PersonaTwin: A Multi-Tier Prompt Conditioning Framework for Generating and Evaluating Personalized Digital Twins

Sihan Chen,[1] **John P. Lalor,**[2,3] **Yi Yang,**[4] **Ahmed Abbasi**[2,3]
[1]Viterbi School of Engineering, University of Southern California
[2]Human-centered Analytics Lab, University of Notre Dame
[3]Department of IT, Analytics, and Operations, University of Notre Dame
[4]Department of Information Systems, Business Statistics and Operations Management, HKUST
schen976@usc.edu, john.lalor@nd.edu, imyiyang@ust.hk, aabbasi@nd.edu

Abstract

While large language models (LLMs) afford new possibilities for user modeling and approximation of human behaviors, they often fail to capture the multidimensional nuances of individual users. In this work, we introduce `PersonaTwin`, a multi-tier prompt conditioning framework that builds adaptive digital twins by integrating demographic, behavioral, and psychometric data. Using a comprehensive data set in the healthcare context of more than 8,500 individuals, we systematically benchmark `PersonaTwin` against standard LLM outputs, and our rigorous evaluation unites state-of-the-art text similarity metrics with dedicated demographic parity assessments, ensuring that generated responses remain accurate and unbiased. Experimental results show that our framework produces simulation fidelity on par with oracle settings. Moreover, downstream models trained on persona-twins approximate models trained on individuals in terms of prediction and fairness metrics across both GPT-4o-based and Llama-based models. Together, these findings underscore the potential for LLM digital twin-based approaches in producing realistic and emotionally nuanced user simulations, offering a powerful tool for personalized digital user modeling and behavior analysis.

1 Introduction

Large language models (LLMs) present exciting opportunities for user modeling, behavior analysis, and understanding and improving the human condition. Opportunities abound across an array of contexts including healthcare, education, etc. For instance, a pressing healthcare challenge is the development of conversational systems that truly account for the nuanced experiences and identities of individual patients (Davenport and Kalakota, 2019; Jiang et al., 2017). In telemedicine or mental health coaching scenarios, clinicians require tools that adapt dynamically to each patient's demographic, behavioral, and psychological profile, rather than offering generic responses. Although large language models such as GPT-4 (Achiam et al., 2023) and Llama-3-70b (Dubey et al., 2024) have shown substantial improvement in natural language processing tasks–demonstrated by benchmarks in medical QA datasets, automated note taking, and patient triage use cases–they still struggle to model the multifaceted nature of personal identity in real world settings (Laranjo et al., 2018). Numerous persona-based conversational frameworks have begun to address this gap by incorporating basic user attributes into language models, and studies have demonstrated modest gains in engagement and trust when even minimal demographic cues are included (Abuelezz et al., 2024). However, many of these frameworks remain limited by static or simplistic representations that fail to capture evolving factors. In the health setting, for instance, these frameworks fail to represent health behaviors over time, emotional states during stressful events, and shifting attitudes toward medical professionals (Huang et al., 2024; Guo and Chen, 2024).

To overcome these limitations, we draw on the concept of digital twins, originally popularized in engineering, to represent physical systems virtually (Grieves, 2014). In our adaptation, a "digital twin" for conversational AI is a virtual replica of a user (e.g., a patient) that encapsulates not only demographic information (e.g., age, gender, and socioeconomic status), but also behavioral data (e.g., physical activity, dialogue habits, and compliance with medications), along with psychological attributes (e.g., anxiety levels, trust, and perceived literacy) (Meijer et al., 2023; Lukaniszyn et al., 2024). This framework is particularly relevant in scenarios such as mental health chatbots or chronic disease management systems, where the emotional and psychological realism of the dialogue can directly impact patient adherence and satisfaction.

However, creating such multidimensional and adaptive representations raises several methodolog-

Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM² 2025), pages 774–788

ical hurdles. First, many existing language-based approaches provide only a narrow view of a user's identity, focusing predominantly on stylistic or linguistic features while neglecting deeper demographic or psychometric attributes. Second, static systems do not adapt to shifting contexts, including new symptoms or a gradual erosion of trust, resulting in repetitive or misaligned conversations. Third, there is a lack of comprehensive evaluation benchmarks that jointly measure factual correctness, emotional coherence, and alignment with actual user expressions. For instance, in clinical contexts, much of the previous NLP work has focused on factual accuracy, leaving emotional nuance and user alignment underexplored (Jiang et al., 2017).

To address these challenges, we introduce `PersonaTwin`, a multi-tier prompt conditioning framework that systematically integrates demographic, behavioral, and psychological data into a comprehensive digital twin. Our approach employs a structured methodology in which each level of user information is processed and encoded into the model prompt (Lester et al., 2021; Chen et al., 2024). `PersonaTwin` consists of two parts, **Multi-tiered Conditioning for Digital Twin Creation** and **Conversation Update Loop**. In the first part, step 1 involves mapping person-level persona metadata to persona information tiers such as demographics, behavioral, and psychological information; whereas step 2 initializes the digital twin. In the second part, the instantiated digital twin is iteratively updated with the real person's previous conversational responses to psychometric questions (e.g., related to numeracy, anxiety, etc.). This layered technique enables the model to produce simulated dialogues that are not only contextually relevant but also capable of reflecting shifting user states as new data are introduced (Reimers and Gurevych, 2019). We tested our framework using a large-scale psychometric dataset of more than 8,500 respondents (Abbasi et al., 2021), which provides a rich combination of survey-based measures, user-generated text, and demographic information. By incorporating real responses on health numeracy, medical visit anxiety, and trust in healthcare providers, we ensure that our simulations reflect authentic user experiences while maintaining privacy through deidentification and ethical safeguards (Cascella et al., 2023).

To rigorously evaluate `PersonaTwin`, we implemented a dual-pronged strategy. First, we employ state-of-the-art text similarity metrics to measure how closely the digital twin-generated output matches the actual user responses (Reimers and Gurevych, 2019; Song et al., 2020; Wang et al., 2020). Second, we use a downstream NLP prediction task to examine the efficacy of the generated twins, relative to the actual users, in terms of the fine-tuned model's predictive performance and fairness assessments across key demographic dimensions (Hardt et al., 2016; Barocas et al., 2023).

