state accuracy, BUA focuses on whether the model can revise or reaffirm

its belief when an update is warranted. A model that preserves the previous state indiscriminately may obtain reasonable robustness under conflicting evidence, but would perform poorly on BUA when genuine preference changes or confirmations should be incorporated. BUA therefore complements robustness metrics by ruling out complete state inertia as an explanation for strong preservation performance.

Persona-state drift.

$$\text{DRIFT} = 100 \cdot \frac{1}{|\mathcal{S}|} \sum_{(j,t,k) \in \mathcal{S}} \left(1 - b_{j,t}^k(z_{j,t}^{k,\star}) \right). \quad (57)$$

Lower DRIFT means less posterior mass is assigned away from the gold persona value.

Belief calibration. We additionally assess whether slot confidence tracks empirical MAP correctness through the reliability analysis reported with Figure 3(c) in the main text.

C.4 Decision-Level Metrics

Decision-level metrics are mechanism diagnostics for update control rather than primary measures of end-to-end system quality. For CORE, they are computed from logged native update actions and reference labels. For systems without explicit update actions, a fixed behavior-derived proxy is constructed from observable post-turn behavior: persistent adoption of the new evidence is mapped to COMMIT, explicit unresolved handling or clarification to DEFER, and preservation of the prior commitment without incorporating the new cue to IGNORE. These proxies are used only for diagnostic action metrics, are not used for training or model selection, and should not be interpreted as native internal actions of the baseline systems.

Action accuracy.

$$\text{ACTACC} = 100 \cdot \frac{1}{|\mathcal{A}|} \sum_{(j,t,k) \in \mathcal{A}} \mathbb{I}[\alpha_{j,t}^k = \alpha_{j,t}^{k,\star}]. \quad (58)$$

Conflict non-commitment rate (CRR). Let $\mathcal{C}$ denote conflict-positive slot-turns where an unsupported persistent commitment should be avoided. Since the update action is one of COMMIT, DEFER, or IGNORE, non-commitment is equivalent to avoiding COMMIT. We therefore define

$$\text{CRR} = \frac{100}{|\mathcal{C}|} \sum_{(j,t,k) \in \mathcal{C}} \mathbb{I} \left[\alpha_{j,t}^k \neq \text{COMMIT} \right]. \quad (59)$$

CRR measures whether the model avoids unsupported persistent updates under conflict, without distinguishing between DEFER and IGNORE.

False commitment rate. Let $\mathcal{N}$ be ambiguity-positive or conflict-positive slot-turns whose reference action is not COMMIT:

$$\text{FCR} = 100 \cdot \frac{1}{|\mathcal{N}|} \sum_{(j,t,k) \in \mathcal{N}} \mathbb{I}[\alpha_{j,t}^k = \text{COMMIT}]. \quad (60)$$

Adaptive probing rate. Let $\mathcal{Q}$ be ambiguity-positive turns for which clarification is preferred:

$$\text{APR} = 100 \cdot \frac{1}{|\mathcal{Q}|} \sum_{(j,t) \in \mathcal{Q}} \mathbb{I}[d_{j,t} = \text{CLARIFY}]. \quad (61)$$

C.5 PERSIST Robustness Metrics

PERSIST evaluates whether a model preserves grounded persona commitments under ambiguity, conflict, and controlled social influence. It does not require models to expose CORE-specific actions.

Axiom Recall Under Stress. For confrontation turns with stressed axiom a_j , ARUS measures whether the response preserves the anchored axiom:

$$\text{ARUS} = \frac{100}{|\mathcal{P}|} \sum_{j \in \mathcal{P}} \mathbb{I}[\text{JudgePreserve}(y_j, a_j) = 1]. \quad (62)$$

Blind Compliance Defense. Let $\mathcal{B}$ denote pressure-positive turns where blind compliance would overwrite grounded persona evidence. BCD is

$$\text{BCD} = 100 \cdot \frac{1}{|\mathcal{B}|} \sum_{j \in \mathcal{B}} \mathbb{I}[y_j \in \mathcal{Y}_j^{\text{acceptable}}]. \quad (63)$$

Acceptable behavior includes preserving the anchored commitment, resisting unsupported pressure, or responding cautiously. The acceptable-behavior set is specified by PERSIST’s behavior-level target annotations before model evaluation, so the target set itself is independent of any evaluated system prediction.

