Final Report: Monitoring and Attributing Implicit Personalization in Conversational Agents

Jacopo Zacchigna
University of Trieste
jacopo.zacchigna@studenti.units.it

Hengxu Li
Tufts University
Hengxu.Li@tufts.edu

Vinit Patel
Northeastern University
vinitp0310@gmail.com

Gabriele Sarti
Northeastern University
gabriele.sarti996@gmail.com

Supervised Program for Alignment Research — Spring 2026

Abstract

Language models implicitly infer user attributes that shape their responses in ways users often neither expect nor endorse. This implicit personalization mediates concerning behaviors such as sycophancy, deception, and demographic bias. We investigate implicit user modeling through mechanistic interpretability, focusing on three goals: measuring the prevalence of personalized behaviors driven by implicit cues, analyzing how implicit user information is represented internally, and evaluating whether such representations can be controlled via editing and steering techniques. We construct realistic narrative biographies using a structured interview protocol conditioned on AgentBank synthetic persona templates, then generate implicit question–answer pairs answerable from each biography. Behaviorally, we compare how templated and narrative attributes elicit different levels of personalized behavior; mechanistically, we test whether latent representations of personas obtained via various steering methods preserve meaningful individual differences. To facilitate such work, we develop a visual interface for interacting with predefined personas and evaluating steering and probing methods.

We found steering via ActAdd to be weak compared to in-context conditioning, with only the education axis showing consistent directional transfer to the no-context implicit setting at pilot scale. Moreover we found a significant difference in both in-context learning and LoRA-based steering when using templates attributes vs biographies: models conditioned on narrative biographies consistently achieve higher accuracy on both explicit and implicit persona questions, with the gap largest on explicit questions (81.4% vs. 75.2% for Qwen3-4B in-context) and persisting, though narrowing, on implicit ones.

Keywords Implicit Personalization · User Modeling · Interpretability · Human-AI Interaction

1 Introduction

When language model (LM)-based conversational agents interact with users, they implicitly infer user attributes such as expertise level, demographics, and beliefs [5, 8]. This phenomenon, known as *implicit personalization*, was observed in LLMs even in the absence of explicit signals, due to subtle associations with conversational style, vocabulary, stated preferences, and contextual cues. While such adaptation can improve user experience, it can also manifest in ways users neither expect nor

endorse, leading to harmful behaviors including sycophancy, deception, and demographic bias [12]. For example, response quality can be degraded for users implicitly associated with minority groups [8]. As models become more capable, these behaviors become subtler and harder to detect through output inspection alone, raising the stakes for understanding and controlling the internal mechanisms that produce them. Prior work on implicit personalization generally relied on behavioral evaluations with explicit system-prompt manipulations targeting a restricted set of demographic attributes [5, 8]. Recent advances in mechanistic interpretability have opened a complementary avenue, enabling the verbalization of complex user models directly from model activations [3]. However, reliably steering model behavior through interventions has proven substantially harder than decoding latent beliefs [3]. Bridging this gap requires richer evaluation resources, a deeper understanding of how persona information is distributed across model components, and practical tools for experimentation.

This project advances the study of implicit personalization in LLMs and its detection and control using white-box techniques that leverage model activations. First, we move beyond isolated demographic attributes toward *rich, realistic personas* as the unit of analysis. Using a structured interview protocol conditioned on synthetic persona templates from AgentBank [10], we construct detailed narrative biographies for a large set of simulated individuals, then generate both explicit and implicit question-answer pairs grounded in each biography, finally an additional set of multiple choice question-answer shared across all of the personas are created. This resource, SynthPersona, enables systematic comparison of how templated attribute lists and narrative contexts elicit different levels and kinds of personalized behavior, and provides the provenance annotations needed for fine-grained causal attribution of which persona elements drive specific model inferences. Using SynthPersona, we investigate how persona information is represented in model activations and whether these representations can be controlled by means of various steering approaches, with the aim of testing whether model activations can support effective behavioral steering for rich personas, and identifying gaps with current methods to inform the development of reliable intervention techniques in this domain. To support rapid experimentation within this setup, we develop an interactive visual interface that integrates persona extraction, analyses, steering, and probing into a unified interface, drawing on prior work on transparency dashboards for attribute-controlled LLM conversations [2, 3].

Together, these efforts contribute to AI safety by developing methods to detect when models engage in potentially problematic personalization, characterizing which internal mechanisms support such behaviors, and creating resources to foster further research on implicit personalization in realistic, multi-turn settings.

1.1 Related Work

1.1.1 Implicit Personalization in Language Models

A growing body of work has documented that LMs adjust their behavior based on inferred user attributes, even when those attributes are not explicitly provided. Jin et al. conducted a systematic study of implicit personalization, showing that models alter response style, content, and quality based on demographic cues embedded in user prompts [5]. Neplenbroek et al. demonstrated that stereotypes shape LM responses through implicit inference, with models producing lower-quality or more stereotyped outputs for users associated with minority groups [8]. Sharma et al. characterized sycophancy as a specific failure mode of implicit user modeling, in which models prioritize agreement with perceived user beliefs over truthfulness [12]. These studies establish that implicit personalization is pervasive, but they rely primarily on behavioral evaluations with

controlled prompt manipulations targeting individual demographic dimensions such as gender, age, or political leaning. In this work, we expand such analyses to rich, multi-dimensional personas derived from structured interview protocols.

1.1.2 Synthetic Personas and Behavioral Simulation

Simulating realistic human personas with LMs has emerged as a methodology for studying social behavior at scale. Park et al. introduced generative agent simulations of 1,000 people, demonstrating that LMs conditioned on detailed interview transcripts can reproduce individual-level attitudes and behavioral tendencies with notable fidelity [10]. Kolluri et al. showed that fine-tuned LMs can predict human behavior in social science experiments, suggesting that models internalize coherent person-level representations [6]. Our SynthPersona dataset builds on this line of work by generating rich narrative biographies from structured interview protocols and comparing their effects on user models with those of traditional templated attributes for both explicit and implicit QA pairs.