Our key contributions are: (i) we introduce `PersonaTwin`, a multi-tier framework that integrates demographic, behavioral, and psychological data to generate adaptive digital twins, enhancing realism with LLM-driven personal insights, (ii) we generate 8,500+ digital twins representing diverse personas and validate response fidelity using conditioned experiments and advanced similarity metrics, and (iii) we conduct a rigorous downstream evaluation of models trained/tested on generated personas versus actual users and show that the persona-based models achieve comparable predictive power and fairness outcomes.[1]

2 Related Work

2.1 Simulative Persona Construction and the Importance of Digital Twins

A pioneering study by Park et al. (2023) laid the groundwork for persona-based conversational systems by simulating a small town of 25 virtual characters using simplified models of human cognition to enable dialogue. Recent advances in generative agents have begun to explore the ability of LLMs to emulate more precise human behaviors. For instance, Park et al. (2024) simulate survey responses for 1,000 individuals based on audio interviews with participants. Similarly, Xu et al. (2024) benchmark LLM agents on consequential real-world tasks. In parallel, Chuang et al. (2024) develop digital twins using a belief network to capture open-domain dimensions–such as those revealed in the Controversial Beliefs Survey–broadening the scope of persona construction. Moreover, Shao et al. (2023) propose Character-LLM, an approach that crawls online records and stories of historical or fictional figures to serve as persona inputs, thereby enriching the contextual and experiential background of the simulated agents. Additionally, as discussed in (Meister et al., 2024), "steering methods" offer

[1]Our code is available on GitHub: `https://github.com/nd-hal/psych-agent-llm`.

promising strategies to guide the behavior of simulated agents. However, these studies often face two major challenges: (1) some approaches rely solely on unstructured text inputs yet lack the precise control needed to ensure consistency in the perspectives from which user content is drawn; and (2) other methods incorporate structured data, but primarily focus on personal background without delving deeply into the internal psychological traits and behavioral dynamics of individuals.

Furthermore, Salemi et al. (2024) introduce LaMP, a comprehensive benchmark and retrieval-augmentation framework that conditions LLMs on fine-grained user profiles–spanning classification and generation tasks–to produce personalized outputs, demonstrating significant gains in both zero-shot and fine-tuned settings. Meanwhile, Sorokovikova et al. (2024) provide empirical evidence that LLMs (e.g. Llama-2, GPT-4, Mixtral) can simulate stable Big Five personality traits, revealing the potential of LLM-driven agents to model intra-individual psychological characteristics with consistency across varied prompts.

Our work addresses this challenge by taking a fine-grained, high-dimensional approach to simulating individual personas. We integrate psychological, behavioral, personal background, and linguistic style information to construct digital twins that capture the nuanced and evolving nature of real human identities. By leveraging authentic user inputs as benchmarks, our framework explores replication of core behavioral patterns and individual variability that is typically lost in more simplistic, one-dimensional models. We demonstrate the potential for enriched representations for generating digital twins that better reflect real human behavior.

2.2 Evaluation Metrics for Fairness and Authenticity in Generative Agents

Evaluating generative agents requires robust metrics that capture not only the linguistic quality but also the downstream efficacy and fairness of the generated responses. Many studies have adopted LLM-based evaluation methods, either by leveraging off-the-shelf or fine-tuned LLMs, or by incorporating human evaluators, to assess the authenticity of generated text (Jandaghi et al., 2024; Mendonça et al., 2024; Park et al., 2023; Chiang and Lee, 2023). Although these methods have demonstrated promising results, they are not without drawbacks. In scenarios involving large volumes of language data, extensive human evaluation quickly becomes both cost- and time-inefficient. Furthermore, the performance of these evaluation frameworks relies heavily on the underlying LLMs, which may harbor inherent biases or produce unpredictable outputs (Lin and Chen, 2023). Moreover, traditional automatic metrics such as BLEU, ROUGE, METEOR, and CIDEr (Papineni et al., 2002; Lin, 2004; Banerjee and Lavie, 2005; Oliveira dos Santos et al., 2021) often fall short in capturing deeper semantic alignment and social fairness (Zhang et al., 2020).

To overcome these challenges, embedding-based metrics, particularly those leveraging BERT, have emerged as a promising balance between effectiveness and efficiency. For example, BERTScore (Zhang et al., 2020) computes semantic similarity by comparing contextual embeddings of generated texts with those of reference texts, thereby capturing nuances that traditional n-gram metrics often miss. Moreover, Zhu and Bhat (Zhu and Bhat, 2020) introduce GRUEN, a reference-less framework that leverages a BERT-based model to reliably assess the linguistic quality of generated text. Additionally, several studies have extended BERT-based evaluations beyond mere semantic alignment. For instance, Lalor et al. (2022) fine-tuned BERT and RoBERTa models and assessed fairness via disparate impact scores across multiple demographic attributes. These applications underscore how BERT and its variants can provide a robust and efficient framework for evaluating both the authenticity and fairness of generative agents, offering a viable alternative to more resource-intensive LLM-based or human evaluation strategies (Lin and Chen, 2023).

3 Methodology

In this section, we introduce the structure of our proposed framework, `PersonaTwin` (§3.1), and then detailed our evaluation metrics (§3.2).