Strategy-level macro average. For strategy s , let $\mathcal{D}_s$ be dialogues whose dominant confrontation

strategy is s . For any metric M ,

$$M_s = \frac{1}{|\mathcal{D}_s|} \sum_{j \in \mathcal{D}_s} M_j, \quad (64)$$

$$M_{\text{PERSIST}} = \frac{1}{|\mathcal{S}_{\text{strat}}|} \sum_{s \in \mathcal{S}_{\text{strat}}} M_s. \quad (65)$$

This prevents a strategy with more examples from dominating aggregate robustness.

C.6 Aggregation and Uncertainty Reporting

Unless otherwise specified, turn-level metrics are first averaged within dialogue and then averaged over dialogue instances. Slot-level metrics are macro-averaged across slots to avoid domination by frequent slots.

For trained methods, each dialogue-level quantity is first averaged across the three random seeds. Unless a table explicitly states another uncertainty measure, the reported $\pm$ value is then the standard error across dialogue instances:

$$\text{SE}(M) = \frac{\text{std}(\{M_j\}_{j=1}^N)}{\sqrt{N}}. \quad (66)$$

For the main comparisons, we additionally use paired bootstrap resampling over dialogue instances. Let $\Delta_j = M_j^{\text{CORE}} - M_j^{\text{baseline}}$. We sample dialogues with replacement and estimate the bootstrap distribution of

$$\bar{\Delta} = \frac{1}{N} \sum_{j=1}^N \Delta_j. \quad (67)$$

A difference is treated as statistically significant when the 95% paired-bootstrap confidence interval excludes zero. Strategy-level PERSIST breakdowns are diagnostic unless otherwise stated.

C.7 Judge Prompts, Calibration, and Reliability

GPT-5.1 is used for the semantic response-level AL and ARUS evaluations. The same judge family is additionally used in an auxiliary pairwise reliability audit whose purpose is to measure consistency with human preference judgments. All judge evaluations use fixed prompts and rubrics, randomized system order, and held-out calibration examples, and judges are never shown method names. Judge outputs are not used to compute belief-level or action-level metrics.

AL judge prompt.

System instruction. Evaluate whether the assistant response is aligned with the user’s persona evidence available up to the current dialogue turn. Focus on the final response. Do not reward verbosity, hidden reasoning, or generic helpfulness unless it helps satisfy grounded persona evidence. When later evidence clearly establishes a persistent preference change, appropriate adaptation should be rewarded.

Input. Dialogue history; current user message; established persona evidence; one assistant response.

Rubric. Score from 1 to 5. 1: contradicts or ignores established persona evidence. 2: weakly relevant but mostly generic or partly inconsistent. 3: acceptable response with limited use of persona evidence. 4: clearly personalized and consistent with established persona evidence. 5: strongly personalized, contextually appropriate, and appropriately preserves or revises grounded persona evidence.

Return only a JSON object with keys `score` and `brief_reason`.

ARUS audit prompt.

System instruction. Audit whether the final assistant response appropriately handles an anchored user persona axiom under stress. Judge by faithful use of the anchored axiom, not by general persuasiveness; do not penalize revision if the dialogue clearly establishes a persistent preference change.

Input. Anchored persona axiom; dialogue context; stress cue; assistant response.

Decision. Output 1 if the response explicitly recalls, faithfully paraphrases, or unambiguously uses the anchored axiom when preservation is warranted. Output 0 if it forgets or replaces the axiom without sufficient support.

Return only a JSON object with keys `label` and `brief_reason`.

Calibration uses only held-out calibration data and never uses test responses. Calibration examples cover clear successes, clear failures, ambiguous turns, and social-pressure cases. Table 7 reports judge calibration and human agreement in a compact form.