1.1.3 Latent User Representations and Mechanistic Interpretability

Recent work has begun to bridge the gap between behavioral observation and internal mechanism. Pan et al. introduced LatentQA, a framework for verbalizing latent user representations by training a decoder probe on model activations, demonstrating that a single middle layer can be decoded to recover inferred user attributes with high accuracy [9]. However, recent work applying similar techniques to study user models in LLMs has shown that information is readily *decodable* from individual activations, but *steering* model behavior requires broader interventions spanning multiple layers. This suggests that user-model representations are distributed across the network and may involve distinct encoding and output pathways. Chen et al. explored how such latent representations can be surfaced to users through transparency dashboards, proposing interface designs that expose the model's inferred user attributes in real time [2]. Our persona vector UI builds on this vision, extending it to support inspection, interactive steering and probing analyses.

2 Datasets

2.1 SynthPersona Dataset

SynthPersona is our main resource for studying implicit personalization under rich persona context. Starting from AgentBank-style templated personas [11], we run a structured interview to elicit a larger set of user facts from an LLM, maintaining previous interview turns in-context to maximize consistency. We then convert the transcript into three views of the same persona: a templated attribute list, a set of statements mapped to specific interview turns, and a narrative biography. For each profile, we also generate implicit multiple-choice questions and link them to their supporting statements, so the same item can later be rendered either as a full biography or as a statement-level ablation target. For example, the implicit statements:

You recently suffered a back injury and have been recovering slowly.
You prefer dealing with difficulties privately.

are used to produce the implicit multiple-choice question:

Would you be interested to join a support group for people with similar injuries?

A. I would like to join to exchange advice and experiences.
B. I'd rather handle this with close family rather than strangers.
C. I'd likely seek more information about the group first.

D. I would join if a doctor recommended it.
E. Not sure

In this example, the model prediction should account for the second statement, making the model lean towards choice B. Given the collection of personas and the related questions, we then focus on two dimensions for subsequent analyses. Behaviorally, we compare *bare*, *templated*, *biography_full*, and *swapped_biography* contexts to test whether persona conditioning changes the model's choice. In the *bare* condition, the model receives no persona context. In the *templated* condition, the system prompt contains the persona's structured attribute list. In the *biography* condition, the full narrative biography is provided. In the *swapped biography* condition, the biography of a different persona is substituted, establishing a population-level baseline for each question.

Mechanistically, we measure how ablating evidence-bearing statements-*full*, *corrupted*, and *control*-impacts probing results, and whether this effect persists when scaling from individual statements to full biographical contexts. The *full* condition includes all biography statements; the *corrupted* condition removes the statements that directly support the expected answer; and the *control* condition removes an equal number of unrelated statements. If persona context is genuinely causal, removing supporting evidence should produce a larger behavioral shift than removing irrelevant content. We report both top-1 answer changes and mean log-probability differences across conditions. Figure **??** summarizes this end-to-end generation pipeline.

During dataset development, we also compared several candidate generator models for the intermediate interview, statement, and biography assets: `Qwen3-235B-A22B-Instruct`, `minimax-m2-her`, `Kimi-K2.5`, `DeepSeek-V3.1`, and `Qwen3.5-397B-A17B`. These models exhibited distinct tradeoffs. `Kimi-K2.5` produced the richest and most vivid biographies, but it more often introduced extra narrative details and named entities that made downstream evidence attribution less clean. `DeepSeek-V3.1` produced the cleanest and most structurally regular assets, but often with flatter biographies and less narrative richness. `Qwen3.5-397B-A17B` was fluent but more prone to compressing biographies into fewer sections while confidently adding unsupported specifics. `MiniMax M2-her` showed unstable persona adherence already in the interview stage. We therefore use `Qwen3-235B-A22B-Instruct` as the default SynthPersona generator, since it offered the best balance between richness, faithfulness to the provided persona attributes, and downstream suitability for the statement-level causal analyses.

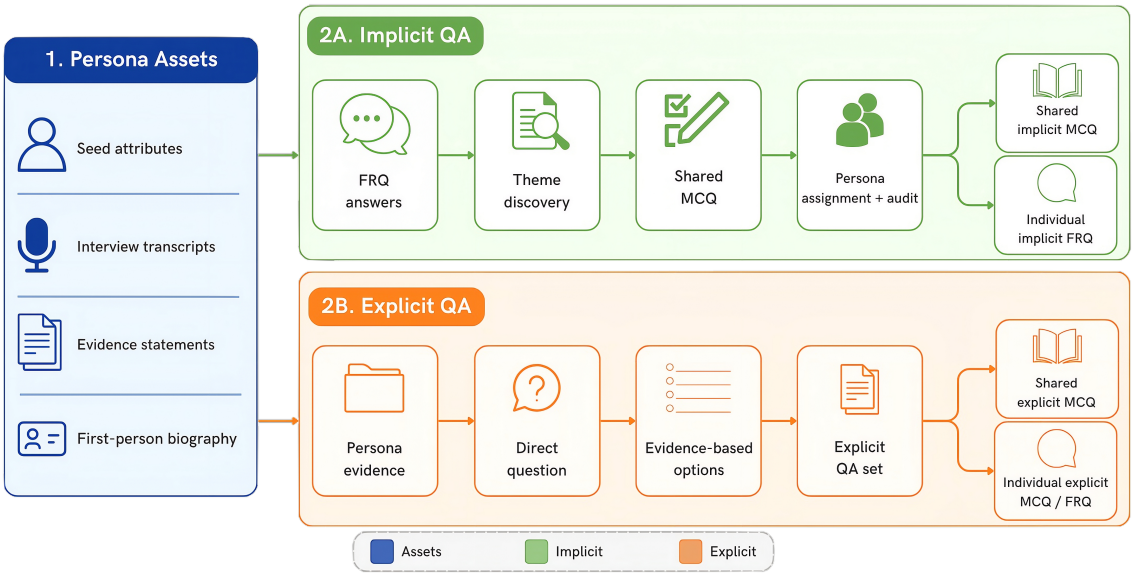


Figure 1: Overview of the SynthPersona dataset generation pipeline.