3.1 Digital Twin Construction

In this section, we detail our two-stage methodology for constructing and refining digital twins using large language models (LLMs). We denote our framework by `PersonaTwin`. In Stage 1, we create an initial digital twin by integrating multi-dimensional user data into a structured prompt for LLM. In Stage 2, we iteratively update the digital twin based on new user input and conversation data, thus capturing temporal changes in user states. The

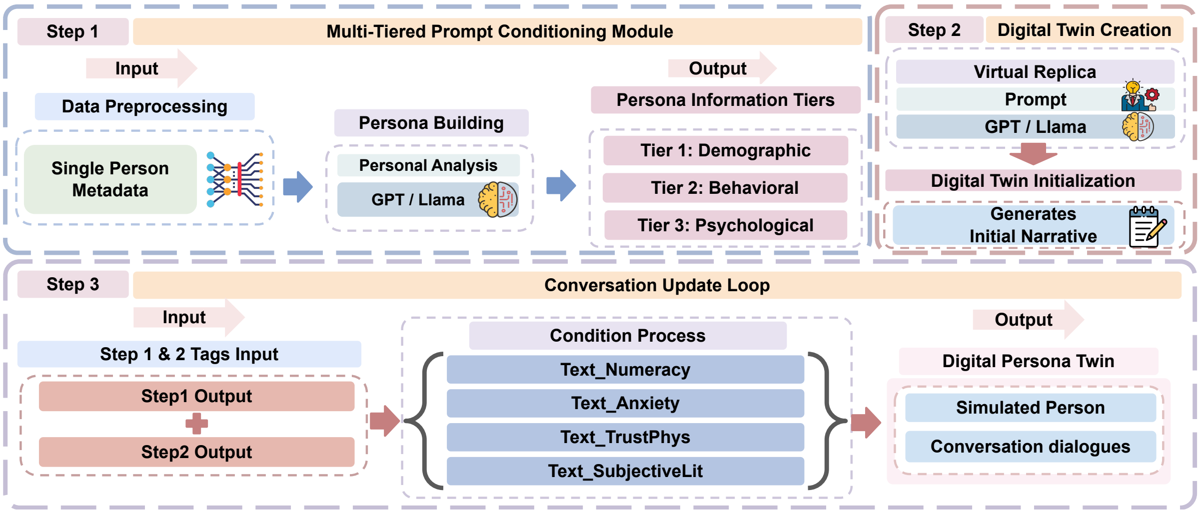

Figure 1: An overview of the PersonaTwin framework, including 1) Multi-Tiered Prompt Conditioning Module, 2) Digital Twin Creation, and 3) Conversation Update Loop.

overall process for constructing and refining digital twins is formally detailed in Figure 1.

3.1.1 Digital Twin Initialization

For the first step of initializing digital twins, we systematically collect and preprocess heterogeneous user data: $D = \{d_1, d_2, \ldots, d_N\}$, where each d_i can represent a demographic (age, race, income), behavioral (physical activity, dietary habits, medication adherence), or psychological (trust, anxiety levels, literacy, numeracy) attributes. A preprocessing function $I(\cdot)$ converts and normalizes all these inputs into a structured representation:

$$X = I(D) = \left[X_{\text{dem}},\, X_{\text{beh}},\, X_{\text{psy}}\right]. \qquad (1)$$

Here, $X_{\text{dem}}, X_{\text{beh}}, X_{\text{psy}}$, respectively, encode the demographic, behavioral, and psychological data in vectorized or categorical form.[2]

Multi-Tiered Template Functions. Unlike a simple concatenation of all features, PersonaTwin employs three dedicated template functions: Template_dem, Template_beh, and Template_psy, each tuned to capture domain-specific nuances. These functions provide additional context such as causal phrases (e.g., "Because the person has high anxiety..."), relevant guidelines, or rhetorical questions that nudge the LLM to infuse the output with emotional tone and factual correctness.

[2]Refer to Appendix A.1 for further details.

Formally,

$$\begin{aligned} P_{\text{dem}} &= \texttt{Template\_dem}(X_{\text{dem}}), \\ P_{\text{beh}} &= \texttt{Template\_beh}(X_{\text{beh}}), \\ P_{\text{psy}} &= \texttt{Template\_psy}(X_{\text{psy}}). \end{aligned} \qquad (2)$$

where each template can *rewrite, summarize, or highlight* the most critical aspects of the data. For example, if X_{psy} indicates a high anxiety level, Template_psy might produce text emphasizing the user's tendency to worry about medical procedures, thus improving emotional realism.

Initial Digital Twin Generation. We concatenate the tier-specific prompts to form a composite prompt:

$$P = \text{Concat}(P_{\text{dem}}, P_{\text{beh}}, P_{\text{psy}}), \qquad (3)$$

which is passed to a selected LLM $G(\cdot)$ to obtain the initial digital twin T_0:

$$T_0 = G(P). \qquad (4)$$

This *initialization* step produces a coherent user narrative or persona that encapsulates the baseline demographic, behavioral, and psychological characteristics.

3.1.2 Conversation Data Integration and Dynamic Update Loop

Although the initial digital twin T_0 provides a rich snapshot of the focal user, it cannot reflect changes in user states or additional data acquired over time. This motivates our second stage, where we iteratively integrate user's conversations (e.g., with psychiatrists) into our digital twin framework.

Conversation Update Mechanism. At each iteration t, the user query Q_t corresponds to one of the four types of prompts in Table 1 (i.e., `Text_Numeracy`, `Text_Anxiety`, `Text_TrustPhys`, or `Text_SubjectiveLit`). We obtain a corresponding user response R_t, which may be drawn from real user data or a newly simulated input. An update function U refines the digital twin T_t as follows:

$$T_{t+1} = U(T_t, Q_t, R_t). \quad (5)$$

In practice, U rechecks each prompt template to integrate relevant changes. For example, if R_t indicates a dose increase for a medication, `Template_beh` is updated to reflect this new regimen. In contrast, if the user contradicts an earlier statement (e.g., previously denied smoking but now mentions occasional use), `Template_beh` reconciles these by prioritizing the recent self-report while tagging older statements as "possible past data." This conflict resolution policy ensures that the most up-to-date information prevails, although older data are retained for longitudinal context.

Multi-Tiered Prompt Conditioning Experiments. Rather than simply updating static persona templates, we devised eight distinct subsample conditions, denoted by

$$T' = \{T'_1, T'_2, \ldots, T'_8\}. \quad (6)$$

to assess how well `PersonaTwin` generates realistic user responses under varying degrees of known personal and conversational information pertaining to the focal user. These conditions are based on two factors: (1) whether the simulated person receives their paired users' three persona information tiers (i.e., demographic, behavioral, psychological); (2) whether the simulated person receives some or none of the four potential conversation updates (called few-shot if yes, zero-shot if no).