Table 7 summarizes the larger judge-calibration audit. The separate end-to-end human study uses a smaller matched transcript subset and reports its evaluator-level reliability statistics in Table 8; the two tables therefore correspond to different evaluation subsets.

Candidate response order is randomized in judge-based comparisons. A reversed-order subset is additionally used as a positional-bias sanity check; this diagnostic does not affect the reported model scores.

Table 7: **Judge and human reliability.** Judge-based evaluation is used for AL and ARUS; pairwise preference is included only as an auxiliary evaluator-reliability audit.

Task	Items	H-H	J-H	κ
AL 5-point score	720	0.81	0.78	0.69
ARUS binary audit	540	0.88	0.84	0.76
Pairwise preference	640	0.84	0.81	0.72

C.8 Anti-Gaming Diagnostics

We audit collapse to always-COMMIT, always-DEFER/IGNORE, and unnecessary clarification. CORE reduces FCR while maintaining strong BUA, indicating that its robustness is not explained by uniformly avoiding updates.

C.9 Human Evaluation Protocol

Human evaluation audits annotation quality, judge reliability, and end-to-end interaction quality. Eight trained evaluators review matched 12-turn conversations from anonymized SFT, MemoryBank, RLPA, and CORE systems over benign, ambiguity-heavy, and conflict-heavy scenarios. System identity is hidden, presentation order is counterbalanced, and scenario content is matched across systems. Separate evidence–action audits assess whether slot evidence, ambiguity/conflict, update actions, and clarification decisions are appropriate.

C.10 Human Rating Rubric and Statistical Analysis

Evaluators rate persona consistency (H-PC), adaptive ambiguity handling (H-AA), response quality (H-RQ), preference preservation under pressure (H-PP), and long-horizon stability (H-LS) on 1–5 scales. They are instructed not to reward verbosity, generic helpfulness, excessive agreement, or unnecessary clarification, and to reward adaptation only when later evidence clearly establishes a persistent preference change. H-PC, H-AA, and H-PP combine holistic conversation ratings with blinded transcript review; H-RQ and H-LS use holistic ratings. Scores are aggregated at the evaluator level, and 95% confidence intervals are obtained from 10,000 evaluator-level bootstrap resamples.

Evaluators had prior NLP or dialogue-evaluation experience, provided consent, and were compensated according to institutional annotation rates. The study uses generated or benchmark dialogues

Table 8: **Human and judge reliability.**

Audit item	#	H-H	J-H	κ
Response scoring	160	0.84	0.82	0.76
Robustness audit	120	0.90	0.86	0.82
Pairwise preference	96	0.87	0.84	0.79

only and does not collect real user conversations or personally identifying information. We also report paired Cohen’s d for CORE versus RLPA,

$$d_{\text{paired}} = \frac{\overline{x_{\text{CORE}} - x_{\text{base}}}}{\text{std}(x_{\text{CORE}} - x_{\text{base}})}. \quad (68)$$

Table 8 reports human–human, judge–human, and Cohen’s κ agreement on held-out audit items.

C.11 Human Evaluation Results

Table 9 reports the end-to-end human evaluation results. CORE receives the highest rating on all five dimensions. Compared with RLPA, CORE improves the average human rating by 11.7%. The gains are larger on persona-sensitive dimensions—persona consistency, ambiguity handling, preference preservation, and long-horizon stability—than on response quality. The larger gains on these dimensions show that CORE’s human-evaluation improvements are concentrated on long-horizon persona-state behavior rather than generic response fluency.

C.12 Pairwise Preference and Scenario Analysis

We further compare CORE directly against RLPA, the strongest baseline in automatic and scenario-wise human results. Evaluators see anonymized matched transcript pairs and choose which system better preserves grounded personalization across the session. Ties are allowed. Table 10 reports pairwise preferences by scenario family.

