DO LANGUAGE MODELS HAVE A MODEL OF YOU? DISCOVERING AND ENHANCING LATENT REPRESENTATIONS OF USER PERSONAS

Anonymous authors
Paper under double-blind review

ABSTRACT

Humans carry working models of the people we know, and these models are useful because they persist across contexts and support predictions about future behavior. Language-model personalization is usually evaluated very differently: almost entirely by whether the model produces an adapted response. This creates a central blind spot. A system can achieve strong personalized output quality while maintaining a user state that is weakly recoverable, poorly calibrated, or unstable across different histories of the same person. We therefore distinguish *behavioral personalization*, which concerns whether outputs look adapted to a user, from *representational personalization*, which concerns whether the system maintains a recoverable and reusable user-state estimate that explains behavior across interactions. We formalize representational personalization through *persona identifiability*: a useful user state should be predictive of held-out behavior, recoverable from past interactions, and stable across partial histories of the same user. To study this distinction, we introduce HUMANLM-PROBE, a controlled diagnostic protocol built from HumanLM user profiles and interaction records under a strict public/private split, and PERSONA-LR, an EM-inspired **l**atent **r**efinement procedure that couples behavior prediction with persona inference over explicit user-state estimates. Across controlled recovery, calibration, history-stability, user-shuffling, long-history, and real-user personalization evaluations, we find that output quality and user-model quality are related but not equivalent: models with similar or strong downstream personalization can differ substantially in the recoverability and stability of their inferred user states. Moreover, explicitly optimizing for identifiable user states improves recovery, calibration, cross-history consistency, and held-out user behavior prediction while remaining competitive or stronger on downstream personalization tasks. These results suggest that personalization should be evaluated not only by what a model says, but also by the quality of the user model that supports its behavior.

1 INTRODUCTION

We routinely infer models of others from limited evidence; even brief observations can support predictions about interpersonal outcomes. Theory-of-mind research studies how people attribute beliefs, desires, intentions, and percepts to explain and predict behavior (Ambady & Rosenthal, 1992; Premack & Woodruff, 1978; Frith & Frith, 2005; Baker et al., 2017). More broadly, cognitive accounts of abstraction emphasize that people build compact, structured representations from sparse observations (Tenenbaum et al., 2011). These implicit models shape what we expect of others, what we ask them for, and how we interpret their responses.

Language models are now deployed in settings that seem to require similar capabilities, including assistants, tutors, recommenders, and long-horizon agents acting on users' behalf. In these settings, effective personalization requires user representations that generalize across interactions. Yet modern systems can adapt behavior through in-context learning, demonstrations, or external memory mechanisms (Brown et al., 2020; Min et al., 2022; Packer et al., 2023). Such output adaptation, though potentially useful, has limited consistency and transparency. Evidence from persona simulation shows that models can follow assigned personas while collapsing distinct profiles into similar behavioral

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

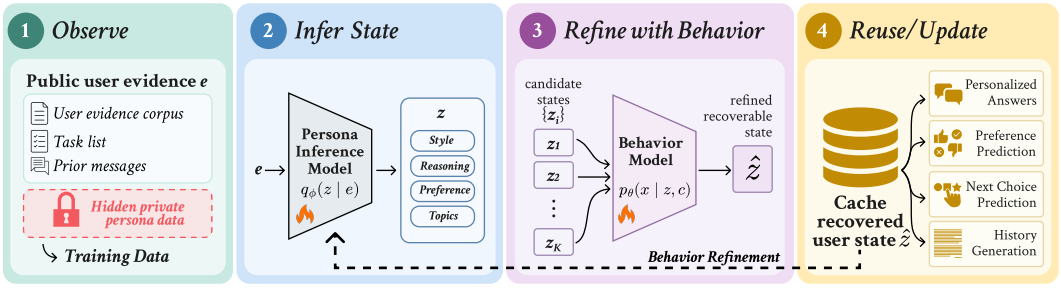

Figure 1: Standard personalization methods can improve responses by conditioning on recent context or retrieved history, but this does not require an explicit user state that persists beyond the current input. PERSONA-LR instead infers a structured user-state estimate from observable interaction history, reranks candidate states by how well they explain held-out behavior, and evaluates state consistency across partial histories.

modes (Xiao et al., 2026). This creates a central ambiguity: a model may behave as if it knows the user while relying primarily on local context.

Most prior work evaluates personalization through downstream behavior, using explicit personas, inferred profiles, or memory mechanisms to adapt model outputs (Li et al., 2016; Zhang et al., 2018; Cho et al., 2022; Weston et al., 2014; Packer et al., 2023). Instruction tuning and preference optimization methods likewise improve response quality without directly testing whether models maintain stable user representations (Ouyang et al., 2022; Rafailov et al., 2023). By contrast, representation analysis in machine learning studies what information is encoded inside learned models beyond their external behavior (Bengio et al., 2013; Alain & Bengio, 2016; Belinkov & Glass, 2019; Hewitt & Manning, 2019; Rogers et al., 2020). We argue that personalization should similarly be evaluated not only by what models say, but by the quality of user representations they encode.

We frame this gap as a distinction between two notions of personalization. *Behavioral personalization* means outputs appear adapted to the user. *Representational personalization* means that the system maintains an explicit, reusable user-state estimate that helps explain behavior across interactions. We study this through the principle of *persona identifiability*: useful user representations should **(i)** predict future behavior, **(ii)** be recoverable from past behavior, and **(iii)** remain consistent under different partial views of the user's history, such as prefixes or subsets of prior interactions.

To study representational personalization using persona identifiability, we make three contributions:

1. **HUMANLM-PROBE**, a controlled diagnostic protocol derived from HumanLM (Wu et al., 2026) for evaluating persona identifiability. HUMANLM-PROBE separates public behavioral evidence from private latent factors, enabling direct measurement of (a) predictivity of future behavior, (b) recoverability from past interactions, and (c) stability across partial histories.
2. **PERSONA-LR**, an EM-inspired latent refinement framework for representational personalization. PERSONA-LR couples a behavior model $p_\theta(x \mid z, c)$ with a persona inference model $q_\phi(z \mid e)$, optimizing user states for predictivity, recoverability, and consistency.
3. **Evidence that output-level personalization does not fully characterize user-model quality.** Across controlled user-state recovery, calibration, cross-history stability, user-shuffling interventions, and external personalization benchmarks, we find that models with similar or strong downstream personalization can differ substantially in whether they maintain recoverable, user-specific, and stable representations. Explicitly optimizing these representational properties improves both user-state quality and held-out personalization, showing the value of this approach. **These findings support our central claim that personalization should be evaluated not only by what a model outputs, but also by the user representation underlying those outputs.**

2 RELATED WORK

Personalization in language models. Personalization has long been studied through speaker embeddings, explicit user profiles, user-factor adaptation, and persona conditioning in dialogue

systems (Li et al., 2016; Zhang et al., 2018; Cho et al., 2022; Lynn et al., 2017). More recent work studies LLMs for user simulation, personalized assistants, private-user-data understanding, and profile-based question answering (Wu et al., 2026; Zhao et al., 2025; Tan et al., 2025; Su et al., 2025). Modern LLM systems often shape behavior through memory, instruction tuning, constitutional feedback, and preference optimization (Packer et al., 2023; Ouyang et al., 2022; Bai et al., 2022; Rafailov et al., 2023). These approaches can improve task utility or response quality, but are typically evaluated through output behavior rather than through the recoverability or stability of induced user representations. Related work also shows that persona prompting can collapse distinct profiles into similar behaviors, highlighting limits of output-level fidelity alone (Xiao et al., 2026).

Benchmarks for personalized systems. A growing literature evaluates whether language models use user context effectively. LaMP introduced profile-grounded personalized tasks (Salemi et al., 2024), while LongLaMP extends this setting to longer histories and personalized long-form generation (Kumar et al., 2024). PersonaLens evaluates conversational assistants using user profiles and interaction histories (Zhao et al., 2025), while PersonaBench evaluates models' understanding of personal information from synthetic private user data (Tan et al., 2025). PANDORA provides Reddit histories labeled with personality and demographic attributes (Gjurković et al., 2021). These benchmarks mainly measure task success, response quality, private-information understanding, or trait inference. By contrast, we study explicit user-state estimates that are predictive, stable, and recoverable from behavior.

Interpretability and representation analysis. Representation learning fundamentally emphasizes useful latent abstractions (Bengio et al., 2013), and understanding representations has been a recent focus across machine learning. For instance, probing methods analyze information encoded in intermediate representations (Alain & Bengio, 2016). In NLP, such methods have been used to recover syntactic, semantic, and task-relevant structure from hidden states (Belinkov & Glass, 2019; Hewitt & Manning, 2019; Rogers et al., 2020). Later work studies transformer specialization, factual recall, latent knowledge, and mechanistic structure (Voita et al., 2019; Geva et al., 2023; Burns et al., 2022; Elhage et al., 2022). Our focus is operational interpretability of user representations: whether user models can be inspected, recovered, and evaluated.

Latent user modeling, memory. Outside language modeling, recommendation systems represent users as latent states inferred from interaction histories through matrix factorization, sequential modeling, and self-attention (Koren et al., 2009; Rendle et al., 2012; Hidasi et al., 2015; Kang & McAuley, 2018). Prior work has explored mechanisms for extending model behavior beyond a single prompt, including memory networks (Weston et al., 2014; Sukhbaatar et al., 2015), LLM memory systems (Packer et al., 2023), generative agents (Park et al., 2023), reasoning-and-acting frameworks (Yao et al., 2022), and tool-use learning (Schick et al., 2023). However, storing more history is not equivalent to learning a coherent, recoverable user representation. Explicit user modeling also raises privacy concerns such as memorization and training-data extraction (Carlini et al., 2021).