Specifically, we define T'_1 as ***Persona Oracle***, where the system prompt includes persona information tier data and the all four conversation updates are revealed, thus serving as a maximum informed oracle. We then introduce four ***Persona Few-shot*** variants, T'_2, T'_3, T'_4, and T'_5, each withholding one of the four real responses to test the model's ability to infer missing content from partial context. Next, T'_6, labeled ***Persona Zero-shot***, omits all real answers entirely, requiring the LLM to generate plausible responses purely from the user's personal attributes. In contrast, T'_7, named ***Few-shot Oracle***, removes all demographic and behavioral cues but supplies the actual four responses, allowing the model to ground its simulation in user statements while lacking direct persona data. Finally, T'_8, the ***Zero-shot*** condition, excludes both personal information and true answers, evaluating how the model performs with virtually no contextual cues. Evaluating each digital twin across $\mathcal{T}'$ and the four queries in Table 1 allows us to gauge the influence of different configurations on the coherence, precision and consistency of the simulated person responses.

Table 1: Q&A Prompts for Digital Twin Updates

Question Dimension	Prompt
Numeracy	*"In a few sentences, please describe an experience in your life that demonstrated your knowledge of health or medical issues."*
Anxiety	*"In a few sentences, please describe what makes you feel most anxious or worried when visiting the doctor's office."*
Trust in Physician	*"In a few sentences, please explain the reasons why you trust or distrust your primary care physician. If you do not have a primary care physician, please answer in regard to doctors in general."*
Subjective Health Literacy	*"In a few sentences, please describe to what degree do you feel you have the capacity to obtain, process, and understand basic health information and services needed to make appropriate health decisions?"*

3.2 Evaluation Metrics

3.2.1 Simulated Person Response Similarity

Let t be a text document (e.g., a patient response), and let $f : \mathcal{T} \to \mathbb{R}^d$ be an embedding function provided by a pre-trained language model such as `BERT_CLS`, `MiniLM-L6-v2`, or `mpnet-base-v2`. For any text t, we obtain its embedding vector $\mathbf{v}$ via $\mathbf{v} = f(t)$.

In our setting, each user is asked one of four domain-specific questions pertaining to a specific health dimension (Table 1). Let $t_{\text{gen}}(q)$ be the LLM-generated response and $t_{\text{true}}(q)$ the corresponding ground-truth response for question q. We map each to its embedding space, yielding

$$\mathbf{v}_{\text{gen}}(q) = f\big(t_{\text{gen}}(q)\big) \quad (7)$$
$$\mathbf{v}_{\text{true}}(q) = f\big(t_{\text{true}}(q)\big) \quad (8)$$

We then compute their cosine similarity,

$$\text{sim}\big(t_{\text{gen}}(q), t_{\text{true}}(q)\big) = \frac{\langle \mathbf{v}_{\text{gen}}(q),\, \mathbf{v}_{\text{true}}(q)\rangle}{\|\mathbf{v}_{\text{gen}}(q)\|\;\|\mathbf{v}_{\text{true}}(q)\|}, \quad (9)$$

where $\langle \cdot, \cdot \rangle$ denotes the dot product, and $\| \cdot \|$ denotes the Euclidean norm. This similarity measure lies in the interval $\big[-1, 1\big]$, with higher values indicating stronger alignment between the generated response and the ground-truth text.

3.2.2 Downstream Prediction and Fairness

We assess the downstream prediction power and fairness of models fine-tuned using the simulated persona-twins versus actual users by drawing on the methodology described in Lalor et al. (2022), which focuses on quantifying prediction metrics and related intersectional biases across multiple demographic dimensions. Our evaluation framework includes the following components:

Model Fine-Tuning and Hyperparameter Settings. We fine-tuned BERT model for five epochs using a batch size of 32, a learning rate of 1e-5, and a weight decay of 0.01. The model that achieved the lowest validation loss was saved as the final model. This approach balances training quality and overfitting prevention. For each experimental setting, we conducted five-fold cross validation.

Performance and Fairness Metrics. In addition to standard performance metrics such as AUC, F1 score, mean squared error (MSE), and Pearson's correlation coefficient, we evaluated fairness using a series of disparate impact (DI) metrics. Specifically, DI scores were computed for individual demographic attributes: age, gender, race, education, and income, as well as for their intersectional combinations. These metrics help to reveal any biases in model predictions across different subgroups.

Collectively, the downstream prediction task is intended to highlight the inference potential and fairness of models trained on the constructed digital persona twins relative to the actual users.

4 Experiments

4.1 Datasets

For this study, we utilized the psychometric dataset from Abbasi et al. (2021). The dataset comprises survey-based psychometric measures alongside user-generated text, gathered from over 8,500 respondents. The primary psychometric dimensions measured include trust in physicians, anxiety visiting the doctor's office, health numeracy, and subjective health literacy. These dimensions are critical to understanding user behavior in healthcare and were selected based on their relevance to an array of health outcomes. The English-language dataset offers a rich blend of structured survey responses and unstructured text, including detailed demographic information (e.g., age, gender, race, education, and income) alongside psychometric and behavior measures. This enables a comprehensive analysis of how human factors influence text-based responses (e.g., Zhou et al., 2023; Gohar and Cheng, 2023; Dai et al., 2024; Van der Wal et al., 2024).[3]

4.2 Models

To generate the simulated responses t_{gen}, we employ two LLMs: `GPT-4o` and `Llama-3-70b`.[4] We then use each of the three pre-trained models (`bert-base-uncased`, `MiniLM-L6-v2`, and `mpnet-base-v2`) as embedding functions f to assess the quality of the generated text from multiple representational perspectives (Reimers and Gurevych, 2019). This way, we evaluate how faithfully the model output matches the user's actual responses, and also examine the robustness of our similarity scores to variations in the underlying embedding space.