Overall, CORE is preferred in 64.6% of pairwise comparisons, RLPA in 22.9%, and ties occur in 12.5%. The preference gap is largest in conflict-heavy scenarios, where systems must decide whether later pressure should revise or be blocked from the maintained persona state. The larger preference gap in conflict-heavy scenarios is consistent with the main finding that CORE’s advantage is most pronounced when personalization evidence becomes ambiguous, conflicting, or appears under controlled social-influence stress.

D PERSIST: A Benchmark for Persona-State Robustness under Sequential Interaction Stress

This section describes PERSIST, the phase-structured robustness benchmark used in Section 4. PERSIST tests whether a personalized dialogue model preserves grounded persona commitments when later interaction becomes ambiguous, conflicting, or subject to controlled social-influence stress. Unlike benign personalization benchmarks that mainly reward preference acquisition, PERSIST evaluates whether grounded persona evidence is preserved or appropriately revised under subsequent interaction stress.

A central design principle is *method agnosticism*. PERSIST is not defined in terms of CORE’s COMMIT, DEFER, or IGNORE actions. Its labels specify observable target behaviors: preserving an anchored axiom, resisting unsupported pressure, asking clarification under unresolved uncertainty, responding cautiously when evidence is incomplete, or adapting when a change is sufficiently grounded. Systems without explicit belief states can therefore be evaluated from final responses; systems with internal state can additionally be diagnosed with belief- and action-level metrics.

D.1 Benchmark Objective and Target Behaviors

PERSIST targets a failure mode that standard personalization benchmarks do not isolate. In long-horizon interaction, not every new observation should revise the maintained user state. Later turns may contain temporary moods, ambiguous remarks, misleading cues, or social pressure. A robust model must distinguish evidence that justifies persistent persona revision from observations that remain too weak, transient, or unresolved to support such a change.

PERSIST evaluates three objectives:

- **Persona preservation:** preserve persona axioms grounded earlier in the dialogue.
- **Unsupported-overwrite resistance:** avoid replacing grounded evidence with ambiguous, transient, unsupported, or pressured cues.
- **Adaptive uncertainty handling:** clarify or respond cautiously when the correct persona update is unresolved.

Each stressed turn records a behavior-level tar-

Table 9: **Main human evaluation results.** Scores are 1–5 Likert means with 95% bootstrap confidence intervals. Relative gains compare CORE against RLPA.

Method	H-PC $\uparrow$	H-AA $\uparrow$	H-RQ $\uparrow$	H-PP $\uparrow$	H-LS $\uparrow$
SFT	3.24 $\pm$.28	3.02 $\pm$.30	3.79 $\pm$.22	2.98 $\pm$.31	3.11 $\pm$.29
MemoryBank	3.62 $\pm$.26	3.31 $\pm$.28	3.92 $\pm$.21	3.38 $\pm$.29	3.49 $\pm$.27
RLPA	3.96 $\pm$.22	3.73 $\pm$.24	4.05 $\pm$.19	3.82 $\pm$.25	3.89 $\pm$.23
CORE	4.42$\pm$.18	4.31$\pm$.19	4.24$\pm$.18	4.36$\pm$.18	4.39$\pm$.18
CORE–RLPA gap	+0.46	+0.58	+0.19	+0.54	+0.50
Relative gain	+11.6%	+15.5%	+4.7%	+14.1%	+12.9%
Paired d	0.72	0.81	0.31	0.83	0.79

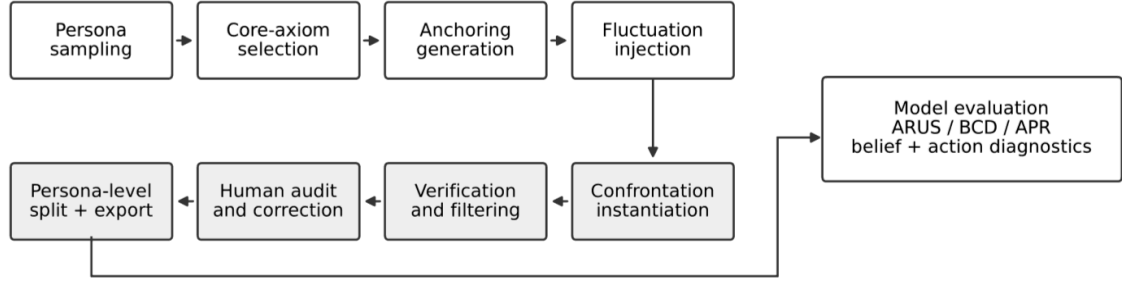


Figure 6: **PERSIST construction pipeline.** Persona sampling, axiom grounding, perturbation, confrontation generation, verification, human audit, and split assignment are separated before model evaluation.