3 METHOD

Problem Setup Let N denote user episodes, with $i \in \{1, \ldots, N\}$. We observe $D = \{(x_i, c_i, e_i)\}_{i=1}^{N}$ where e_i denotes the observable user evidence available for inferring the user state, such as prior messages. The task context c_i specifies the current problem, excluding the user state. Depending on the data available, this may include task type, input query, item or decision context, visible-history budget, evidence indices, and schema metadata used to construct the model prompt. Finally, $x_i = (x_{i1}, \ldots, x_{iT})$ denotes the target behavior to be predicted or explained, such as held-out responses, choices, ratings, decisions, or structured trait targets.

Our goal is to learn user states $z_i^{(t)} \in \mathcal{Z}$ that summarize persistent behavioral regularities across interactions. In principle, z may be represented as structured traits, free-form summaries, symbolic states, or continuous embeddings. In our experiments, user states are primarily instantiated as structured persona objects serialized as text. For explicit recovery evaluations, these states are parsed into numeric trait vectors for scoring. Appendix D shows the concrete instantiation of (e_i, c_i, x_i, z_i) for the HUMANLM-PROBE.

PERSONA-LR We aim to learn user states that are behaviorally useful, recoverable from observable evidence, and stable across different views of the user (which we operationalize as partial histories).

Conditioning on user history or retrieved memories can improve outputs, but does not by itself guarantee a reusable user representation. PERSONA-LR therefore links inferred user states to held-out behavior while keeping those states explicit and inspectable.

PERSONA-LR couples a behavior model $p_\theta(x \mid z, c)$ with an inference model $q_\phi(z \mid e)$. The behavior model predicts target behavior x from a candidate user state z and task context c, while the inference model proposes user states from observable evidence. Candidate states are evaluated by how well they explain behavior, and the behavior-selected states become targets for inference-model updates.

This separation is necessary because the user state is not directly observed. The behavior model grounds the inference model's proposals by scoring their predictive fit. Repeating this process allows the inferred state itself to improve over refinement rounds.

This resembles EM in that latent states are inferred under a current model and then used to update model parameters. However, classical EM is not directly applicable because our user states may be structured or free-form text, and our objectives include recoverability and cross-history consistency rather than only marginal likelihood. PERSONA-LR is therefore EM-inspired: it proposes candidate states with q_ϕ, reranks them using behavioral likelihood under p_θ, and alternates teacher-forced SFT updates coupled through behavior-selected user-state targets.

Training Objectives PERSONA-LR uses coupled post-training objectives to make the desired properties of the user state explicit during refinement. At evaluation time, the model receives only the allowed public evidence and task context; hidden benchmark factors are withheld. Let $z_i^{(t)}$ denote the user state for example i at refinement round t. Let $f_i^{(t)}$ denote behavior-model feedback available before candidate generation, such as the score or perplexity of the current state. Candidate scores are computed later during reranking. The inference input is $u_i^{(t)} = \left(e_i,\ z_i^{(t)},\ f_i^{(t)}\right)$.

The **behavior model** is trained to predict the observed trajectory conditioned on the current cached user state and task context:

$$\mathcal{L}_{pred} = -\sum_i \log p_\theta(x_i \mid z_i^{(t)}, c_i). \tag{1}$$

This term encourages the cached user state to contain information useful for explaining or predicting held-out behavior. Then, the **inference model** predicts a refinement target from $u_i^{(t)}$:

$$\mathcal{L}_{rec} = -\sum_i \log q_\phi\left(\tilde{z}_i^{(t)} \mid u_i^{(t)}\right). \tag{2}$$

The inference model proposes candidates from $u_i^{(t)}$; the behavior model scores them and selects the recovery target, $\tilde{z}_i^{(t)} = \hat{z}_i^{(t)}$. To encourage consistency across partial histories, we optionally regularize inference outputs from two evidence subsets of the same user:

$$\mathcal{L}_{cons} = \sum_i d\left(q_\phi(\cdot \mid e_i^a), q_\phi(\cdot \mid e_i^b)\right),$$

where e_i^a and e_i^b are different prefixes or subsets of the same user's history.

We also introduce an anchor stabilizer against abrupt drift. The anchor uses the same inference input $u_i^{(t)}$, but targets the pre-refresh cached state:

$$\mathcal{L}_{anchor} = -\sum_i \log q_\phi\left(z_i^{(t)} \mid u_i^{(t)}\right). \tag{3}$$

This is a teacher-forced sequence loss on the serialized user state. When enabled, it encourages preservation of $z_i^{(t)}$ while the recovery loss moves inference toward $\hat{z}_i^{(t)}$.

The behavior and inference models are updated with separate but coupled SFT objectives:

$$\mathcal{L}_\theta = \mathcal{L}_{pred}, \qquad \mathcal{L}_\phi = \mathcal{L}_{rec} + \lambda_{cons}\mathcal{L}_{cons} + \lambda_{anchor}\mathcal{L}_{anchor}. \tag{4}$$

The consistency and anchor terms are auxiliary training-time stabilizers (disabled variants set the corresponding coefficient to zero). Both updates are teacher-forced causal language-model SFT steps. The updates are coupled through the behavior-selected target $\hat{z}_i^{(t)}$.

Behavior-Grounded Persona Refinement The losses in Section 3 define how the behavior and inference models are updated once a target user state is fixed. We still need a procedure for choosing that target state. PERSONA-LR does this through behavior-grounded reranking. Given inference input $u_i^{(t)} = (e_i, z_i^{(t)}, f_i^{(t)})$, the inference model proposes a finite set of candidate user states:

$$C_i^{(t)} = \{z_{i,1}^{(t)}, \ldots, z_{i,K}^{(t)}\}, \qquad z_{i,k}^{(t)} \sim q_\phi(\cdot \mid u_i^{(t)}).$$

These candidates are used only to choose the refinement target $\tilde{z}_i^{(t)}$ for the next inference-model SFT update. Each candidate is scored by a selection criterion:

$$s_i(z) = -\frac{1}{T_i} \log p_\theta(x_i \mid z, c_i) + \alpha r(z)$$

where the first term measures how well the candidate state explains the observed behavior under the current behavior model, and $r(z)$ is a lightweight validity penalty for malformed, overly long, or poorly formatted states. The behavior-selected state is

$$\hat{z}_i = \arg\min_{z \in C_i} s_i(z),$$

and the next inference target is set to $\tilde{z}_i = \hat{z}_i$.

Alternating Refinement Procedure PERSONA-LR uses alternating block updates over the coupled SFT objectives. Before refinement begins, the inference model initializes each cached user state from observable evidence. Each refinement round then first updates the behavior model with a teacher-forced loss on observed behavior conditioned on the current cached user state. The inference model proposes candidate states from the current inference input, and the behavior model reranks them by predictive fit. The selected states become supervised targets for the inference-model SFT step, with optional consistency and anchor terms. After this update, the cache is refreshed with the behavior-selected states for the next refinement round. A detailed implementation-level description of the alternating refinement loop is provided in Appendix B.

CODE RELEASE. Full training code is released anonymously at `https://anonymous.4open.science/r/Persona-LR-Training-0BC5`.

4 EXPERIMENTAL SETUP

We evaluate whether PERSONA-LR improves both recoverable user representations (via HUMANLM-PROBE) and downstream personalization via direct preference prediction (LaMP-3 (Salemi et al., 2024)), long-history personalization (LongLaMP (Kumar et al., 2024)), and real-user histories (PersonaConvBench (Li et al., 2025), PANDORA (Gjurković et al., 2021)).

Benchmarks HUMANLM-PROBE is a diagnostic protocol constructed from HumanLM (Wu et al., 2026), covering 1,500 users with latent persona factors over style, reasoning, task preference, and topic preference. HumanLM persona annotations define hidden factors $z_i^\star$, while interaction records provide the behavioral evidence used to infer them.

We group records by user, map persona fields into a structured latent-state schema, and split interactions into visible history and held-out behavior. We evaluate persona recovery, next-choice prediction, persona updating, and personalized generation.

Baselines We compare PERSONA-LR against Base Qwen3-8B, its underlying backbone (Yang et al., 2025); Qwen3-8B + SFT, which controls for ordinary supervised tuning without behavior-grounded refinement; and HumanLM Base (Wu et al., 2026), a model trained explicitly for human behavior modeling.

We also include similarly sized open-weight instruction models: Llama-3-8B, Mistral-7B, DeepSeek-7B, and Yi-1.5-9B (Grattafiori et al., 2024; Jiang et al., 2023; Bi et al., 2024; Young et al., 2024). Where relevant, we compare against prompt-based personalization and long-context conditioning (Brown et al., 2020; Packer et al., 2023). All methods receive the same histories, task inputs, and prompts, and all initial user states are inferred from observable evidence only.

Evaluation protocol At inference time, all methods receive only the observable user evidence e_i and task context c_i. Benchmark-defined latent user factors $z_i^\star$ are never provided as model inputs and are used only for benchmark construction and post-hoc evaluation. Likewise, held-out evaluation targets are not available to the model at inference time. During PERSONA-LR training, observed behavior may be used to score candidate user states during behavior-grounded refinement, but the targets reserved for final evaluation remain held out. All methods are evaluated under the same evidence and task-context conditions.