4.3 Results on Fidelity of Responses

In this section, we report the similarity scores obtained from two groups of models, 4o-based and Llama-based, across five experimental conditions (*Persona Oracle*, *Few-shot Oracle*, *Persona Few-shot*, *Persona Zero-shot*, and *Zero-shot*) on four tasks (`Anxiety`, `Numeracy`, `Literacy`, and `TrustPhys`). Detailed data for 4o-based models and Llama-based models appear in Table 2. Figure 2 offers a visual comparison of performance across all tasks and conditions.

`PersonaTwin` Compared With Baselines. Our primary focus is on scenarios where the "twin" model does not receive the correct answers. In these cases, we compare three conditions: *Persona Few-shot* (which retains the full structure of

[3] The data collection protocol for (Abbas and Lichouri, 2021) was approved by the University of Virginia IRB-SBS under SBS Number 2017014300.

[4] Refer to Appendix A.2 for implementation details.

Condition	bert_CLS				sbert_MiniLM				sbert_mpnet			
	Anxiety	Numeracy	Lit	TrustPhys	Anxiety	Numeracy	Lit	TrustPhys	Anxiety	Numeracy	Lit	TrustPhys
						GPT-4o						
Persona Oracle	**0.952**	0.952	**0.970**	**0.965**	**0.535**	**0.291**	0.586	**0.589**	**0.599**	**0.361**	**0.647**	**0.683**
Few-Shot Oracle	0.946	0.951	0.968	0.962	0.504	0.285	**0.587**	0.562	0.575	0.354	0.644	0.660
Persona Few-shot	0.949*	**0.953***	0.968*	0.961*	0.490	0.272*	0.553	0.536*	0.556	0.337*	0.620*	0.641*
Persona Zero-shot	0.939*	0.943	0.964*	0.952	0.491	0.227	0.500	0.515	0.554	0.292	0.582	0.624
Zero-Shot	0.937	0.942	0.962	0.954	0.492	0.240	0.553	0.513	0.562	0.299	0.612	0.620
						Llama-3-70b						
Persona Oracle	**0.957**	**0.959**	**0.971**	**0.961**	**0.526**	0.325	**0.571**	**0.600**	**0.600**	0.383	**0.615**	**0.689**
Few-Shot Oracle	0.955	0.958	**0.971**	0.960	0.510	**0.330**	0.564	0.593	0.582	**0.385**	0.604	0.683
Persona Few-shot	0.955*	0.956*	0.969*	0.956*	0.486*	0.291*	0.544*	0.545*	0.555*	0.346	0.595*	0.650*
Persona Zero-shot	0.941	0.949*	0.966	0.956*	0.476	0.282*	0.517*	0.506	0.533*	0.327	0.577*	0.623*
Zero-Shot	0.931	0.942	0.967	0.950	0.476	0.277	0.510	0.503	0.522	0.306	0.533	0.609

Table 2: Similarity scores for GPT-4o (top) and Llama-3-70b (bottom) models across different conditions. $*$ indicates similarity scores significantly higher than the zero-shot baseline ($p < 0.05$).

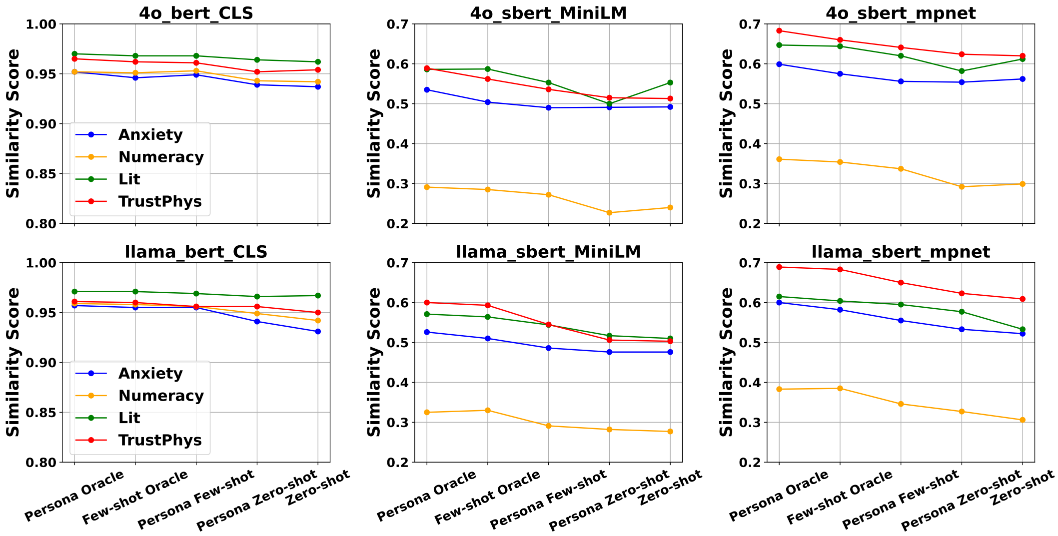

Figure 2: Comparison of similarity scores for 4o-based and Llama-based models under different conditions. The top row corresponds to the 4o-based models, and the bottom row corresponds to the Llama-based models. Each subplot includes results for the tasks: Anxiety, Numeracy, Lit, and TrustPhys.