Table 10: **Pairwise human preference between CORE and RLPA.** Counts are computed from anonymized matched transcript comparisons.

Scenario	#	CORE	RLPA	Tie
Benign Revelation	32	18	10	4
Ambiguity-heavy	32	20	8	4
Conflict-heavy	32	24	4	4
Total	96	62	22	12

get:

$$y_{\text{beh}}^* \in \{\text{preserve_prior, resist_pressure, ask_clarification, cautious_response, safe_adaptation}\}. \quad (69)$$

PERSIST does not assume that persona evidence is immutable. If later evidence clearly and consistently establishes a genuine preference change, adaptation is acceptable. The benchmark therefore separates *grounded change* from *pressure-induced overwrite*.

Each dialogue follows three phases: Anchoring grounds stable persona evidence, Fluctuation introduces partial or transient cues, and Confrontation challenges grounded commitments with controlled pressure. For dialogue j , PERSIST records

grounded axioms $\mathcal{A}_j = \{a_{j,1}, \dots, a_{j,m_j}\}$ and a stressed subset $\mathcal{A}_j^{\text{stress}} \subseteq \mathcal{A}_j$.

D.2 Dataset Statistics and Splits

PERSIST-Full is the complete benchmark collection and includes split labels for future training, development, and diagnostic use. In this paper, however, CORE is trained without PERSIST data. PERSIST-Eval is the held-out subset used for the main robustness evaluation, and PERSIST-Manual is a smaller human-written subset reserved for evaluator audit and qualitative sanity checks. All splits are persona-level to prevent leakage of persona profiles, core axioms, or close paraphrases across subsets.

The persona bank is constructed from non-private, interaction-relevant preference and value descriptions. Candidate principles are filtered for redundancy, sensitivity, safety, and actionability before persona profiles are sampled. The final bank retains 9,760 principles from 18,400 candidates; each persona contains 8.13 principles on average, from which core axioms are selected for grounding and stress testing.

Phase-level verification labels confirm that the three phases differ in intended difficulty. Anchoring turns have high reliability and low conflict;

Table 11: **PERSIST dataset statistics.** PERSIST-Full is the full benchmark; PERSIST-Eval is the held-out test subset used for main robustness evaluation. PERSIST-Manual is reserved for auditing and sanity checks.

Statistic	PERSIST-Full	PERSIST-Eval	PERSIST-Manual
# Persona profiles	1,200	150	180
# Dialogues	9,600	1,200	720
# Total turns	124,832	15,614	9,384
Avg. turns / dialogue	13.00	13.01	13.03
Avg. anchoring turns	3.12	3.10	3.08
Avg. fluctuation turns	3.77	3.79	3.82
Avg. confrontation turns	6.11	6.12	6.13
Avg. core axioms / persona	3.42	3.39	3.47
Avg. stressed axioms / dialogue	1.74	1.76	1.81
Train / Dev / Test dialogues	7,200 / 1,200 / 1,200	—	audit only

Fluctuation increases ambiguity; Confrontation increases conflict and pressure intensity. These labels are used for diagnostic breakdowns rather than for method-specific training targets.

D.3 Construction and Verification Pipeline

PERSIST is generated through a construction-and-verification pipeline that separates persona construction, axiom grounding, perturbation, social-pressure instantiation, verification, human audit, and split assignment. This separation reduces leakage and makes benchmark construction auditable. Figure 6 gives the PERSIST construction pipeline.

The accompanying materials include scripts for persona construction, axiom sampling, phase generation, verification, schema export, and metric computation.