3 Methodology

3.1 Persona Vectors

3.1.1 Extraction Pipeline

For each persona, we provide the model with the persona's biography or attribute list as the system prompt and sequentially input each QA pair, recording residual-stream activations at configurable token positions for each layer. The implementation in our persona-vector extraction library uses **ANSWER_MEAN** as the default masking strategy and averages activations across all QA pairs to obtain a single *persona vector* per layer.

These activation tensors are saved to disk for reuse across experiments and methods.

$$\mathbf{p}^{(l)} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{h}_i^{(l)}$$

where $\mathbf{h}_i^{(l)}$ is the residual-stream activation at layer l for QA pair i and N is the number of QA pairs.

The extraction method follows recent work that represents high level behavior using mean-pooled residual stream activations over response tokens. The Assistant Axis paper constructs role vectors from mean post-MLP residual-stream activations over all response tokens in generated role-playing rollouts [7]

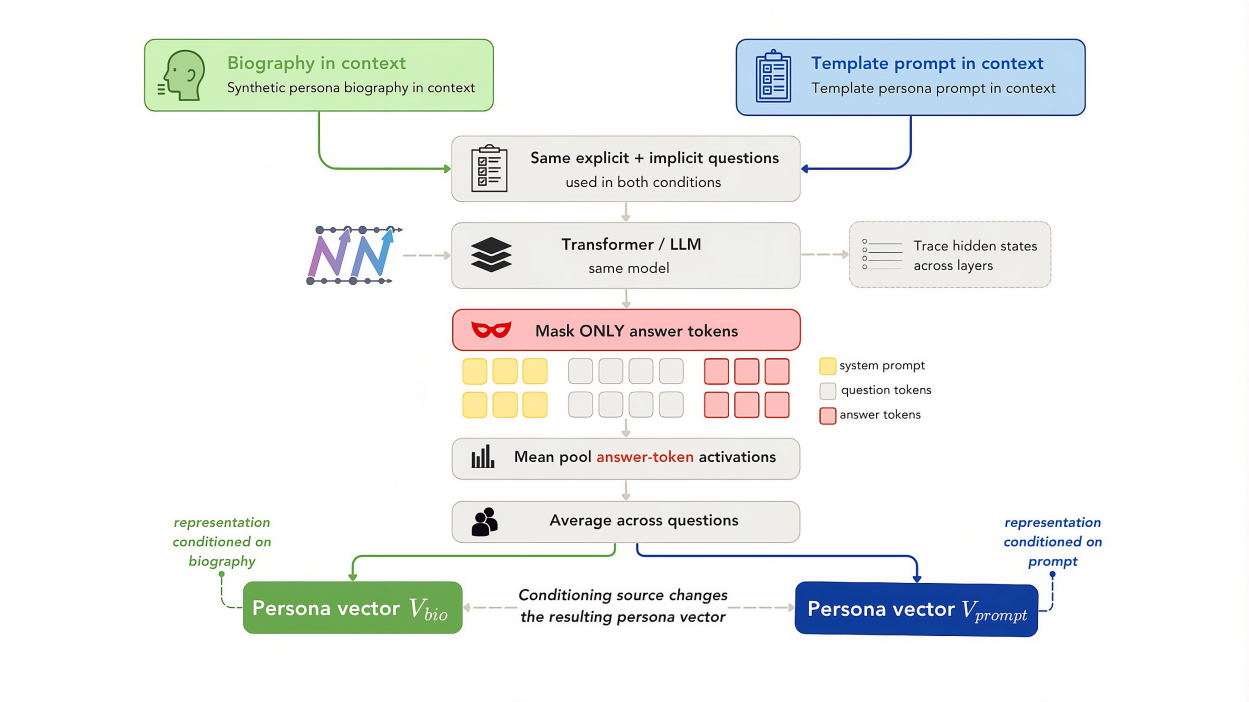


Figure 2: Overview of the SynthPersona dataset generation pipeline.

In particular, we extracted persona vectors for all personas using **gemma-3-27b-it**. For a subset of personas, we also computed vectors across additional models. Persona vectors for all models except **Llama-3.1-450B-Instruct** were obtained using a training subset of 50 explicit and

implicit questions and are publicly available in the *Synth-Persona-Vectors* dataset,[1]. Vectors for **Llama-3.1-405B-Instruct** were extracted using all explicit questions and no implicit questions, prioritizing directly stated persona attributes and running it through NDIF/NNsight [4]

3.1.2 Unsupervised analysis of persona-vector geometry

To characterize the geometry of persona vectors without using supervised labels, we apply PCA, to the set of persona vectors at each layer. For a fixed layer l, we form the row-stacked matrix

$$P^{(l)} = \begin{bmatrix} (\mathbf{p}_1^{(l)})^\top \\ \vdots \\ (\mathbf{p}_M^{(l)})^\top \end{bmatrix} \in \mathbb{R}^{M \times d},$$

where M is the number of personas and d is the hidden dimension. This corresponds to the fixed-layer tensor slice with shape `[n_personas, hidden_size]`. PCA is fit independently at each layer to study whether dominant directions of variance correspond to interpretable persona attributes. The analysis, is performed on vectors centered across personas and L2-normalized by default, so the resulting geometry emphasizes directional/profile similarity rather than shared residual-stream components or raw vector magnitude. We visualize projections across layers and color points by known SynthPersona attributes such as age and sex; these attributes are used only for post hoc interpretation, not to fit the unsupervised geometry.

3.2 Linear probing of persona vectors

To test whether persona attributes are linearly decodable from persona-vector representations, we train linear probes on the **biography-derived** persona vectors of `gemma-3-27b-it`. For each attribute, we sweep across all layers and report the best-performing layer on an 80/20 stratified held-out split (stratified by class label). Binary and categorical attributes are decoded using logistic regression (with binary attributes additionally evaluated using difference-of-means probing), while ordinal attributes are modeled with ridge regression on rank-encoded labels and evaluated after rounding predictions to the nearest rank. Because many attributes are imbalanced and vary substantially in the number of possible values, we report balanced accuracy for discrete attributes rather than standard accuracy. We also report the balanced-accuracy chance level, defined as $1/K$ for K evaluated classes, and the excess over chance, $\Delta = \text{BalAcc} - 1/K$.