Metrics We choose metrics to separate properties that can otherwise be conflated: whether user states are recoverable from observable evidence, stable across partial histories, useful for predicting held-out behavior, supportive of downstream generation quality, and responsive to user evidence.

For explicit persona recovery, models output a structured trait vector $\hat{z}_i$, which is compared against the controlled target vector $z_i^\star$. We report Spearman correlation for rank-order trait recovery, MAE and RMSE for numeric error, and Brier score and ECE for probabilistic calibration.

For stability, we compare states inferred from different partial histories of the same user against states inferred from different users. We report the agreement gap ($\text{Agreement}_{\text{same user}} - \text{Agreement}_{\text{different user}}$) where larger values indicate that partial observations of the same user produce more consistent inferred states than observations from different users.

For behavior prediction, models predict held-out user behavior x_i, such as ratings, choices, or decisions. We report Accuracy for discrete predictions and MAE/RMSE for ordinal or numeric targets. We additionally use correct-state and shuffled-state controls to test whether predictive improvements depend on matching the inferred state to the correct user.

For generation tasks, we report ROUGE-L F1 against reference outputs. For PANDORA, we report mean trait correlation and parse rate, capturing both trait prediction quality and structured-output validity. For history-scaling experiments, we report Δ from no-history to full-history conditions.

5 RESULTS

We evaluate PERSONA-LR across five complementary settings: (i) direct behavior prediction on LaMP-3; (ii) long-history personalization on LongLaMP; (iii) real-user histories on PersonaConvBench and PANDORA; (iv) controlled user-state recovery on HUMANLM-PROBE; and (v) robustness analyses under increasing evidence and bootstrap uncertainty. Additional analyses of robustness, layerwise probing, clustering, and history scaling appear in the Appendix.

5.1 BEHAVIORAL UTILITY ON EXTERNAL BENCHMARKS

We first test held-out behavior prediction on **LaMP-3**. Unlike explicit recovery metrics, this evaluates whether learned user states are behaviorally useful on an external benchmark.

Model	Accuracy↑	MAE↓	RMSE↓
PERSONA-LR	**0.659 ± 0.002**	**0.424 ± 0.006**	**0.745**
Mistral-7B Instruct	0.642	0.426	0.773
Yi-1.5-9B Chat	0.599	0.592	1.062
humanLM Base	0.590	0.509	0.889
PERSONA-LR (SFT Only)	0.578 ± 0.015	0.523 ± 0.007	0.925
Llama-3-8B Instruct	0.525	0.534	0.834
DeepSeek-7B Chat	0.482	0.634	0.976
Base Qwen3-8B	0.449	0.845	1.349

Table 1: LaMP-3 behavior prediction results. For PERSONA-LR and the SFT-only control, we report mean ± standard deviation across three seeds. Other baselines are reported from the original evaluation. Dark gray = best; light gray = second-best.

PERSONA-LR achieves the strongest LaMP-3 behavior prediction, reaching 0.659 ± 0.002 accuracy and 0.424 ± 0.006 MAE across three seeds. Compared with the strongest baseline by accuracy,

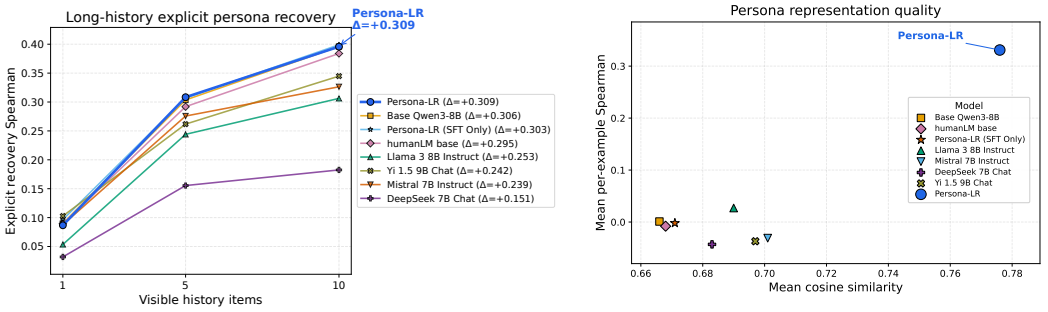

Figure 2: **Left:** Explicit user-state recovery improves as longer histories become available. **Right:** Downstream utility and user-state recoverability are related but distinct; strong task performance does not always imply strong internal user modeling.

Mistral-7B Instruct, accuracy improves from 0.642 to 0.659, showing that more recoverable user states can also improve held-out preference prediction.

5.2 Interpretability on Real-User Personalization Tasks

We next evaluate practical personalization where gold latent traits are unavailable, using long-history tasks and naturally occurring Reddit user histories. In these settings, user-model quality must be inferred indirectly through transfer performance, history scaling, and structured trait prediction.

Model	Abstract(K = 100)	Topic(K = 100)	Review@100	Mean(K = 100)
PERSONA-LR	**0.221**	**0.139**	0.128	**0.166**
Llama-3-8B	0.206	0.130	**0.143**	0.160
Mistral-7B	0.210	0.135	0.126	0.157
DeepSeek-7B	0.209	0.123	0.128	0.153
Qwen3-8B	0.198	0.121	0.133	0.151
PERSONA-LR (SFT Only)	0.198	0.124	0.139	0.150
humanLM Base	0.198	0.124	0.127	0.150

Table 2: LongLaMP ($N = 1000$). Dark gray = best overall. Light gray = second-best per column.

PERSONA-LR achieves the best overall LongLaMP score and leads two of three subtasks. Relative to the SFT-only control, Mean(K = 100) improves by 10.7%, showing that stronger user representations need not reduce generation quality. Figure 2 shows that explicit user-state recovery improves with longer histories, while utility and recoverability remain related but distinct. Figure 3 shows the same pattern at the method level: PERSONA-LR occupies the strongest joint region across utility and recoverability.

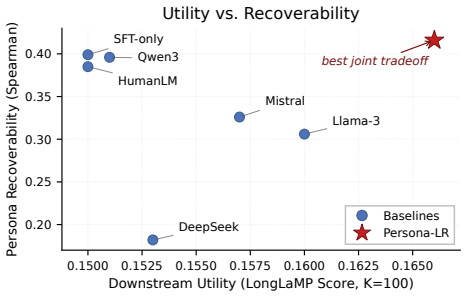

Figure 3: PERSONA-LR gives the best joint tradeoff overall.

PERSONA-LR also leads real-user benchmarks. On PersonaConvBench, larger gains from added history suggest stronger use of accumulated user evidence. On PANDORA, PERSONA-LR attains the highest trait correlation with perfect parse reliability. Across these non-synthetic settings, **PERSONA-LR remains competitive or strongest overall**.

5.3 Controlled User-State Recovery

We finally use **HUMANLM-PROBE** as a controlled diagnostic setting where benchmark-defined latent user factors are known, allowing direct measurement of *explicit user-state recoverability* that is not available on real-user datasets. Although the latent factors are controlled, the observed trajectories are grounded in HumanLM interaction data, reducing reliance on purely templated synthetic behavior. Our goal here is not to interpret the recovered coordinates as

validated psychological measurements, but to test whether observable behavior supports user-state estimates that are recoverable, stable across partial histories, and useful for predicting behavior.

Model	PCB Full	PCB Δ	PANDORA Corr.	Parse
PERSONA-LR	**0.3025**	**+0.1562**	**0.059**	**100**
Mistral-7B	0.2537	+0.1090	0.024	99.5
DeepSeek-7B	0.2333	+0.0859	0.022	81.0
humanLM Base	0.1968	+0.0615	0.047	64.5
PERSONA-LR (SFT Only)	0.1968	+0.0614	0.000	0.0
Llama-3-8B	0.1873	+0.0423	-0.021	5.5
Qwen3-8B	0.1601	+0.0629	0.038	15.0

Table 3: PersonaConvBench + PANDORA. Dark gray = best. Light gray = second-best.

Model	Spearman↑	MAE↓	RMSE↓	Brier↓	ECE↓
Base Qwen3-8B	0.396	0.372	0.472	0.337	0.369
PERSONA-LR (SFT Only)	0.399 ± 0.002	0.341 ± 0.007	0.444 ± 0.002	0.311 ± 0.004	0.349 ± 0.002
PERSONA-LR	$\mathbf{0.416 \pm 0.004}$	$\mathbf{0.304 \pm 0.003}$	0.426 ± 0.002	$\mathbf{0.260 \pm 0.003}$	$\mathbf{0.304 \pm 0.002}$
HumanLM Base	0.385	0.378	0.481	0.342	0.363
Yi-1.5-9B Chat	0.345	0.318	**0.416**	0.274	0.306
Mistral-7B	0.326	0.567	0.647	0.585	0.609
Llama-3-8B	0.306	0.470	0.543	0.453	0.497
DeepSeek-7B	0.182	0.568	0.645	0.584	0.606

Table 4: HUMANLM-PROBE controlled user-state recovery. For PERSONA-LR and the SFT-only control, we report mean ± standard deviation across three seeds. Other baselines are reported from the original evaluation. Dark gray = best overall; light gray = second-best per column.