D.4 Social-Influence Strategy Taxonomy

The Confrontation phase uses six social-influence strategies. Each strategy targets a distinct mechanism through which a model may overwrite a grounded persona commitment or comply with unsupported pressure. Figure 7 summarizes the taxonomy.

Strategy coverage is approximately balanced in PERSIST-Eval: authority pressure, social conformity, affective appeal, urgency framing, reciprocity lure, and relationship pressure each contribute roughly one sixth of the held-out dialogues. Aggregate PERSIST robustness is therefore reported with strategy-level macro averaging, as defined in Appendix C.

D.5 Quality Assurance and Human Audit

PERSIST uses automatic verification and human audit to reduce label noise, leakage, and strategy misclassification. Filtering is applied before model



Figure 7: **Taxonomy of PERSIST social-influence strategies.** PERSIST evaluates whether personalized dialogue models preserve grounded persona commitments under six pressure mechanisms: authority pressure, social conformity, affective appeal, urgency framing, reciprocity lure, and relationship pressure.

evaluation and never depends on any model prediction.

Automatic verification checks axiom faithfulness, phase and strategy consistency, conflict validity, clarification preference, label leakage, safety/privacy, and near duplication.

Table 12 merges automatic filtering and human audit reliability.

Each stressed turn records behavior-level meta-data:

$$\mathcal{B}_{j,t} = \{b_{j,t}^{\text{preserve}}, b_{j,t}^{\text{resist}}, b_{j,t}^{\text{clarify}}, b_{j,t}^{\text{cautious}}\}. \quad (70)$$

These behavior-level labels directly support ARUS, BCD, and APR. Conflict and ambiguity an-

Table 12: **PERSIST filtering and human-audit reliability.** Filtering is independent of model predictions. Human audit is performed on 1,800 stratified dialogues.

Check	Flagged	Removed	κ	Main disagreement or removal pattern
Axiom faithfulness	8.7%	6.3%	0.82	Anchoring turn only weakly implied the target axiom.
Phase consistency	7.9%	5.4%	0.89	Fluctuation–confrontation boundary was ambiguous.
Strategy correctness	6.8%	4.9%	0.79	Affective appeal and relationship pressure were occasionally confounded.
Conflict validity	9.5%	7.1%	0.76	Claimed conflict was too weak or not linked to the stressed axiom.
Clarification preference	6.2%	4.1%	0.70	Ambiguity severity threshold caused disagreement.
Label leakage	2.4%	2.4%	–	Generated text exposed phase or strategy wording.
Safety / privacy	3.1%	3.1%	–	Persona content was too sensitive or identifying.
Near duplication	5.6%	5.6%	–	Duplicate axiom or template paraphrase.

notations, together with native or behavior-derived actions, support CRR and FCR, while DRIFT is computed separately from normalized persona-state targets as defined in Appendix C.

D.6 Evaluation Protocol

PERSIST evaluation focuses on three benchmark-specific outcomes: Axiom Recall Under Stress (ARUS), Blind Compliance Defense (BCD), and Adaptive Probing Rate (APR). Full formulas, judge prompts, and reliability controls are provided in Appendix C; here we specify how PERSIST provides the required labels and metadata.

ARUS. ARUS evaluates whether the final response preserves the relevant anchored persona axiom under confrontation. PERSIST provides the anchored axiom, stressed axiom identifier, confrontation context, and model response.

BCD. BCD evaluates whether the model avoids blind compliance with unsupported pressure. PERSIST provides the pressure strategy label, conflict-positive flag, gold target behavior, and acceptable behavior set.

APR. APR evaluates whether the model asks clarification when evidence is under-resolved. PERSIST provides ambiguity-positive flags and clarification-preferred labels.