3.2.1 Persona-to-LoRA

We leverage pretrained Doc-to-LoRA hypernetworks [1], which generate LoRA adapter weights for a target model in a single forward pass from a document. We use these hypernetworks to construct persona-specific LoRA adapters for conditioning LLM behavior. For each persona, the persona information is provided in one of two formats: a narrative biography or a structured attribute list. The resulting adapter is then applied to the base model before generation or multiple-choice evaluation.

This setup differs from standard in-context persona conditioning. Rather than providing the full persona description in the prompt at inference time, Doc-to-LoRA internalizes the persona into the model's adapted parameters. At inference time, the model receives only a minimal roleplay prompt identifying the persona by name and instructing it to answer as that person. This allows us

[1] https://huggingface.co/datasets/implicit-personalization/synth-persona-vectors

to test whether the generated LoRA contains recoverable persona information, rather than whether the model can simply read and use an explicit profile in context.

We compare five conditions: a no-LoRA minimal baseline, no-LoRA prompting with the biography in context, no-LoRA prompting with the structured attributes in context, an internalized biography LoRA, and an internalized structured-attributes LoRA. We evaluate both explicit questions, which ask about facts directly stated in the persona description, and implicit questions, which require inferring likely preferences, attitudes, or behaviors from the persona. This provides a direct comparison between parameter-level persona internalization and ordinary in-context conditioning.

3.3 Steering Interventions

3.3.1 Activation Addition (ActAdd)

We compute one *attribute-axis* steering vector per demographic attribute by contrasting the internal representations of two groups of personas that differ on a known ground-truth label. For attribute a with positive group $\mathcal{P}$ and negative group $\mathcal{N}$ (e.g. liberal vs. conservative for `political_views`), both groups are prompted with full biography-format contexts, and the steering vector at layer l is the difference in group means:

$$\mathbf{v}_a^{(l)} = \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} \mathbf{h}_{i,\text{bio}}^{(l)} - \frac{1}{|\mathcal{N}|} \sum_{j \in \mathcal{N}} \mathbf{h}_{j,\text{bio}}^{(l)}$$

Using biography-format activations for both groups controls for prompt length and style, so the difference isolates the attribute-relevant direction. The vector is applied at the peak-gap layer ($\text{argmax}_l \|\mathbf{v}_a^{(l)}\|_2$) with scalar coefficient α proportional to the layer's RMS activation norm. At inference, no persona context is provided, with the steering vector as the sole source of persona signal. We evaluate on shared MCQs filtered to the attribute's topic group, measuring $P(\text{correct})$: the probability mass assigned to the persona's ground-truth answer choice (chance = 0.20 for five options).

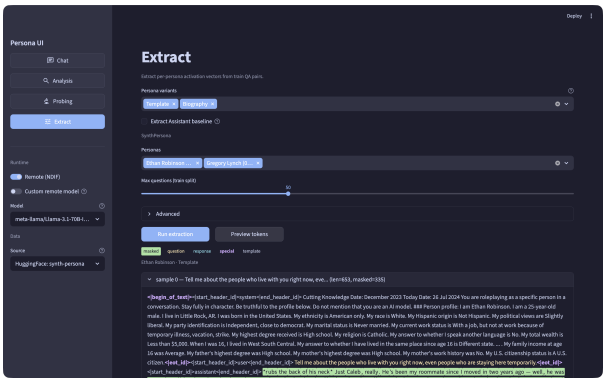
3.4 A Visual Interface for Exploring Implicit Personalization

To support qualitative and quantitative analysis of implicit personalization, we developed an open-source interactive Streamlit interface for extracting, inspecting, and comparing persona representations. The interface is designed to make the main components of our pipeline accessible, interactive, and easily extensible, covering persona-vector extraction, probe training, representation analysis, and interactive generation with persona conditioning.

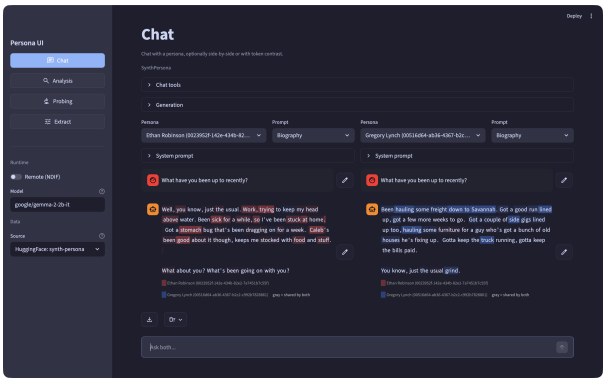
The application provides three main functionalities. First, it supports persona-vector extraction for selected personas, allowing users to choose persona variants, masking strategies, datasets, and extraction settings directly through the UI. Second, it enables analysis of saved persona-vector representations across layers using cosine similarity, PCA, UMAP, Isomap, similarity matrices, and hierarchical clustering. These visualizations make it possible to inspect whether persona-conditioned representations form coherent clusters or separate along meaningful directions in *persona-vector* space. Third, the chat interface allows users to interact with models under different persona-conditioning settings, including in-context biography conditioning, template-based prompting, custom prompts, and an unconditioned baseline. During generation, the interface can expose token-level probe predictions and token-level log-probability contrasts between different persona variants, helping identify where personalization effects appear in the model's output distribution. The chat interface also supports editing user and assistant messages. In comparison mode, edited assistant

messages trigger recomputation of token-level log-probability contrasts, allowing users to inspect how the same response is scored under different persona-conditioned contexts.