Recoverability. PERSONA-LR improves over the SFT-only control across all recovery and calibration metrics. Spearman rises from 0.399 ± 0.002 to 0.416 ± 0.004, while MAE drops from 0.341 ± 0.007 to 0.304 ± 0.003. Calibration also improves, with Brier decreasing from 0.311 ± 0.004 to 0.260 ± 0.003 and ECE from 0.349 ± 0.002 to 0.304 ± 0.002.

Dependence on user-specific evidence. A recoverable state should depend on the correct user's history rather than only on global priors or the fixed trait schema. We therefore shuffle histories across users while keeping the evaluation procedure unchanged. Recovery collapses under this intervention: PERSONA-LR's Spearman falls from 0.416 to 0.039, while MAE increases from 0.307 to 0.443. The same pattern holds for the SFT-only control.

Method	Condition	Spearman↑	MAE↓	RMSE↓	Brier↓	ECE↓
PERSONA-LR	Original	0.416	0.307	0.428	0.263	0.306
PERSONA-LR	Shuffled	0.039	0.443	0.543	0.430	0.454
PERSONA-LR (SFT Only)	Original	0.391	0.348	0.444	0.305	0.343
PERSONA-LR (SFT Only)	Shuffled	0.023	0.447	0.543	0.440	0.479

Table 5: Shuffled-history control on HUMANLM-PROBE. Histories are reassigned across users while the trait schema and evaluation procedure remain fixed. The large degradation indicates that recovery depends on matched user evidence rather than schema structure alone.

Stability across partial histories. History scaling alone does not establish that a user-state estimate is stable: recovery could improve with more evidence even if different partial views of the same user induce inconsistent states. We therefore directly compare states inferred from different partial histories of the same user against states inferred from different users. For each user and visible-history budget $B \in \{1, 2, 4, 8\}$, we construct multiple partial views of that user's available history, infer a user state from each view, and compare same-user state similarity against different-user similarity. We define $\mathrm{AgreementGap}(B) = \mathrm{Sim}_{\mathrm{same}}(B) - \mathrm{Sim}_{\mathrm{different}}(B)$ where B is the number of visible history items. A positive gap indicates that two partial histories belonging to the same user yield more similar inferred states than histories belonging to different users. Results are averaged over 10 resampling seeds.

The agreement gap rises from 0.003 at $B = 1$ to 0.563 at $B = 8$, showing that additional user evidence makes partial histories of the same user yield increasingly consistent state estimates. This directly measures the stability component of persona identifiability.

B	Agreement Gap↑
1	0.003
2	0.066
4	0.159
8	0.563

Table 6: Matched partial-history stability for PERSONA-LR across visible-history budgets B, averaged over 10 seeds; higher gaps indicate stronger same-user consistency.

User-specific predictive utility. We next test whether the inferred state is useful specifically because it represents the correct user. We compare behavior prediction under three conditions: no user state, the inferred state for the correct user, and a state shuffled from another user.

For PERSONA-LR, accuracy increases from 0.253 without a state to 0.367 with the correct inferred state, but returns to 0.253 when that state is replaced by another user's state. This indicates that the predictive gain depends on user-specific information encoded in the inferred state rather than merely on supplying an additional structured persona object.

Taken together, these experiments evaluate the three operational components of persona identifiability. The inferred states are *recoverable* from observable user behavior, become increasingly *stable* across partial histories of the same user, and are *predictive* when matched to the correct user's held-out behavior. Importantly, these claims concern explicit, reusable user-state estimates rather than a mechanistically localized hidden representation inside the language model.

Layerwise probing provides complementary evidence about the model's hidden representations: persona-related signal increases from 0.288 at layer 0 to 0.383 near layer 34 (App. A.3). We treat this analysis as supporting evidence about internal representation structure rather than as the basis of the explicit user-state claim.

Method	State	Acc,↑	MAE↓
PERSONA-LR	No state	0.253	0.363
PERSONA-LR	Correct state	**0.367**	**0.342**
PERSONA-LR	Shuffled state	0.253	0.368

Table 7: Correct-state versus shuffled-state behavior prediction. The correct state improves prediction; shuffling removes the gain.

5.4 SCALING AND ANALYSIS

PERSONA-LR improves as more user evidence becomes available, but history scaling alone does not establish that inferred states are stable or user-specific. We therefore complement scaling with matched partial-history, shuffled-history, and shuffled-state controls. Together, these show that recovery improves with more evidence and depends on correctly matched user information rather than schema structure or simply providing a persona object.

Downstream utility and user-state recoverability remain related but distinct. Figure 3 shows that models with similar downstream performance can differ substantially in recoverability, while PERSONA-LR performs strongly on both. The one-shot/no-refresh ablation further shows that repeated behavior-grounded refinement contributes beyond ordinary SFT or single-step pseudo-labeling.

Limitations. HUMANLM-PROBE uses controlled benchmark-defined latent factors, and our conclusions may depend on the chosen user-state representation and do not capture the full complexity of real users. We discuss these limitations, uncertainty estimation, and privacy risks in Appendix A.

6 CONCLUSION

Personalization research has thus far spent much of its evaluation budget on what models say and almost none on what they appear to believe. Our results show that strong personalized outputs do not necessarily imply strong user models. We take a first step to bridge this gap, proposing principled strategies to enhance and study user models. We do not claim to have solved this: our benchmark is synthetic in its latents, and explicit user models raise profiling risks that schema design alone likely won't resolve. We hope, however, that this work encourages a representation-centered perspective on personalization: one that values utility, interpretability, and accountability together.

AI USE STATEMENT

In this work, we used generative AI tools to assist with editing the paper to improve clarity and readability, including rephrasing, grammar checking, and LaTeX formatting. We also used generative AI tools to create and edit software code in several parts of the codebase. We have reviewed all AI-assisted work and verified AI-assisted code for correctness. We take responsibility for the final content of this work, including text, code, and other artifacts produced with the aid of generative AI.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Nalini Ambady and Robert Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological bulletin*, 111(2):256, 1992.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Chris L Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B Tenenbaum. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(4): 0064, 2017.

Yonatan Belinkov and James Glass. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72, 2019.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Itsugun Cho, Dongyang Wang, Ryota Takahashi, and Hiroaki Saito. A personalized dialogue generator with implicit user persona detection. In *Proceedings of the 29th international conference on computational linguistics*, pp. 367–377, 2022.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Chris Frith and Uta Frith. Theory of mind. *Current biology*, 15(17):R644–R645, 2005.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235, 2023.

Matej Gjurković, Vanja M Karan, Iva Vukojević, Mihaela Bošnjak, and Jan Šnajder. Pandora talks: Personality and demographics on reddit. In *Proceedings of the ninth international workshop on natural language processing for social media*, pp. 138–152, 2021.

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

John Hewitt and Christopher D Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138, 2019.

Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939*, 2015.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lelio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothee Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Wang-Cheng Kang and Julian McAuley. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*, pp. 197–206. IEEE, 2018.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra, Alireza Salemi, Ryan A Rossi, Franck Dernoncourt, Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang, Shubham Agarwal, et al. Longlamp: A benchmark for personalized long-form text generation. *arXiv preprint arXiv:2407.11016*, 2024.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and William B Dolan. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 994–1003, 2016.

Li Li, Peilin Cai, Ryan A Rossi, Franck Dernoncourt, Branislav Kveton, Junda Wu, Tong Yu, Linxin Song, Tiankai Yang, Yuehan Qin, et al. A personalized conversational benchmark: Towards simulating personalized conversations. *arXiv preprint arXiv:2505.14106*, 2025.

Veronica Lynn, Youngseo Son, Vivek Kulkarni, Niranjan Balasubramanian, and H Andrew Schwartz. Human centered nlp with user-factor adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1146–1155, 2017.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pp. 11048–11064, 2022.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G Patil, Ion Stoica, and Joseph E Gonzalez. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526, 1978.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*, 2012.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works. *Transactions of the association for computational linguistics*, 8:842–866, 2020.

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

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36:68539–68551, 2023.

Hang Su, Yun Yang, Tianyang Liu, Xin Liu, Peng Pu, and Xuesong Lu. Personalized question answering with user profile generation and compression. *Findings of the Association for Computational Linguistics: EMNLP*, pp. 4744–4763, 2025.

Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. End-to-end memory networks. *arXiv preprint arXiv:1503.08895*, 2015.

Juntao Tan, Liangwei Yang, Zuxin Liu, Zhiwei Liu, Rithesh RN, Tulika Manoj Awalgaonkar, Jianguo Zhang, Weiran Yao, Ming Zhu, Shirley Kokane, et al. Personabench: Evaluating ai models on understanding personal information through accessing (synthetic) private user data. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 878–893, 2025.

Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285, 2011.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 5797–5808, 2019.

Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014.

Shirley Wu, Evelyn Choi, Arpandeep Khatua, Zhanghan Wang, Joy He-Yueya, Tharindu Cyril Weerasooriya, Wei Wei, Diyi Yang, Jure Leskovec, and James Zou. Humanlm: Simulating users with state alignment beats response imitation. *arXiv preprint arXiv:2603.03303*, 2026.

Yunze Xiao, Vivienne J Zhang, Chenghao Yang, Ningshan Ma, Weihao Xuan, and Jen-tse Huang. The chameleon's limit: Investigating persona collapse and homogenization in large language models. *arXiv preprint arXiv:2604.24698*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, et al. Yi: Open foundation models by 01.ai. *arXiv preprint arXiv:2403.04652*, 2024.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 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 (Volume 1: Long Papers)*, pp. 2204–2213, 2018.