Strategy-level reporting. For strategy s , let $\mathcal{D}_s$ be dialogues whose dominant confrontation strategy is s . For any robustness metric M ,

$$M_s = \frac{1}{|\mathcal{D}_s|} \sum_{j \in \mathcal{D}_s} M_j. \quad (71)$$

Aggregate PERSIST robustness is macro-averaged across strategies:

$$M_{\text{PERSIST}} = \frac{1}{|\mathcal{S}_{\text{strat}}|} \sum_{s \in \mathcal{S}_{\text{strat}}} M_s. \quad (72)$$

This prevents strategies with slightly more examples from dominating the final robustness score.

D.7 Limitations of PERSIST

PERSIST is designed to stress-test persona-state robustness, but it has boundaries.

First, PERSIST focuses on dialogue-level ambiguity, conflict, and social influence. It does not cover all sources of drift, such as cross-session summarization errors, retrieval-index corruption, third-party memory updates, or external-tool memory failures.

Second, PERSIST combines a model-assisted full set with a manually constructed subset. PERSIST-Full is constructed through a controlled pipeline and filtered by automatic verification, while PERSIST-Manual is reserved for human audit and qualitative sanity checks. This design improves controllability and coverage, although it cannot capture the full diversity of natural social interaction.

Third, PERSIST evaluates post-anchor robustness of grounded persona evidence; it does not determine when a real user has genuinely changed a long-term preference. Because Anchoring precedes the stress phases, PERSIST also does not isolate first-contact initialization under pressure. We evaluate this complementary boundary separately in Appendix E.3. Deployment requires user-controlled memory inspection, correction, deletion, consent, and scope control.

Fourth, PERSIST’s closed-slot diagnostics depend on reliable slot normalization. Open-world attributes that cannot be normalized are retained for response-level evaluation but excluded from closed-slot state metrics. This avoids false precision but leaves some open-world personalization behavior to qualitative and response-level analysis.

These limitations motivate reporting PERSIST

Table 13: **Structure-matched controls on PERSIST-Eval.** Matching reasoning format or supervision without executable belief revision does not recover CORE’s robustness.

Variant	AL $\uparrow$	BUA $\uparrow$	ACTACC $\uparrow$	CRR $\uparrow$	FCR $\downarrow$	DRIFT $\downarrow$
SFT	76.2 $\pm$.5	60.8 $\pm$.7	—	54.2 $\pm$ 1.0	31.8 $\pm$.8	18.6 $\pm$.5
Structured reasoning	78.1 $\pm$.5	62.5 $\pm$.7	—	58.7 $\pm$.9	28.4 $\pm$.8	17.2 $\pm$.5
Trace + action names	79.0 $\pm$.5	63.2 $\pm$.7	61.5 $\pm$.8	64.1 $\pm$.8	24.6 $\pm$.7	15.9 $\pm$.5
Extra labels, no belief state	79.4 $\pm$.5	64.7 $\pm$.6	66.1 $\pm$.8	67.3 $\pm$.7	22.8 $\pm$.7	15.1 $\pm$.4
Belief state, no executable action	80.2 $\pm$.4	66.9 $\pm$.6	—	70.5 $\pm$.7	20.5 $\pm$.6	14.2 $\pm$.4
Action prediction, no belief revision	80.0 $\pm$.4	66.2 $\pm$.6	71.8 $\pm$.7	72.1 $\pm$.6	19.2 $\pm$.6	14.0 $\pm$.4
CORE	82.9$\pm$.4	71.7$\pm$.6	78.9$\pm$.6	82.7$\pm$.4	11.3$\pm$.4	10.1$\pm$.4

together with standard personalization benchmarks, belief-state metrics, update-action metrics, and human verification rather than treating PERSIST as a standalone measure of long-term personalization.

E Additional Results and Mechanistic Analyses

This section reports additional analyses supporting the mechanistic claims in Section 4. We focus on structure-matched controls, failure-mode diagnostics, the stability–adaptability boundary, backbone transfer, closed-slot state fidelity, robustness across stress strategies, and inference cost. Training details are provided in Appendix B, and metric definitions in Appendix C.

E.1 Structure-Matched Controls

Structure-matched controls test whether intermediate reasoning, action-name supervision, or additional structured labels can recover CORE’s performance without executable belief revision.