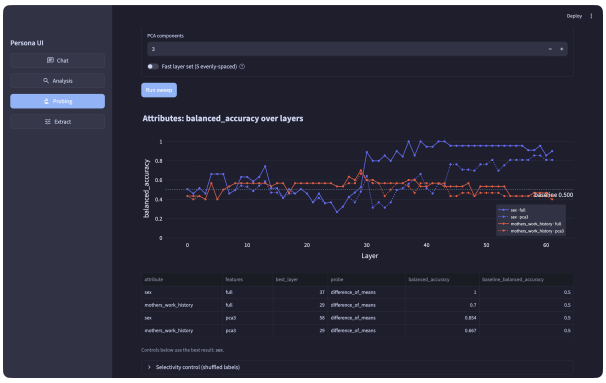
The interface is implemented in Streamlit and is publicly available on Hugging Face[2]. Figure 3 shows representative pages of the application, including persona-vector extraction, interactive chat, probe training, and representation visualization.



(a) Persona-vector extraction interface.



(b) Interactive chat interface with token contrast



(c) Probe training and evaluation interface.

(d) 3D PCA projection for *Llama-3.1-70B-Instruct*.

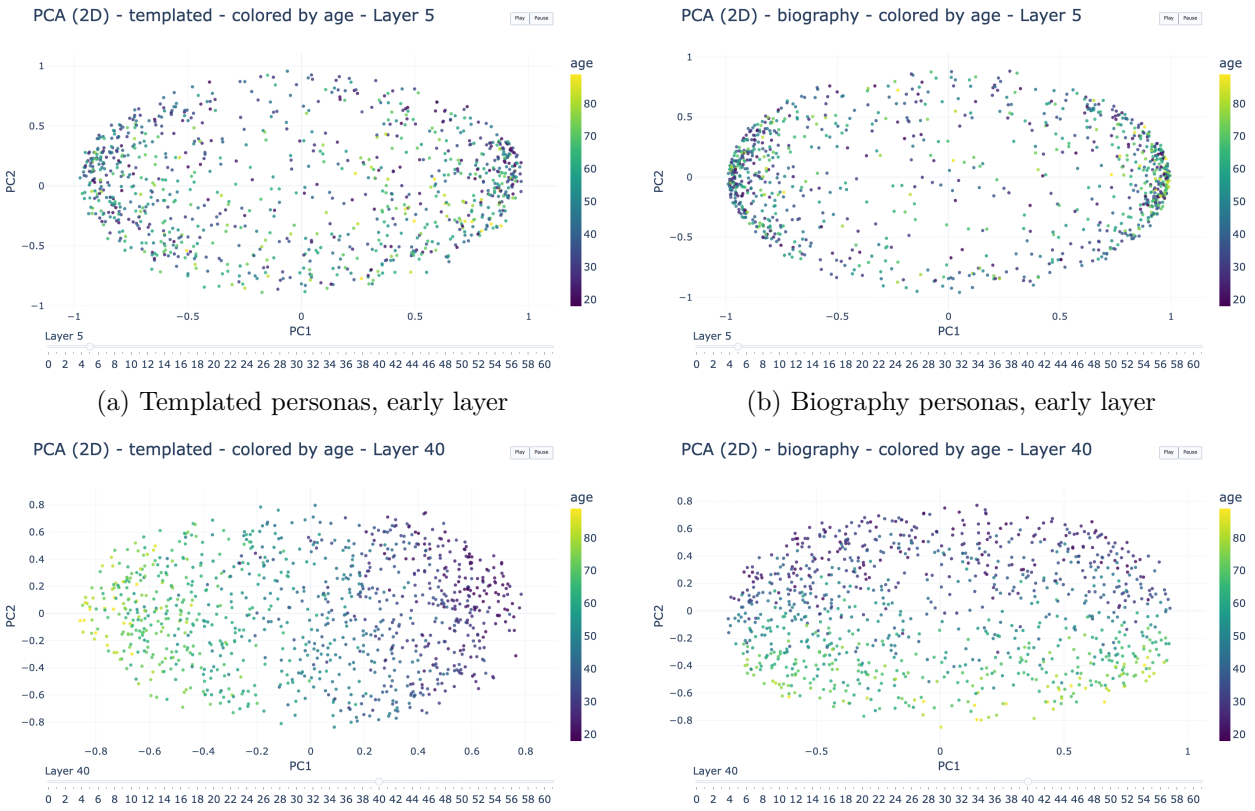
Figure 3: Streamlit application for exploring implicit personalization. The interface supports persona-vector extraction, interactive persona-conditioned chat, probe training and evaluation, and visualization of hidden-state representations.

4 Results

4.1 Unsupervised geometry of persona vectors

Unsupervised projections suggest that a small number of directions capture structured variation across personas. In PCA visualizations of `gemma-3-27b-it` biography-derived persona vectors, the *age* attributes visibly align with the first two principal components. In early layers, age does not appear to be systematically organized in the PCA plane; by the middle-to-late layers, however, the projection develops a smoother age gradient for both templated and biography-derived persona

[2] https://huggingface.co/spaces/implicit-personalization/persona-ui



(a) Templated personas, early layer

(b) Biography personas, early layer

(c) Templated personas, late layer

(d) Biography personas, late layer

Figure 4: PCA projections of persona vectors colored by age across early and late transformer layers for both templated and biography derived personas in `gemma-3-27b-it`. Early layers don't organize by age, whereas late layers develop a smooth age gradient aligned with the principal directions.

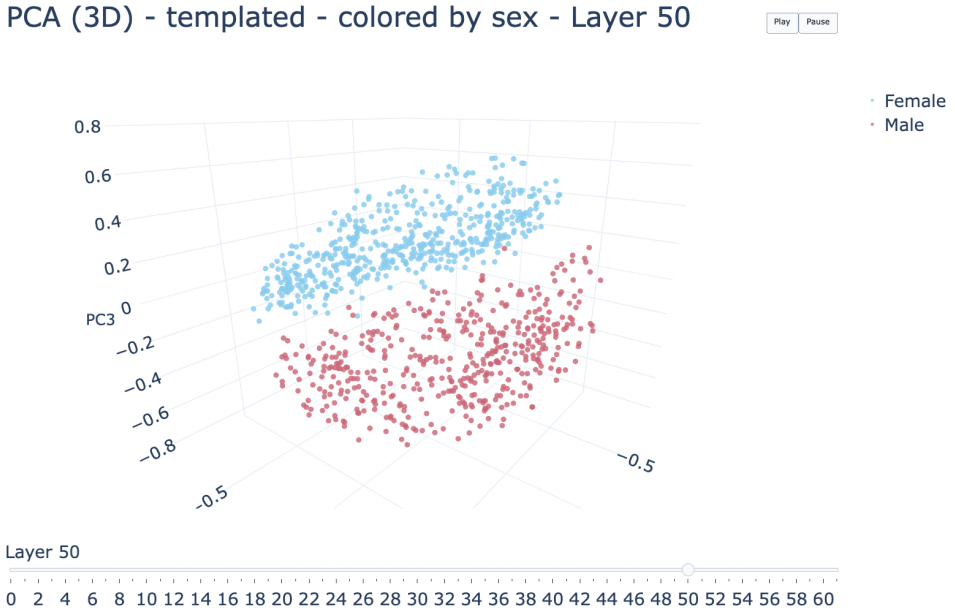


Figure 5: 3D PCA projection of templated persona vectors, colored by sex attribute.