Zheng Zhao, Clara Vania, Subhradeep Kayal, Naila Khan, Shay B Cohen, and Emine Yilmaz. Personalens: A benchmark for personalization evaluation in conversational ai assistants. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18023–18055, 2025.

APPENDIX

This appendix reports additional benchmark construction details, robustness checks, representation analyses, qualitative examples, and implementation details for PERSONA-LR. Which is effective personalization requires more than adapting outputs to recent context. Strong personalized systems should instead learn user representations that are predictive of future behavior, recoverable from past evidence, discriminative across users, well calibrated, and responsive to new information.

Beyond the benchmark results in the main paper, we present additional analyses spanning benchmark construction, statistical robustness, layerwise representation structure, user-specific geometry, cross-history refinement, trait-family behavior, downstream personalization under long histories, HumanLM comparisons, failure modes, qualitative examples, and implementation details.

A LIMITATIONS

HUMANLM-PROBE is a diagnostic benchmark and uses benchmark-defined latent factors rather than full human identity. Its controlled design enables direct measurement of persona recoverability, but it does not capture the full ambiguity, fluidity, or social complexity of real users. Recoverability also depends on the chosen user-state schema and representation family; different schemas, continuous embeddings, or free-form summaries may yield different conclusions.

Our real-user evaluations partially address this gap, but they still simplify user modeling. Real users exhibit temporal drift, conflicting preferences, sparse evidence, multilingual behavior, and context-dependent identities that are only partly captured by the benchmarks considered here. In addition, our uncertainty estimates use bootstrap confidence intervals over held-out evaluation examples because repeated full multi-seed PERSONA-LR training runs are computationally expensive. These intervals should therefore be interpreted as evaluation-sample robustness rather than full training-run variance.

Finally, making user representations more explicit can improve auditing and controllability, but it may also increase risks of profiling, sensitive attribute inference, or representation leakage if misused. Future systems should pair representational personalization with privacy safeguards, minimization, access control, and user consent.

B PERSONA-LR ALTERNATING REFINEMENT ALGORITHM

This appendix gives the implementation-level training loop for PERSONA-LR. The algorithm is used during post-training refinement. At evaluation time, the trained models receive only the allowed public evidence and task context; hidden benchmark factors, held-out targets, and training losses are not provided as model inputs.

PERSONA-LR maintains an explicit cached user-state estimate $z_i^{(t)}$ for each training example or user i at refinement round t. The cache is initialized from observable public evidence and then updated through behavior-grounded refinement. Each refinement round contains four main stages: behavior-model SFT, candidate generation by the inference model, behavior-model reranking, and inference-model SFT toward the selected state. The selected state then refreshes the cache for the next round.

Initialization. For each example i, the initial cached user state $z_i^{(0)}$ is decoded from observable public evidence e_i. In the main public-evidence setting, this initialization does not use hidden benchmark factors $z_i^\star$. The hidden factors are reserved for controlled benchmark construction and post-hoc scoring. If oracle or noisy-oracle initialization is used in a diagnostic experiment, it should be reported as a separate diagnostic variant and excluded from the main public-evidence comparisons.

Behavior-model SFT. At the beginning of each refinement round, the behavior model is updated while the cached user states are fixed. The behavior model receives the current cached state $z_i^{(r)}$ and task context c_i, and is trained to predict the observed held-out behavior x_i:

$$\mathcal{L}_{pred} = -\sum_i \log p_\theta(x_i \mid z_i^{(r)}, c_i).$$

Algorithm 1: PERSONA-LR Alternating Refinement

Input: $\mathcal{D} = \{(x_i, c_i, e_i)\}_{i=1}^{N}$; models p_θ, q_ϕ; budget K; rounds R; weights $\lambda_{cons}, \lambda_{anchor}$

Output: Updated θ, ϕ and states $\{z_i^{(R)}\}_{i=1}^{N}$

// Initialize each user state from observable public evidence

for $i = 1$ **to** N **do**

  $z_i^{(0)} \leftarrow \mathrm{Init}_{q_\phi}(e_i)$ // decode initial state from public evidence

end

// Alternate between SFT updates and state refresh

for $r = 0$ **to** $R - 1$ **do**

  // Update behavior model with current cached states fixed

  $\theta \leftarrow \arg\min_\theta \mathcal{L}_{pred}(\{z_i^{(r)}\}_{i=1}^{N})$

  // Propose and behavior-rerank candidate states for each user

  for $i = 1$ **to** N **do**

    $f_i^{(r)} \leftarrow \mathrm{Feedback}_{p_\theta}(x_i, c_i, z_i^{(r)})$ // current-state feedback

    $u_i^{(r)} \leftarrow (e_i, z_i^{(r)}, f_i^{(r)})$ // inference input

    $C_i^{(r)} \leftarrow \{z_{i,k}^{(r)} \sim q_\phi(\cdot \mid u_i^{(r)})\}_{k=1}^{K}$ // candidate states

    $\hat{z}_i^{(r)} \leftarrow \arg\min_{z \in C_i^{(r)}} s_i(z)$ // select state with best behavioral fit

    $\tilde{z}_i^{(r)} \leftarrow \hat{z}_i^{(r)}$ // use selected state as inference target

  end

  // Update inference model toward selected states, with optional stabilizers

  $\phi \leftarrow \arg\min_\phi \mathcal{L}_{rec} + \lambda_{cons}\mathcal{L}_{cons} + \lambda_{anchor}\mathcal{L}_{anchor}$

  // Refresh cached user states for the next round

  $z_i^{(r+1)} \leftarrow \tilde{z}_i^{(r)}$ for all i

end

return $\theta, \phi, \{z_i^{(R)}\}_{i=1}^{N}$

This is a teacher-forced causal language-model objective. The purpose of this step is to make the behavior model sensitive to whether a candidate user state explains the observed behavior.

Feedback construction. After the behavior-model update, the current cached state is evaluated by the behavior model to produce feedback $f_i^{(r)}$. This feedback can include the behavior model's score, likelihood, or perplexity under the current cached state. Candidate-specific behavior scores are not part of $f_i^{(r)}$ at this point; they are computed later during reranking. The inference input is then

$$u_i^{(r)} = (e_i, z_i^{(r)}, f_i^{(r)}).$$

Thus, the inference model sees observable evidence, the previous cached estimate, and feedback about how well that estimate explains behavior.

Candidate generation. The inference model proposes K candidate user-state refinements from the current inference input:

$$C_i^{(r)} = \{z_{i,k}^{(r)} \sim q_\phi(\cdot \mid u_i^{(r)})\}_{k=1}^{K}.$$

These candidates are possible updated persona states for the same user/example i. The cache is maintained separately per user or example, so candidate generation and anchoring do not transfer states across users.

Behavior-grounded reranking. The behavior model scores each candidate by how well it explains the observed behavior:

$$s_i(z) = -\frac{1}{T_i} \log p_\theta(x_i \mid z, c_i) + \alpha r(z),$$

where the first term is the normalized negative log-likelihood of the observed behavior under candidate state z, and $r(z)$ is a lightweight validity or formatting penalty. The behavior-selected state is

$$\hat{z}_i^{(r)} = \arg \min_{z \in C_i^{(r)}} s_i(z).$$

This reranking step is non-differentiable. It is used only to choose the supervised target for the inference-model update.

Inference-model SFT. The selected state becomes the recovery target:

$$\tilde{z}_i^{(r)} = \hat{z}_i^{(r)}.$$

The inference model is then updated by teacher-forced supervised fine-tuning:

$$\mathcal{L}_{rec} = -\sum_i \log q_\phi(\tilde{z}_i^{(r)} \mid u_i^{(r)}).$$

This trains the inference model to produce the user-state estimate that the behavior model selected as most predictive of the observed behavior. The behavior model therefore shapes the inference model through target construction, not through a differentiable gradient path.

Optional stabilizers. The consistency and anchor terms are auxiliary training-time stabilizers. The consistency term encourages compatible inferred states from different partial histories of the same user. The anchor term, when enabled, uses the pre-refresh cached state $z_i^{(r)}$ as an auxiliary teacher-forced target:

$$\mathcal{L}_{anchor} = -\sum_i \log q_\phi(z_i^{(r)} \mid u_i^{(r)}).$$

The anchor target is the previous cached estimate, not the hidden benchmark persona. Intuitively, the recovery loss moves the inference model toward the behavior-selected updated state, while the anchor stabilizer discourages unsupported drift away from the previous cached estimate.

Cache refresh. After the inference-model update, the cache is refreshed:

$$z_i^{(r+1)} \leftarrow \hat{z}_i^{(r)}.$$

The next refinement round uses this selected state as the current cached user state. For a different user or example, the algorithm uses that user's own cache; persona states are not carried across users.

Training versus evaluation. The entire loop above is a training-time refinement procedure. At evaluation time, models are loaded from their trained checkpoints and receive only the allowed public evidence and task context. Hidden benchmark factors and held-out behavior labels are used only for scoring. The training loop does not require reinforcement learning: both model updates are ordinary teacher-forced SFT updates, and reranking is a non-differentiable target-selection step.