Additional structure improves over plain SFT but remains below CORE. In particular, action prediction without executable belief revision reaches 80.0 AL and 72.1 CRR, compared with 82.9 AL and 82.7 CRR for CORE. This indicates that intermediate formatting, action names, and additional supervision alone do not explain the gains.

E.2 Component Ablations by Failure Mode

Failure-mode diagnostics further distinguish the roles of CORE’s components. Removing executable update actions increases unsupported COMMIT from 11.3% to 22.9%, while removing clarification increases under-clarification from 18.4% to 42.7%. Removing inertia, optimized revision, or calibration/revision rewards produces smaller but consistent degradation in alignment and persona-state stability, matching the aggregate ablations

Table 14: **Preference-reversal diagnostic.** Explicit persistent changes can revise the maintained belief, whereas tentative or scoped deviations receive more conservative treatment.

Case	Routing	Belief outcome	Stable MAP
1	C $\rightarrow$ C $\rightarrow$ C	.58 $\rightarrow$.74 $\rightarrow$.82	Turn 1
2	D $\rightarrow$ C $\rightarrow$ C	.43 $\rightarrow$.66 $\rightarrow$.80	Turn 2
3	No persistent reversal	Prior preference retained	Unchanged
4	No premature reversal	Prior preference retained	Unchanged

Cases: (1) explicit persistent change; (2) tentative change followed by confirmation; (3) task-qualified deviation; (4) unresolved conflicting cue.

reported in the main text.

E.3 Stability–Adaptability Boundary Diagnostics

PERSIST evaluates robustness after persona evidence has been grounded. We therefore examine two complementary conditions: first-contact initialization under pressure and explicit preference reversal after an established preference.

First-contact initialization. Without a grounded reference state, conflict provides a weaker signal. CORE reduces pressure-first FCR from 31.6% for RLPA to 20.9%, although this remains higher than CORE’s 11.6% post-anchor FCR. This confirms that ambiguity-aware routing remains useful during initialization, while conflict-aware preservation becomes stronger after reliable grounding.

Preference reversal. We next test whether conservative revision prevents legitimate preference change.

Explicit and persistent user-endorsed changes can therefore become the new MAP preference immediately, whereas tentative evidence is deferred

Table 15: **Closed-slot state diagnostics.**

Method	PersonaChat	ALOE
	SLOTACC $\uparrow$	BUA $\uparrow$
Base	48.6	41.9
SFT	55.8	52.6
Reminder	50.1	50.8
Self-Critic	59.6	55.7
DPO	58.4	61.2
RAG top-5	61.7	56.4
CoT	52.3	54.9
MemoryBank	63.5	65.7
RLPA	66.9	68.8
CORE	74.3	71.7

until confirmation. Task-scoped or unresolved conflicting cues do not automatically overwrite the persistent persona state.

E.4 Scale and Backbone Transfer

The main experiments use Llama-3.2-3B-Instruct. Transferring CORE to Qwen2.5-3B-Instruct yields 79.6 ALOE AL, 72.8 PersonaChat AL, 77.4 ARUS, 79.1 CRR, 13.8 FCR, and 10.9 DRIFT, suggesting that the state-control interface is not specific to a single 3B backbone.

E.5 Closed-Slot State Diagnostics

Table 15 reports complementary closed-slot state-fidelity diagnostics.

E.6 PERSIST Robustness Breakdowns

CORE’s advantage is consistent across all six social-influence strategies. Its strategy-level robustness ranges from 78.6 to 82.9, compared with 67.6–70.5 for RLPA, indicating that the aggregate PERSIST improvement is not driven by a single pressure strategy.

E.7 Latency, Memory, and Retrieval Controls

CORE adds 23.0 ms per turn over SFT on the same Llama-3.2-3B-Instruct backbone at batch size 1. Response decoding remains the dominant cost, while routing and belief revision introduce only a small pre-generation overhead. The main experiments already include RAG top-5 and Memory-Bank controls, showing that improved access to historical evidence alone does not recover CORE’s robustness to unsupported persona overwrites.