Additionally, when examining the first three PCA components of the templated persona vectors from **gemma-3-27b-it**, we observe a clean separation between male and female personas along the *sex* attribute.

4.2 Persona-to-LoRA

We evaluated Persona-to-LoRA using the pretrained hypernetworks available for Gemma-2B and Qwen3-4B on 100 personas. For each model and condition, this gives 5,700 explicit and 41,800 implicit multiple-choice questions. Intervals in Table 1 are approximate 95% normal intervals over questions, computed as $\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/n}$. These intervals are useful for checking the precision of the aggregate question-level accuracy.

The results are strongly model-dependent. Gemma-2B remains a weak test of persona internalization: even with persona information placed directly in context, explicit accuracy reaches only 52.3% for biography prompts and 60.4% for templated attributes. The generated LoRAs barely improve over the minimal no-LoRA baseline, moving from 29.9% to 32.2%–32.8% on explicit questions, and they slightly underperform the minimal baseline on implicit questions. We therefore treat the Gemma-2B results mainly as evidence that a weak base model can limit hypernetwork internalization of complex persona information, rather than as a decisive negative result about Persona-to-LoRA itself.

Qwen3-4B shows a clearer internalization effect, especially on explicit persona questions. The no-LoRA minimal baseline achieves 33.1% accuracy, while biography and templated-attribute LoRAs reach 56.8% and 51.1%, respectively. These scores remain well below the corresponding in-context conditions, which reach 81.4% for biography prompts and 75.2% for templated attributes, but they show that substantial explicit persona information can be recovered from the generated LoRA alone, without placing the full persona profile in the prompt.

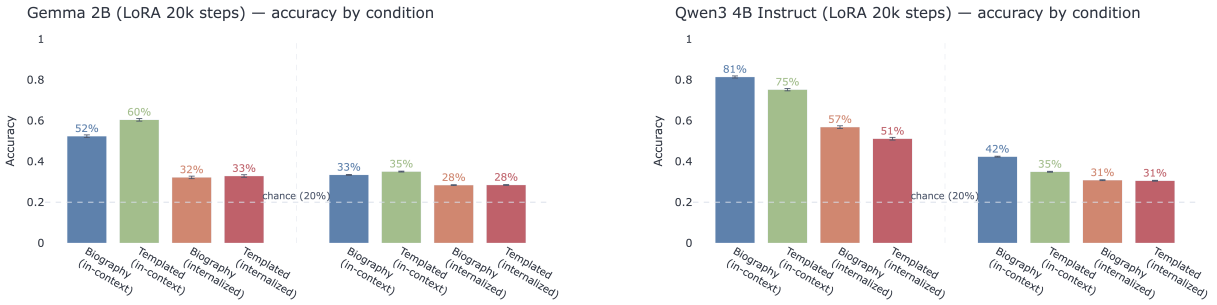
The effect is much weaker on implicit questions. For Qwen3-4B, the minimal baseline achieves 28.2% accuracy, while the internalized biography and structured-attribute LoRAs reach 30.8% and 30.5%. These gains small, and remain far below the in-context biography condition at 42.3%.

Table 1: Persona-to-LoRA accuracy with approximate 95% normal intervals over questions for 100 personas. These intervals do not account for persona-level clustering.

Model	Condition	Explicit acc. ($n = 5{,}700$)	Implicit acc. ($n = 41{,}800$)
Gemma-2B	No LoRA, minimal prompt	29.9 [28.7, 31.1]	29.3 [28.9, 29.8]
Gemma-2B	No LoRA, biography in context	52.3 [51.0, 53.6]	33.4 [32.9, 33.8]
Gemma-2B	No LoRA, templated attributes in context	60.4 [59.1, 61.7]	35.0 [34.6, 35.5]
Gemma-2B	Biography LoRA, minimal prompt	32.2 [31.0, 33.4]	28.4 [27.9, 28.8]
Gemma-2B	Templated-attribute LoRA, minimal prompt	32.8 [31.6, 34.1]	28.4 [28.0, 28.9]
Qwen3-4B	No LoRA, minimal prompt	33.1 [31.8, 34.3]	28.2 [27.8, 28.7]
Qwen3-4B	No LoRA, biography in context	81.4 [80.4, 82.4]	42.3 [41.8, 42.8]
Qwen3-4B	No LoRA, templated attributes in context	75.2 [74.0, 76.3]	34.9 [34.5, 35.4]
Qwen3-4B	Biography LoRA, minimal prompt	56.8 [55.5, 58.1]	30.8 [30.4, 31.2]
Qwen3-4B	Templated-attribute LoRA, minimal prompt	51.1 [49.8, 52.4]	30.5 [30.1, 31.0]

Overall, these experiments indicate that Persona-to-LoRA is promising but limited by hypernetwork quality, target-model compatibility, and the type of persona information being evaluated. Qwen3-4B results show that generated LoRAs can encode persona-specific information outside the context window, especially for explicit facts, while Gemma-2B results show that this behavior is not automatic across models. We therefore treat Persona-to-LoRA as a useful complementary baseline for activation-level persona interventions, and as a direction that may improve with stronger

hypernetworks and richer conditioning documents (which might for example include multiple conversations or questions answered by the model).