C HUMANLM-PROBE CONSTRUCTION, CONTROLS, AND SCOPE

HUMANLM-PROBE is a controlled diagnostic benchmark designed to test whether hidden user structure can be inferred from observed behavioral trajectories. Its purpose is not to simulate the full ambiguity or fluidity of real human identity, but to isolate whether language models can learn latent user states that are predictive of future behavior, recoverable from partial evidence, and reusable across interactions. Although latent evaluation factors are controlled, behavioral trajectories are instantiated using behaviorally grounded data-generation procedures informed by HumanLM-style raw interaction corpora rather than purely hand-authored templates. This reduces reliance on unrealistic templated synthetic users while preserving controlled hidden-state evaluation.

The released benchmark contains 1,500 synthetic users partitioned into 1,224 training users, 120 validation users, and 156 held-out test users. In total, the benchmark contains 6,000 task examples spanning four task families, together with an additional history-scaling variant containing 18,000

examples across visible-history settings $k \in \{1, 5, 10\}$. User splits are disjoint, preventing leakage of user-specific trajectories across train and test evaluation.

Each user is associated with a latent persona represented by 16 continuous traits in $[0, 1]$, organized into four balanced families:

- **Style** (4): communication style and behavioral tendencies,
- **Reasoning** (4): analytical or decision-making preferences,
- **Task Preference** (4): preferred task types or activity patterns,
- **Topic Preference** (4): recurring topical interests.

These latent traits influence generated behavior probabilistically rather than deterministically. As a result, no single interaction fully reveals the hidden persona, requiring models to aggregate evidence across multiple observations rather than rely on isolated cues.

Each public user history contains ten ordered interaction events with timestamps. Tasks expose only prefixes or subsets of these histories, allowing controlled tests of partial observability and evidence accumulation.

Evaluation follows a strict public/private protocol. At inference time, models receive only public histories and task inputs. Hidden personas, private targets, latent constraints, and evaluation labels are withheld and used exclusively for scoring. Supported task targets include persona distribution estimation, persona updating, next-choice prediction, and personalized generation.

To reduce trivial shortcuts, hidden labels are removed from public artifacts, user splits are seeded and deterministic, and additional probe-control settings include shuffled-persona-target and global trait-permutation variants.

HUMANLM-PROBE is intentionally diagnostic and has limitations. The current release is synthetic, interaction length is fixed to ten visible events, and some public histories may contain semantically informative cue fields. Accordingly, results should be interpreted as evidence about identifiable user-state learning under controlled conditions rather than as a complete measure of real-world personalization.

Statistic	Value
Total users	1,500
Train / Dev / Test	1,224 / 120 / 156
Main task examples	6,000
History-scaling examples	18,000
Traits per user schema	16
Trait families	4
Visible events per user	10
History buckets	$k \in \{1, 5, 10\}$

Table 8: HUMANLM-PROBE summary statistics.

D PERSONA STATE PARAMETERIZATION

The PERSONA-LR framework is agnostic to the representation family of user state $z \in \mathcal{Z}$. In principle, z may be implemented as structured traits, free-form summaries, symbolic variables, or continuous embeddings.

In the experiments of this paper, user states are represented as structured key–value persona objects serialized as JSON text. Numeric attributes are normalized to $[0, 1]$, while categorical or descriptive fields are represented as short textual spans. This allows both the behavior model and inference model to consume user states through standard tokenizer inputs without requiring a fixed-dimensional latent vector.

For controlled HUMANLM-PROBE experiments, hidden benchmark factors consist of 16 continuous traits spanning style, reasoning, task preference, and topic preference dimensions. A representative synthetic state is:

```
{"analytical":0.82, "risk_averse":0.74,
"sports_interest":0.11, "verbosity":0.63,
"price_sensitive":0.91}
```

In real-user settings, inferred states are free-form structured summaries derived only from observed histories. A representative inferred state is:

```
{"planning_oriented":0.79, "budget_aware":0.88,
"prefers_direct_answers":0.73, "technical_interest":"high",
"tone":"concise"}
```

During training and inference, these JSON states are serialized into text prompts and consumed autoregressively by both the behavior and inference models.

This separation allows the framework to remain representation-agnostic while enabling concrete implementations across synthetic and real-user benchmarks.

Structured states vs. continuous embeddings. We primarily instantiate user states as structured JSON profiles rather than dense continuous vectors. This choice enables direct inspection, explicit recoverability scoring, schema-level consistency checks, and lightweight reranking during refinement. Continuous embeddings are a natural alternative and may offer stronger compression or smoother optimization, but are less directly interpretable and harder to evaluate under explicit persona-identifiability metrics. Comparing structured and continuous parameterizations is an important direction for future work.

Example instantiation. As an example, suppose a public history prefix contains the turns "Is there a counterexample?" and "Show the math." These turns form part of e_i, the observable evidence used to infer the user state. The corresponding task context c_i specifies that the example is an explicit persona-recovery task, how many history items are visible, and which history indices are exposed. The evaluation target x_i is a structured trait vector, such as high mass on analytical reasoning and math preference. The model prediction is an inferred user state z_i in the same schema, represented as a structured persona object and parsed into numeric traits for evaluation.

D.1 Bootstrap Confidence Intervals and Sources of Variation

We complement the point estimates reported in the main paper with bootstrap 95% confidence intervals computed over per-example evaluation outputs using 1,000 resamples. This analysis measures sensitivity to the composition of the held-out evaluation set and provides an estimate of evaluation-sample uncertainty.

The bootstrap procedure does not capture every source of variation in the full PERSONA-LR training pipeline. In particular, it should not be interpreted as an estimate of unrestricted end-to-end training variance arising from model initialization, stochastic candidate generation, or the trajectory of iterative user-state refinement. We therefore use the bootstrap analysis as a complementary robustness check rather than as a substitute for repeated full retraining.

For the reported main experiments, we fix the public/private splits, prompt templates, decoding and scoring procedures, candidate-set sizes, perturbation scale, and other evaluation settings. Where multiple seeds are reported in the main paper, the corresponding mean and standard deviation reflect the source of randomness varied in that experiment. The bootstrap analysis presented here instead asks a different question: whether the observed pairwise differences remain stable when the held-out evaluation examples are resampled.

Table 9 reports bootstrap estimates of pairwise PERSONA-LR gains against three relevant baselines: the pretrained Base Qwen3-8B model, the SFT-only ablation, and HumanLM Base. For each bootstrap replicate, we resample the held-out HUMANLM-PROBE evaluation examples with replacement and recompute the corresponding metric difference between PERSONA-LR and the comparator. We report the mean pairwise gain together with the empirical 95% bootstrap interval.

The MAE results are the clearest under this analysis. PERSONA-LR improves MAE relative to all three comparators, and the corresponding bootstrap intervals remain strictly positive. Relative to Base

Comparator	HUMANLM-PROBE Spearman Gain↑	HUMANLM-PROBE MAE Gain↑	Downstream Summary
Base Qwen3-8B	$+0.0042\,[-0.0111, 0.0113]$	$+0.0456\,[0.0383, 0.0528]$	LongLaMP $+0.015$; PCB $+0.1424$
PERSONA-LR (SFT Only)	$+0.0107\,[0.0006, 0.0222]$	$+0.0255\,[0.0187, 0.0317]$	LongLaMP $+0.016$; PCB $+0.1057$
HumanLM Base	$+0.0118\,[0.0018, 0.0228]$	$+0.0513\,[0.0448, 0.0576]$	LongLaMP $+0.016$; PCB $+0.1057$

Table 9: Bootstrap 95% confidence intervals for pairwise PERSONA-LR gains. Positive values indicate improvement; for MAE, gain is defined as comparator minus PERSONA-LR. Spearman and MAE gains are computed over 1,000 bootstrap resamples of the held-out HUMANLM-PROBE evaluation set. Downstream values summarize point-estimate differences on LongLaMP and PersonaConvBench and are included for context rather than as bootstrap confidence intervals.

Qwen3-8B, the estimated MAE gain is $+0.0456$ with a 95% bootstrap interval of $[0.0383, 0.0528]$. Relative to the SFT-only control, the gain is $+0.0255$ with interval $[0.0187, 0.0317]$, while the comparison with HumanLM Base yields a gain of $+0.0513$ with interval $[0.0448, 0.0576]$.

The Spearman results are more mixed and should be interpreted accordingly. PERSONA-LR has positive pairwise gains against all three comparators. The bootstrap intervals exclude zero relative to the SFT-only control and HumanLM Base, but the comparison with Base Qwen3-8B overlaps zero. Thus, the strongest statistical evidence concerns improved numeric recovery and calibration rather than a uniform improvement in every notion of user-state recovery.

The downstream columns provide complementary evidence that improved recoverability is not obtained by sacrificing personalization utility. PERSONA-LR maintains or improves performance relative to the same comparators on LongLaMP and PersonaConvBench. These entries are reported as point-estimate differences and should not be interpreted as bootstrap intervals.

Taken together, the bootstrap analysis supports a deliberately narrow conclusion: the principal HUMANLM-PROBE recovery improvements are robust to resampling of the held-out evaluation examples, particularly for MAE. The analysis does not establish complete robustness to all sources of optimization or training stochasticity, and we therefore avoid treating these confidence intervals as estimates of full end-to-end training variance.

E USER-SPECIFIC STRUCTURE OF PERSONA REPRESENTATIONS

A useful persona representation should encode persistent user-specific structure rather than collapse toward generic population averages. If latent user states are meaningful, representations inferred from different histories of the same user should remain more similar to one another than representations inferred from unrelated users. Conversely, if representations mainly capture population-level averages or superficial task artifacts, same-user and different-user similarities should be nearly indistinguishable. A useful user representation should preserve user identity across histories.