Condition	ROUGE	A	N	SL	TP
Persona Oracle	1	0.232	0.201	0.252	0.256
	L	0.201	0.173	0.216	0.224
Few-shot Oracle	1	0.222	0.197	0.249	0.249
	L	0.192	0.171	0.212	0.218
Persona Few-shot	1	0.216	0.193	0.243	0.241
	L	0.185	0.168	0.209	0.211
Persona Zero-shot	1	0.187	0.164	0.206	0.193
	L	0.160	0.146	0.181	0.170
Zero-Shot	1	0.194	0.157	0.192	0.170
	L	0.171	0.144	0.165	0.150

Table 3: ROUGE-1 and ROUGE-L scores for persona-generated text. A: Anxiety, N: Numeracy, SL: Literacy, and TP: TrustPhys

`PersonaTwin`), *Persona Zero-shot* (which provides persona information without iterative dialogues), and *Zero-shot* (a baseline). For example, in the 4o-based models using the SBERT-MPNet metric, the *Persona Few-shot* condition achieves a similarity score of 0.337 on the `Numeracy` task, which is approximately 15% higher than the 0.292 observed under the *Persona Zero-shot* setting. Similarly, for the `Anxiety` task, the *Persona Few-shot* condition's score (0.949 using the BERT-based metric) is about 1.1% higher than that of the *Persona Zero-shot* condition (0.939) and nearly 1.3% higher than the *Zero-shot* condition (0.937). Comparable improvements are seen in the Llama-based models; for instance, using the SBERT-MPNet metric, the average score under *Persona Few-shot* is 0.5365, which represents roughly a 4-9% boost over the corresponding scores of the *Persona Zero-shot* (0.515) and *Zero-shot* (0.4925) conditions. These consistent gains, despite variations across metrics and

Condition	Model	MSE	Pearson's r	F1	AUC	DI_Age	DI_Gender	DI_Race	DI_Education	DI_Income	DI+	DI++
True Response	-	0.30	0.41	0.71	0.71	1.05	1.03	0.89	0.89	0.94	0.95	0.94
Persona Oracle	GPT-4o	0.34	0.32	0.64	0.66	1.08	1.02	0.95	0.84	0.87	0.92	0.89
	Llama-3-70b	0.33	0.35	0.67	0.67	1.14	1.02	0.89	0.82	0.84	0.90	0.88
Few-shot Oracle	GPT-4o	0.36	0.29	0.62	0.65	1.13	1.04	0.94	0.92	0.94	0.92	0.90
	Llama-3-70b	0.33	0.34	0.67	0.67	1.11	1.02	0.88	0.87	0.93	0.95	0.94
Persona Few-shot	GPT-4o	0.36	0.27	0.61	0.63	1.12	1.02	0.94	0.83	0.85	0.91	0.89
	Llama-3-70b	0.35	0.30	0.64	0.65	1.15	1.01	0.89	0.81	0.83	0.90	0.87
Persona Zero-shot	GPT-4o	0.43	0.12	0.44	0.56	1.01	0.98	0.97	0.86	0.89	0.92	0.91
	Llama-3-70b	0.47	0.10	0.47	0.55	1.00	0.99	1.02	0.81	0.80	0.91	0.99
Zero-Shot	GPT-4o	0.47	0.03	0.26	0.51	1.03	0.97	0.98	1.00	1.02	0.99	0.98
	Llama-3-70b	0.47	0.03	0.28	0.51	0.99	0.98	1.01	0.98	1.00	0.99	1.01

Table 4: Performance Metrics for Different Conditions and Models. “True Response” is common across models.

tasks, support the effectiveness of our persona twin approach.

We performed paired t-tests comparing the Persona Few-shot condition against the Zero-shot baseline across all model–metric–task configurations and found that Persona Few-shot significantly outperformed Zero-shot in 20 out of 24 comparisons ($p < 0.05$). This statistical analysis confirms that the observed improvements under the Persona Few-shot condition are robust.

Providing Detailed Persona Information Further Boosts Realism. We also carried out a supplementary experiment in which the correct user/patient answers are provided. In this setting, the *Persona Oracle* condition includes both the real answers and the comprehensive persona module, while the *Few-Shot Oracle* condition supplies the real answers without any persona details. Even with direct access to the actual responses, providing detailed persona information further boosts realism. For instance, in the 4o-based models using the SBERT-MPNet metric, the `Anxiety` score under *Persona Oracle* is 0.599—about 4% higher than the 0.575 observed under *Few-Shot Oracle*. Likewise, on the `TrustPhys` task, the *Persona Oracle* condition achieves a score of 0.683, which is roughly 3.5% higher than the 0.660 score of the baseline. Similar trends are evident in the Llama-based models. These findings show that the persona module prevents verbatim copying and enriches responses with context, enhancing overall fidelity. Interestingly, *Persona Few-shot*, our focal setting devoid of complete answer key information in the experiments, performs relatively close to the two oracle settings on many tasks, for all three similarity measures, across both LLMs.

Task-Specific Differences in Response Fidelity. Our analysis further reveals notable task-dependent differences in response fidelity. Across both model groups and multiple metrics, the `Numeracy` task consistently scores lower than the other tasks. For example, in the 4o-based models using the SBERT-MPNet metric, the `Numeracy` score under the *Persona Oracle* condition is 0.361, while the `TrustPhys` score reaches 0.683, indicating that the TrustPhys responses are nearly 90% higher in similarity. In contrast, the `Anxiety` and `Literacy` tasks typically yield intermediate scores. These task-specific disparities suggest that while our approach is highly effective at generating realistic responses in trust-related and narrative contexts, it remains more challenging to simulate numerical reasoning, which we aim to address in future work.

Lexical Quality Assessment via ROUGE Metrics. Table 3 presents ROUGE-1 and ROUGE-L scores measuring lexical overlap between generated and reference responses across the four psychometric tasks. The results demonstrate consistent superiority of persona-enhanced conditions over baseline approaches. The *Persona Oracle* condition achieves the highest ROUGE scores across all tasks, with ROUGE-1 scores ranging from 0.201 (`Numeracy`) to 0.256 (`TrustPhys`). The *Persona Few-shot* condition maintains competitive performance, achieving ROUGE-1 scores within 6-8% of the oracle condition across tasks.

Particularly noteworthy is the substantial performance gap between persona-enhanced conditions and baseline approaches. For the `TrustPhys` task, the *Persona Few-shot* condition (ROUGE-1 = 0.241) outperforms the *Zero-Shot* baseline (ROUGE-1 = 0.170) by approximately 42%, highlighting the significant contribution of persona information to response quality. Similar patterns emerge across all psychometric dimensions, with the `Numeracy` task again showing the most challenging characteristics, consistent with our earlier similarity score findings.