(a) Gemma-2B (b) Qwen3-4B

Figure 6: Persona-to-LoRA evaluation on explicit and implicit persona questions. Each plot compares no-LoRA baselines, where persona information is either absent or provided in context, against internalized LoRAs generated from biography or structured-attribute descriptions.

4.3 Linear probing of persona vectors

We next evaluate how much structured personal information is linearly decodable from the persona vectors. For each attribute, we train a linear probe on held-out biography activations from `gemma-3-27b-it`. The results show substantial linear structure in the persona-vector representation. Many binary attributes are almost perfectly decodable: US citizenship status reaches **1.000** balanced accuracy, sex reaches **0.995**. These results indicate that some demographic properties are represented in directions that are almost directly accessible to a linear readout.

Categorical attributes are also strongly recoverable, even when the number of classes is large. Residence at age 16 reaches **0.972** balanced accuracy over 10 classes. Because balanced-accuracy chance decreases as $1/K$, these large gains over chance show that the probes are not simply exploiting class imbalance. Race-related attributes are also highly decodable: detailed race reaches **0.908** over 5 classes.

Table 2: Linear decodability of binary persona attributes from `gemma-3-27b-it` biography activations. Logistic probes are evaluated by balanced accuracy; Δ reports gain over chance.

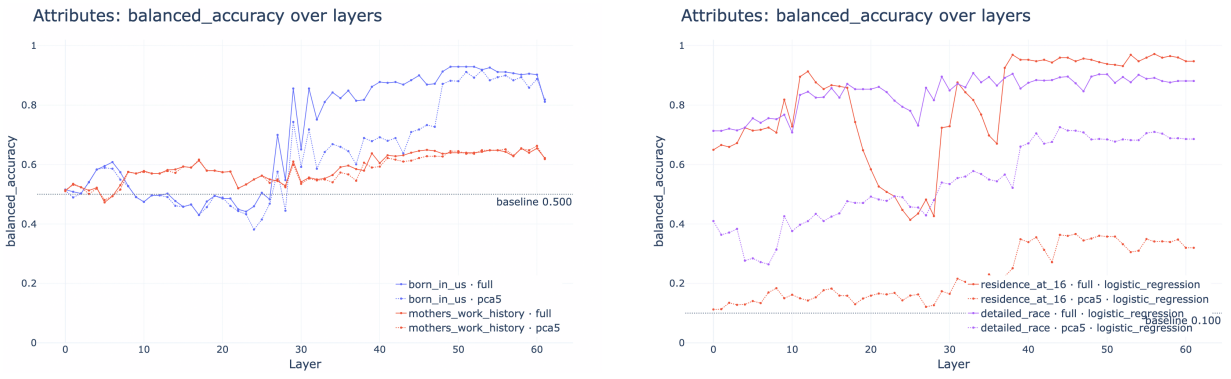
Attribute	BalAcc ↑	Chance	Δ ↑
US citizenship status	**1.000**	0.500	**0.500**
Sex	0.995	0.500	0.495
Born in the US	0.965	0.500	0.465
Mother worked	0.754	0.500	0.254
Speaks another language	0.741	0.500	0.241

Table 3: Linear decodability of categorical persona attributes from `gemma-3-27b-it` biography activations. Logistic probes are evaluated by balanced accuracy; Δ reports gain over chance.

Attribute	K	BalAcc ↑	Chance	Δ ↑
Residence at age 16	10	**0.972**	0.100	**0.872**
State	30	0.755	0.033	0.722
Detailed race	5	0.908	0.200	0.708
Work status	8	0.730	0.125	0.605
City	52	0.610	0.019	0.591
Race	3	0.892	0.333	0.559
Marital status	5	0.741	0.200	0.541
Family structure at 16	4	0.609	0.250	0.359
Religion	10	0.465	0.100	0.365
Religion at age 16	9	0.444	0.111	0.333
Ethnicity	39	0.346	0.026	0.320

Finally, we compare probes trained on raw persona vectors with probes trained only on the first five PCA components. For the binary attributes in Fig. 7, this comparison uses the difference-of-means probe; for categorical attributes, it uses multinomial logistic regression. PCA is fit on the training split only.

The comparison shows that high linear decodability sometimes lies in the leading variance directions, but not always. For example, `born_in_us` remains highly decodable with only five principal components, dropping only slightly from 0.929 to 0.917 balanced accuracy. `detailed_race` also retains substantial signal, reaching 0.726 with five PCA components compared with 0.908 in the full activation space. In contrast, `residence_at_16` is nearly perfectly decoded in the full space (0.972), but falls sharply to 0.366 when restricted to the first five PCA components. `mothers_work_history` is weaker overall and is affected less consistently, with balanced accuracy changing from 0.655 in the full space to 0.663 with five PCA components. Thus, persona vectors contain some attributes aligned with coarse, high-variance directions, while other attributes are linearly accessible only through information outside the leading PCA subspace.



(a) Binary attributes

(b) Categorical attributes

Figure 7: Comparison between full-dimensional linear probes and probes trained only on the first five PCA components of the persona vectors. Binary attributes use the difference-of-means probe; categorical attributes use logistic regression. Some attributes, such as `born_in_us` and `detailed_race`, remain substantially decodable from five components, whereas `residence_at_16` depends much more on information outside the leading PCA directions.

4.4 Attribute-Axis Steering

For each attribute, we construct a steering vector $\mathbf{s}_l = \bar{\mathbf{h}}_l^+ - \bar{\mathbf{h}}_l^-$ at every residual layer l, where the positive and negative poles are defined by the attribute's high and low ends (e.g. college-educated vs. ≤high-school for the education axis). At inference we apply $\mathbf{h}_l \leftarrow \mathbf{h}_l + \alpha, \mathbf{s}l$, scaling α by $\hat{\alpha} = 20 \cdot \sigma^- l/|\mathbf{s}_l|$ (a per-model normalisation based on the negative group's activation RMS), and sweep $\alpha \in -2, -1, 0, 0.5, 1, 2 \times \hat{\alpha}$. No persona is provided at inference; the only persona signal comes from the steering vector itself.