To test this property, we compare cosine similarity between inferred persona vectors drawn from the same user versus vectors drawn from different users. Across all evaluated systems, same-user pairs are consistently more similar than mismatched-user pairs, indicating that user identity information is recoverable from latent states. This provides evidence that even standard pretrained models contain measurable user-specific signal.

However, the magnitude of this separation differs substantially across methods. Base Qwen3-8B achieves a same-user versus different-user gap of 0.073, while PERSONA-LR (SFT Only) modestly improves this to 0.077. HumanLM Base obtains a similar gap of 0.075, suggesting that behavioral personalization alone does not strongly sharpen identity structure.

PERSONA-LR Full achieves the largest separation gap at 0.116, substantially exceeding all baselines. Although its absolute cosine similarities are lower overall, the stronger relative gap indicates better discrimination between users rather than simple representational collapse toward universally high similarity values. In other words, PERSONA-LR appears to organize user representations more selectively, making different users easier to distinguish while preserving within-user consistency.

This result complements the explicit recovery findings in Appendix H. Stronger same-user clustering alone would be insufficient if persona traits were inaccurate, while accurate traits without user separation would suggest weak identity structure. PERSONA-LR improves both, indicating gains in the quality and organization of latent user representations.

Taken together, these findings suggest that PERSONA-LR does not merely produce better outputs or trait estimates; it learns more user-conditioned internal states.

Model	Same-User	Different-User	Gap
Base Qwen3-8B	0.783	0.711	0.073
PERSONA-LR (SFT Only)	0.822	0.744	0.077
PERSONA-LR Full	**0.674**	**0.558**	**0.116**
HumanLM Base	0.848	0.674	0.075

Table 10: Similarity of persona vectors for the same user versus different users. Larger gaps indicate stronger user-specific structure.

F STABILITY–PLASTICITY ACROSS INCREASING HISTORY LENGTH

Effective user modeling requires balancing two competing objectives: *stability* and *plasticity*. A useful persona representation should preserve persistent user characteristics across interactions, while also updating when genuinely informative new evidence becomes available. Excessive stability leads to rigid user states that ignore new observations, whereas excessive plasticity leads to noisy drift and poor long-horizon consistency.

To study this tradeoff, we compare inferred persona vectors as progressively longer user histories become available, measuring cosine similarity between representations estimated from short ($k = 1$), medium ($k = 5$), and longer ($k = 10$) evidence windows. Very high similarity suggests little representational change as additional history is observed, while lower similarity can indicate either meaningful refinement or instability.

Base Qwen3-8B and PERSONA-LR (SFT Only) remain highly similar across evidence lengths, suggesting comparatively static user representations that update only modestly when more evidence is provided. In contrast, PERSONA-LR Full exhibits lower cross-history similarity across all comparisons, including $\text{Sim}(k1, k10) = 0.625$ versus 0.757 for Base Qwen3. Importantly, this greater movement coincides with stronger explicit recovery, sharper trait estimates, and better same-user separation (Sections A.2–A.4). We therefore interpret these changes not as uncontrolled drift, but as evidence-responsive refinement of user states.

These findings suggest that PERSONA-LR achieves a healthier stability–plasticity balance: representations remain user-specific while still adapting substantially when richer histories become available.

Model	$\text{Sim}(k1, k5)$	$\text{Sim}(k5, k10)$	$\text{Sim}(k1, k10)$
Base Qwen3-8B	0.756	0.838	0.757
PERSONA-LR (SFT Only)	0.801	0.864	0.800
PERSONA-LR Full	**0.634**	**0.764**	**0.625**
HumanLM Base	0.814	0.813	0.718

Table 11: Cross-history cosine similarity of inferred persona vectors across increasing evidence lengths. Lower similarity can indicate meaningful refinement when accompanied by stronger recovery and preserved user structure.

G TRAIT-FAMILY ANALYSIS

We group benchmark traits into four semantic families: *style*, *reasoning*, *task preference*, and *topic preference*. This analysis tests whether improvements are concentrated on shallow surface attributes or extend to deeper behavioral dimensions.

PERSONA-LR Full improves over Base Qwen3 across every trait family. The largest reductions in MAE occur on style traits ($0.453 \rightarrow 0.390$), followed by task preference ($0.351 \rightarrow 0.301$), reasoning ($0.384 \rightarrow 0.337$), and topic preference ($0.301 \rightarrow 0.279$). Gains on style and reasoning are particularly notable because these dimensions are often expressed indirectly through behavior rather than explicit lexical cues.

Compared with HumanLM Base, PERSONA-LR also improves consistently across all four groups, suggesting that the benefit is not limited to one narrow trait category. This broad improvement indicates that PERSONA-LR strengthens general user modeling rather than overfitting to a specific subset of persona attributes.

Overall, the results suggest that PERSONA-LR improves deeper behavioral characteristics more strongly than shallow topical descriptors, consistent with the paper's emphasis on representational personalization.

Model	Style	Reasoning	Task	Topic
Base Qwen3-8B	0.453	0.384	0.351	0.301
PERSONA-LR (SFT Only)	0.431	0.360	0.334	0.283
PERSONA-LR Full	**0.390**	**0.337**	**0.301**	**0.279**
HumanLM Base	0.466	0.385	0.355	0.305

Table 12: Trait-family MAE (lower is better). PERSONA-LR Full improves all trait groups, with strongest gains on style and task dimensions.

H REPRESENTATION GEOMETRY ACROSS MODELS

We additionally compare hidden-state geometry across models to test whether internal representations align with persona structure. If user information is meaningfully organized in representation space, nearby hidden states should correspond to similar personas and lower reconstruction error.

PERSONA-LR LM1 achieves the strongest results across all reported metrics. It is the only model with positive hidden-persona correlation ($\rho = 0.036$), obtains the best neighbor match rate (0.111), the highest mean cosine similarity (0.776), and the lowest MAE (0.255). By contrast, several baselines exhibit negative correlation values, suggesting weaker alignment between hidden geometry and user-state structure.

The gains are notable because these metrics are derived from internal representations rather than output behavior. Thus, improvements cannot be explained solely by better prompting or decoding; they indicate that PERSONA-LR changes how user information is encoded internally.

Taken together, the results support the claim that PERSONA-LR improves not only explicit recovery metrics, but also the latent geometry of user representations.

Model	Hidden-Persona ρ	Neighbor Match	Mean Cosine	MAE
Base Qwen3	-0.038	0.056	0.666	0.319
HumanLM Base	-0.038	0.000	0.668	0.319
HumanLM SFT	-0.083	0.000	0.671	0.317
PERSONA-LR LM1	**0.036**	**0.111**	**0.776**	**0.255**
Llama-3-8B	-0.213	0.111	0.690	0.308
Mistral-7B	-0.121	0.000	0.701	0.305

Table 13: Representation geometry comparison. PERSONA-LR LM1 exhibits the strongest alignment between hidden states and persona structure.

I LAMP-3 DIRECT BEHAVIOR PREDICTION

We additionally evaluate whether learned user states improve direct forecasting of future user preferences using **LaMP-3**, a standardized profile-grounded task from the LaMP benchmark suite (Salemi et al., 2024). In our implementation, the benchmark is executed under the pipeline key `lamp3`, with processed rows tagged as `lamp3_rating`. This benchmark complements HUMANLM-PROBE by measuring practical behavioral prediction rather than only latent recoverability.

We use the processed dataset located at `data/processed/lamp3`, constructed from the `LongLaMP/LongLaMP` HuggingFace subset `product_review_user`. The released split contains 400 training examples, 200 validation examples, and 1000 test examples after filtering and canonicalization. Each test example corresponds to a distinct user, with an average observed review-history length of 14.9 prior interactions (minimum 4, maximum 20).

The task is to predict a held-out *1–5 star product rating* for a new review conditioned on prior user review history. Targets are integer labels in $\{1, 2, 3, 4, 5\}$. Inputs consist of the new product review together with up to the most recent 12 historical user reviews, where each history item includes a product summary, prior rating, and short review text.

Unlike free-form generation tasks, LaMP-3 is evaluated via constrained next-token scoring. Models are prompted to return exactly one token from $\{1, 2, 3, 4, 5\}$, and final predictions are selected by softmax-normalized next-token logits over those five labels. No sampling or beam search is used.

We report three complementary metrics: Accuracy, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). Ratings are parsed as integers and evaluated arithmetically. Accuracy measures exact star prediction, while MAE and RMSE capture ordinal distance between predicted and true preferences.

Model	Accuracy↑	MAE↓	RMSE↓
PERSONA-LR	**0.659**	**0.419**	**0.745**
Mistral-7B Instruct	0.642	0.426	0.773
Yi-1.5-9B Chat	0.599	0.592	1.062
humanLM Base	0.590	0.509	0.889
PERSONA-LR (SFT Only)	0.580	0.530	0.925
Llama-3-8B Instruct	0.525	0.534	0.834
DeepSeek-7B Chat	0.482	0.634	0.976
Base Qwen3-8B	0.449	0.845	1.349

Table 14: LaMP-3 direct behavior prediction benchmark. Higher Accuracy and lower MAE/RMSE are better.

PERSONA-LR achieves the strongest overall performance across all reported metrics. Relative to the strongest competing baseline (Mistral-7B Instruct), Accuracy improves from 0.642 to 0.659 while RMSE decreases from 0.773 to 0.745. These findings support the main paper's claim that optimizing identifiable user states improves not only recoverability, but also practical forecasting of future user preferences on a recognized external benchmark.