4.4 Downstream Prediction and Fairness

Table 4 reports selected performance and fairness metrics, including classification metrics (MSE, Pearson's r, F1, and AUC) and demographic parity indices (DI_Age, DI_Gender, DI_Race, DI_Education, and DI_Income). Here, DI+ represents the average fairness metric computed over two-way demographic interactions, while DI++ aggregates the fairness metrics over three-way interactions. Ideally, a DI value of 1 indicates that the positive response rates are balanced across demographic groups; values above 1 suggest an overrepresentation of positive responses, whereas values below 1 indicate underrepresentation.

Looking at the classification performance metrics, notably, *Persona Few-shot* attains error/accuracy/AUC rates that are not only comparable to the two oracle settings, but are also within 6-7 F1/AUC points of those attained using the actual person data (True Response setting). Regarding fairness, in the True Response condition, which reflects human responses, the DI metrics are relatively balanced (with DI+ = 0.95 and DI++ = 0.94), suggesting that the true data is close to evenly distributed across demographic groups. In the Persona Oracle, Few-Shot Oracle, and Persona Few-shot settings, both GPT-4o and Llama-3-70b yield DI values close to those of the True Response baseline, with aggregate metrics (DI+ and DI++) generally ranging between 0.87 and 0.99. This observation indicates that models trained on the generated persona twins do not substantially differ in demographic parity of model outputs. Similar fairness levels are observed for the Persona Zero-shot and Zero-shot settings, albeit with markedly lower prediction and/or classification performance.

Overall, these findings suggest that LLM-based persona twins have potential as a data augmentation and user modeling enrichment strategy for downstream NLP tasks. Although future work is needed to reduce the performance prediction and classification deltas (MSE, Pearson's r, AUC, F1) between Persona Few-shot and True Response, the demographic fairness of the models trained on the twins remain robust, with DI, DI+ and DI++ values near 1 across experimental settings. There may be future opportunities to further enhance twin-based model performance without compromising fairness.

4.5 Big Five Personality Trait Estimation

Table 5 in the appendices presents MSE scores for Big Five personality trait estimation across different experimental conditions. Lower MSE values indicate better alignment between predicted and actual personality traits. Our analysis reveals that the *Persona Oracle* condition consistently achieves the lowest MSE scores across most traits for both model families. For GPT-4o, the *Persona Oracle* condition demonstrates particularly strong performance in estimating Agreeableness (MSE = 1.2388) and Openness (MSE = 1.4314), while showing moderate effectiveness for Extraversion (MSE = 2.1415). Similarly, in Llama-3-70b models, the *Persona Oracle* condition excels in Stability estimation (MSE = 1.7264) and shows competitive performance across other traits.

Notably, the *Persona Few-shot* condition, which is our primary focus as it does not have access to ground truth answers, performs remarkably close to the oracle settings. For instance, in GPT-4o models, the *Persona Few-shot* condition achieves an MSE of 1.6430 for Stability estimation, which is only 3.6% higher than the oracle's 1.7049. This pattern holds consistently across both model families, suggesting that our approach can effectively capture personality nuances even without complete answer information. In contrast, the *Few-shot Oracle* condition, despite having access to correct answers but lacking persona details, shows notably higher MSE scores, particularly for Extraversion and Stability traits, reinforcing the value of incorporating comprehensive persona information.

5 Conclusion

We present PersonaTwin, a multi-tier prompt conditioning framework that enhances digital twin realism and fairness as demonstrated in a healthcare AI context. By combining structured persona encoding with iterative refinement, PersonaTwin generates context-aware responses with competitive downstream performance and fairness potential for fine-tuned NLP models relative to true responses. Extensive evaluations on 8,500 individuals demonstrate significant improvements in simulation fidelity, and maintaining fairness with demographic parity indices consistently ranging between 0.87 and 1.01 across different model architectures. Future directions include expanding psychometric dimensions and enabling real-time adaptation for more accurate downstream predictive power.

Limitations

While PersonaTwin provides a robust foundation for personalized digital twins in healthcare, some areas deserve further attention. First, our framework was evaluated on data in English drawn from a single large-scale psychometric data set. Adapting it to other languages or healthcare settings, particularly those with more complex morphology or differing cultural norms, could involve additional tuning and validation.

Second, although we incorporate multiple tiers of patient information (demographic, behavioral, and psychological), our approach may require certain data formats to be consistently available. In practice, some healthcare settings might present incomplete or heterogeneous records, which could reduce simulation fidelity. Future work could explore data imputation strategies and domain adaptation to maintain robust personalization under such constraints.

Lastly, our fairness checks focus on group-level biases (e.g., by race, age, and income). Although these metrics suggest that deeper contextual data do not inherently exacerbate demographic disparities, we have not exhaustively examined all possible bias dimensions or intersectional factors. Further research could extend these fairness assessments and investigate more granular social determinants of health to ensure that *PersonaTwin* remains equitable between diverse populations of patients.

Ethics Statement

All experimental protocols in this study adhered to established ethical guidelines for handling sensitive health-related data. The psychometric data set we used was fully deidentified and was obtained under appropriate data sharing agreements, ensuring the privacy and confidentiality of the respondents. Moreover, the PersonaTwin multi-tier prompt conditioning approach is designed to mitigate the risk of harmful biases by incorporating fairness assessments that monitor model outputs across sensitive demographic attributes. Although our framework aims to improve personalized healthcare applications, we recognize that any generative technology carries potential misuse risks (e.g., perpetuating biases not captured by our metrics). Consequently, we recommend that health organizations and clinicians applying PersonaTwin maintain rigorous supervision to ensure accountability and respect for patient autonomy and consent. The methods and results reported here comply with the ACL Ethics Policy.[5]

References

Mourad Abbas and Mohamed Lichouri. 2021. TPT: An empirical term selection for Arabic text categorization. In *Proceedings of the Fourth International Conference on Natural Language and Speech Processing (ICNLSP 2021)*, pages 226–231, Trento, Italy. Association for Computational Linguistics.