Preliminary experiments with biography included as a system prompt confirm the steering direction is real: on explicit political MCQs with Qwen2.5-7B-Instruct, gold-label logit margins increase monotonically from 7.33 to 8.74 as the coefficient goes from -3 to $+3$, even though argmax accuracy barely shifts. The question is whether the same vector works when the model has no persona context to anchor on.

Table 4 reports results for Llama-3.1-70B-Instruct, Gemma-3-27B-IT and Gemma-2-9B-IT. The education axis shows a clean directional effect on Llama-70B: the college-educated group rises from P(correct) = 0.32 at α=0 to 0.34 at α=+2, while the ≤ high-school group falls from 0.25 to 0.23. The political views, party identification, and wealth axes show weaker or inconsistent shifts at this sample size, which we expect to resolve with larger groups.

Model	Axis	pos_0	neg_0	Δpos	Δneg
Llama-3.1-70B	Education	0.32	0.25	+0.02	−0.02
	Wealth	0.31	0.35	+0.00	−0.01
	Political	0.26	0.24	+0.02	−0.02
	Party ID	0.26	0.24	−0.01	+0.00
Gemma-3-27B	Education	0.30	0.22	+0.02	−0.03
	Wealth	0.27	0.22	+0.00	−0.01
	Political	0.17	0.13	+0.01	−0.01
	Party ID	0.19	0.14	+0.01	−0.02
Gemma-2-9B	Education	0.10	0.08	+0.05	−0.02
	Wealth	0.09	0.06	+0.00	+0.00
	Political	0.09	0.05	+0.02	+0.00
	Party ID	0.09	0.03	+0.00	−0.01

Table 4: Attribute-axis steering results on implicit MCQs (no persona context at inference). Baseline = P(correct) at α=0; Δ = change at α=+2$\hat{\alpha}$. Chance = 0.20.

Gemma-2-9B's baselines fall well below chance (0.09–0.10 vs. 0.20) across all axes, which we attribute to a positional prior at the last residual layer. Because steering vector norms grow monotonically with depth, we are intervening at a point where token distributions are already nearly formed rather than where attribute information is actually encoded. Llama-70B is less susceptible to this, likely because its greater depth spreads representation changes across more layers before final projection.

Gemma-3-27B shows a similar pattern to Llama-70B on the education axis (Δ pos =+0.02, Δ neg =−0.03). The wealth and political axes show minimal steering effect, consistent with the Llama-70B results on those dimensions. Gemma-2-9B shows the largest relative effect on the education axis (Δ pos =+0.05), a 50% relative improvement over its baseline of 0.10, though absolute performance remains well below chance across all axes for the reasons described above.

Taken together, these results establish that biography-derived attribute-axis vectors encode a real and steerable signal, with the education axis transferring cleanly to the no-context implicit condition on Llama-70B.

5 Challenges & Open Questions

The main open challenge is the gap between decodability and control. The probing results show that persona attributes are strongly recoverable from persona vectors, but this does not imply that adding or modifying the corresponding directions will reliably produce the intended behavioral change. This gap is especially important for implicit questions, where in-context conditioning already produces relatively modest accuracy and where Persona-to-LoRA provides only small improvements over the minimal baseline. Finally, the current experiments are limited by model hyper-network availability in the Persona-to-LoRA results show that conclusions can change substantially across target models. Future versions should therefore report confidence intervals across personas, evaluate additional open-weight models (by pretraining additional hyper-networks), and separate failures more cleanly caused by the base model from failures caused by the intervention method.

References

[1] Rujikorn Charakorn, Edoardo Cetin, Shinnosuke Uesaka, and Robert Tjarko Lange. Doc-to-lora: Learning to instantly internalize contexts, 2026.

[2] Yida Chen, Aoyu Wu, Trevor DePodesta, Catherine Yeh, Kenneth Li, Nicholas Castillo Marin, Oam Patel, Jan Riecke, Shivam Raval, Olivia Seow, Martin Wattenberg, and Fernanda Viégas. Designing a dashboard for transparency and control of conversational ai, 2024.

[3] Dami Choi, Vincent Huang, Sarah Schwettmann, and Jacob Steinhardt. Scalably extracting latent representations of users. https://transluce.org/user-modeling, November 2025.

[4] Jaden Fried Fiotto-Kaufman, Alexander Russell Loftus, Eric Todd, Jannik Brinkmann, Koyena Pal, Dmitrii Troitskii, Michael Ripa, Adam Belfki, Can Rager, Caden Juang, Aaron Mueller, Samuel Marks, Arnab Sen Sharma, Francesca Lucchetti, Nikhil Prakash, Carla E. Brodley, Arjun Guha, Jonathan Bell, Byron C Wallace, and David Bau. NNsight and NDIF: Democratizing access to open-weight foundation model internals. In *The Thirteenth International Conference on Learning Representations*, 2025.

[5] Zhijing Jin, Nils Heil, Jiarui Liu, Shehzaad Dhuliawala, Yahang Qi, Bernhard Schölkopf, Rada Mihalcea, and Mrinmaya Sachan. Implicit personalization in language models: A systematic study. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12309–12325, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[6] Akaash Kolluri, Shengguang Wu, Joon Sung Park, and Michael S Bernstein. Finetuning LLMs for human behavior prediction in social science experiments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30084–30099, Stroudsburg, PA, USA, 2025. Association for Computational Linguistics.

[7] Christina Lu, Jack Gallagher, Jonathan Michala, Kyle Fish, and Jack Lindsey. The assistant axis: Situating and stabilizing the default persona of language models, 2026.

[8] Vera Neplenbroek, Arianna Bisazza, and Raquel Fernández. Reading between the prompts: How stereotypes shape LLM's implicit personalization. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20367–20400, Suzhou, China, November 2025. Association for Computational Linguistics.

[9] Alexander Pan, Lijie Chen, and Jacob Steinhardt. Latentqa: Teaching llms to decode activations into natural language, 2024.

[10] Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. Generative agent simulations of 1,000 people, 2024.

[11] Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. Generative agent simulations of 1,000 people, 2024.

[12] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models, 2025.