All evaluated models use the same processed test set, identical prompting protocol, and the same deterministic next-token scoring procedure. This ensures that differences primarily reflect model representations rather than decoding stochasticity or benchmark-specific prompt variation.

I.1 FAILURE ANALYSIS

We manually inspect common error modes to understand where persona inference systems fail. Two recurring categories are unsupported persona fabrication and structured-output formatting failures.

Unsupported persona fabrication occurs when a model confidently predicts user traits that are weakly supported or contradicted by available history. PERSONA-LR substantially reduces this behavior relative to Base Qwen3, suggesting that its latent states are more tightly grounded in observed evidence rather than relying on generic stereotypes or overconfident priors.

A second category involves formatting errors for structured outputs (e.g., malformed JSON, inconsistent fields, or incomplete trait specifications). PERSONA-LR exhibits only a modest increase in such errors despite producing more informative user states. This indicates that gains in representational quality are not driven by sacrificing output reliability.

These findings are encouraging because stronger personalization systems should ideally improve both accuracy and calibration. A model that predicts richer personas but hallucinates unsupported traits would be risky in practice. PERSONA-LR instead appears to improve factual discipline while strengthening user modeling.

Future work should continue evaluating fairness, uncertainty calibration, and harmful overgeneralization in personalized systems, especially when deployed on real user data.

J Qualitative Persona Examples

To complement aggregate metrics, we show representative examples where models receive the same user history and must infer a reusable user state for predicting held-out behavior. We compare a generic baseline summary against PERSONA-LR. Held-out gains are measured on downstream preference prediction for the corresponding users.

User History	Baseline Summary	PERSONA-LR Summary	Held-out Gain
Finance / planning posts	finance, taxes, shopping, money	analytical, budget-conscious, risk-aware, long-horizon planner who values efficiency, structured tradeoff analysis, and conservative decision making	+12.4% acc.
Film / review posts	movies, tv, reviews, entertainment	values character depth, dislikes clichés, prefers slower pacing, thoughtful narratives, and coherent world-building over spectacle	+9.1% acc.
Technology / coding posts	programming, code, software, computers	systematic problem solver who values clean abstractions, practical debugging steps, efficiency, and direct technical explanations	+10.7% acc.
Fitness / nutrition posts	gym, exercise, diet, health	goal-oriented, consistency-focused user who prefers sustainable routines, measurable progress, and evidence-based recommendations over fads	+8.6% acc.
Travel / lifestyle posts	travel, hotels, food, cities	experience-driven planner who values convenience, local authenticity, budget efficiency, and minimizing logistical friction	+7.9% acc.

Table 15: Representative user-state summaries. Baseline outputs often recover surface topics, while PERSONA-LR more frequently infers decision-relevant preferences and behavioral tendencies that transfer to held-out behavior prediction. Gains denote downstream preference prediction improvement for these users relative to the baseline summary condition.

Across examples, the baseline primarily extracts topical keywords from history, whereas PERSONA-LR recovers higher-level behavioral tendencies that better explain future choices. This mirrors the main quantitative findings: stronger recoverability is associated with improved downstream personalization rather than merely longer or more descriptive summaries.

K Training Cost and Reproducibility

Training used:

- 2500 EM refinement cycles
- 8 $\times$ 48GB GPUs
- bf16 mixed precision
- $\sim$153 hours wall-clock runtime
- $\sim$1224 GPU-hours total compute

Release artifact. The anonymous release artifact includes the PERSONA-LR training implementation and instructions for training the proposed alternating refinement procedure using HumanLM-derived interaction data. The paper and appendix provide the corresponding benchmark construction details, data splits, model baselines, metrics, hyperparameters, compute requirements, and baseline provenance needed to interpret the reported results. External datasets, model checkpoints, and assets whose licenses restrict redistribution are not repackaged; instead, the artifact provides source references and training instructions for users who obtain the required data from the original sources.

L HYPERPARAMETER AND COMPUTE NOTES

PERSONA-LR uses a single default configuration across all main-paper benchmarks. In our implementation, candidate count is controlled by the number of retained LM2 proposals, with small candidate sets used during refinement. Training used 2500 EM refinement cycles on 8×48GB GPUs with bf16 mixed precision, requiring approximately 153 wall-clock hours and 1224 GPU-hours.

Several hyperparameters affect refinement: the number of candidate states K, the LM2 target-loss weight, the consistency or anchoring weight when enabled, and the behavior-reranking penalty. These are modified one at a time in our sweep manifest. Full retraining is required for isolated sensitivity rows because these parameters affect training-time state refresh or LM2 supervision; archived per-candidate outputs were not available for the final reported run. We therefore report the default configuration in the main paper and provide the sweep script and manifest for reproducibility.

Initialization perturbations. To avoid trivial persistence of anchor persona states, initialized personas are perturbed prior to refinement. For normalized numeric persona traits, we apply independent Gaussian noise and clip values to the valid range [0,1]. Unless otherwise stated, the default standard deviation is 0.12. Noise seeds are deterministic per example for reproducibility. Non-numeric fields are left unchanged.

M PERSONA INITIALIZATION AND TRAINING DYNAMICS

PERSONA-LR maintains an explicit user-state assignment for each training example during iterative refinement. Because latent persona variables are not directly observed in most real-user datasets, initial persona states must be constructed from available evidence prior to optimization.

Artifact release scope. We release the PERSONA-LR training implementation and configuration structure to support inspection and procedural reproduction of the refinement framework. Full trained checkpoints, complete cached persona trajectories, and all intermediate training artifacts are not included in the initial release.

Initialization sources and leakage control. For all main reported PERSONA-LR results in Tables 1–4, initial persona states are inferred only from observable public evidence, including user histories, prior interactions, and evidence-derived structured summaries. Hidden benchmark persona factors are never used to initialize, condition, train, rerank, or evaluate the main PERSONA-LR comparisons. In controlled benchmark settings, target-aligned or noisy-oracle initializations are used only in separately labeled diagnostic refinement experiments designed to study refinement dynamics under known latent anchors. These diagnostic runs are excluded from all main benchmark comparisons, ablations, and headline results. Thus, the main PERSONA-LR results follow the same public/private protocol as all baselines: models receive only public histories and task inputs, while hidden personas remain withheld and are used exclusively for scoring.

Initialization perturbations. Exact initialization at an evidence-derived anchor can lead to trivial persistence, where subsequent refinement merely reproduces the starting assignment rather than improving it. To avoid this failure mode, we apply small stochastic perturbations to initialized persona states prior to refinement. For normalized numeric persona traits, we add independent Gaussian noise and clip values to the valid range:

$$z^{(0)} = \mathrm{clip}(z_{\mathrm{evidence}} + \epsilon, 0, 1), \qquad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

Unless otherwise stated, σ is fixed across experiments. Non-numeric structured fields are left unchanged.

Training dynamics. After initialization, PERSONA-LR alternates between persona inference and behavior modeling. Given user evidence e_i, the inference model (LM2) proposes a small candidate set of persona states:

$$C_i = \{z_i^{(1)}, \ldots, z_i^{(K)}\}, \qquad z_i^{(k)} \sim q_\phi(\cdot \mid e_i).$$

The behavior model (LM1) evaluates each candidate using downstream prediction loss on the observed target behavior x_i under context c_i:

$$s(z) = -\log p_\theta(x_i \mid z, c_i) + \alpha r(z),$$

where $r(z)$ is a lightweight validity or complexity penalty. The highest-scoring candidate is selected as the current latent assignment:

$$\hat{z}_i = \arg\min_{z \in C_i} s(z).$$

Parameter updates. LM1 is optimized to better predict observed user behavior conditioned on the selected persona state:

$$\mathcal{L}_{\text{pred}} = -\log p_\theta(x_i \mid \hat{z}_i, c_i).$$

LM2 is optimized to recover behaviorally useful persona assignments from user evidence:

$$\mathcal{L}_{\text{rec}} = -\log q_\phi(\hat{z}_i \mid e_i).$$

When enabled, we additionally apply cross-history consistency regularization so that different evidence subsets from the same user induce similar persona states.

Interpretation. This procedure allows PERSONA-LR to learn user states from behavioral supervision even when explicit persona labels are unavailable. Initialization provides a stable starting point, stochastic perturbations encourage non-trivial refinement, and alternating updates gradually improve both user-state quality and downstream personalization performance.

N BASELINE PROVENANCE

HumanLM baseline provenance. For HumanLM comparisons, we use the publicly released HumanLM ecosystem described by (Wu et al., 2026). Unless otherwise noted, experiments use the authors' released checkpoint `snap-stanford/humanlm-opinion`, a Qwen3-8B based model distributed through the Stanford SNAP Hugging Face repository. This benchmark and model family were introduced as part of HumanLM: *Simulating Users with State Alignment Beats Response Imitation*. HumanLM is particularly relevant because it emphasizes behavioral simulation, providing a strong contrast between output-level imitation and our representation-centered objective. The associated Humanual benchmark suite contains six public user-modeling datasets spanning approximately 26k users and 216k responses across multiple domains, including opinion forums, books, email, politics, dialogue, and news.

Other baselines. All remaining open-model baselines are evaluated using publicly released checkpoints from their official model providers or standard Hugging Face distributions. Unless explicitly stated, we do not apply additional task-specific retraining beyond the evaluation protocols described in the main text.