Ahmed Abbasi, David Dobolyi, John P. Lalor, Richard G. Netemeyer, Kendall Smith, and Yi Yang. 2021. Constructing a psychometric testbed for fair natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3748–3758, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Israa Abuelezz, Mahmoud Barhamgi, Zied El Houki, Khaled M Khan, and Raian Ali. 2024. Qualitative exploration of factors influencing trust and engagement in social engineering: The role of visual and demographic cues. In *2024 11th International Conference on Behavioural and Social Computing (BESC)*, pages 1–8. IEEE.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and machine learning: Limitations and opportunities*. MIT press.

Marco Cascella, Jonathan Montomoli, Valentina Bellini, and Elena Bignami. 2023. Evaluating the feasibility of chatgpt in healthcare: an analysis of multiple clinical and research scenarios. *Journal of Medical Systems*, 47(1):33.

Yi-Pei Chen, Noriki Nishida, Hideki Nakayama, and Yuji Matsumoto. 2024. Recent trends in personalized dialogue generation: A review of datasets, methodologies, and evaluations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13650–13665.

[5] https://www.aclweb.org/portal/content/acl-code-ethics

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Yun-Shiuan Chuang, Krirk Nirunwiroj, Zach Studdiford, Agam Goyal, Vincent V. Frigo, Sijia Yang, Dhavan V. Shah, Junjie Hu, and Timothy T. Rogers. 2024. Beyond demographics: Aligning role-playing LLM-based agents using human belief networks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14010–14026, Miami, Florida, USA. Association for Computational Linguistics.

Yiwei Dai, Hengrui Gu, Ying Wang, and Xin Wang. 2024. Mitigate extrinsic social bias in pre-trained language models via continuous prompts adjustment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11068–11083.

Thomas H. Davenport and R. Kalakota. 2019. The potential for artificial intelligence in healthcare. *Future Healthcare Journal*, 6(2):94–98.

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

Usman Gohar and Lu Cheng. 2023. A survey on intersectional fairness in machine learning: notions, mitigation, and challenges. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 6619–6627.

Michael Grieves. 2014. Digital twin: manufacturing excellence through virtual factory replication. *White paper*, 1(2014):1–7.

Qian Guo and Peiyuan Chen. 2024. Construction and optimization of health behavior prediction model for the older adult in smart older adult care. *Frontiers in Public Health*, 12.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Yuanzhe Huang, Saurab Faruque, Minjie Wu, Akiko Mizuno, Eduardo Diniz, Shaolin Yang, George Dewitt Stetten, Noah Schweitzer, Hecheng Jin, Linghai Wang, et al. 2024. Leveraging the finite states of emotion processing to study late-life mental health. *arXiv preprint arXiv:2403.03414*.

Pegah Jandaghi, Xianghai Sheng, Xinyi Bai, Jay Pujara, and Hakim Sidahmed. 2024. Faithful persona-based conversational dataset generation with large language models. In *Proceedings of the 6th Workshop on NLP for Conversational AI (NLP4ConvAI 2024)*, pages 114–139, Bangkok, Thailand. Association for Computational Linguistics.

Fei Jiang, Yadong Jiang, Hui Zhi, Yahui Dong, Haifeng Li, Sheng Ma, Yuanting Wang, et al. 2017. Artificial intelligence in healthcare: Past, present and future. *Stroke and Vascular Neurology*, 2(4):230–243.

John Lalor, Yi Yang, Kendall Smith, Nicole Forsgren, and Ahmed Abbasi. 2022. Benchmarking intersectional biases in NLP. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3598–3609, Seattle, United States. Association for Computational Linguistics.

Liliana Laranjo, Adam Dunn, Huong Ly Tong, A. Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie Lau, and Enrico Coiera. 2018. Conversational agents in healthcare: A systematic review. *Journal of the American Medical Informatics Association*, 0.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 47–58, Toronto, Canada. Association for Computational Linguistics.

Marian Lukaniszyn, Łukasz Majka, Barbara Grochowicz, Dariusz Mikołajewski, and Aleksandra Kawala-Sterniuk. 2024. Digital twins generated by artificial intelligence in personalized healthcare. *Applied Sciences*, 14:9404.

Charles Meijer, Hae-Won Uh, and Said el Bouhaddani. 2023. Digital twins in healthcare: Methodological challenges and opportunities. *Journal of Personalized Medicine*, 13:1522.

Nicole Meister, Carlos Guestrin, and Tatsunori Hashimoto. 2024. Benchmarking distributional alignment of large language models. *arXiv preprint arXiv:2411.05403*.

John Mendonça, Isabel Trancoso, and Alon Lavie. 2024. Soda-eval: Open-domain dialogue evaluation in the age of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11687–11708, Miami, Florida, USA. Association for Computational Linguistics.

Gabriel Oliveira dos Santos, Esther Luna Colombini, and Sandra Avila. 2021. CIDEr-R: Robust consensus-based image description evaluation. In *Proceedings* *of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 351–360, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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

Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. 2024. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867.

Aleksandra Sorokovikova, Natalia Fedorova, AI Toloka, Sharwin Rezagholi, Technikum Wien, and Ivan P Yamshchikov. 2024. Llms simulate big five personality traits: Further evidence. In *The 1st Workshop on Personalization of Generative AI Systems*, page 83.

Oskar Van der Wal, Dominik Bachmann, Alina Leidinger, Leendert van Maanen, Willem Zuidema, and Katrin Schulz. 2024. Undesirable biases in nlp: Addressing challenges of measurement. *Journal of Artificial Intelligence Research*, 79:1–40.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.

Frank F Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Z Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, et al. 2024. Theagentcompany: benchmarking llm agents on consequential real world tasks. *arXiv preprint arXiv:2412.14161*.

Tianyi Zhang, Vikas Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.

Fan Zhou, Yuzhou Mao, Liu Yu, Yi Yang, and Ting Zhong. 2023. Causal-debias: Unifying debiasing in pretrained language models and fine-tuning via causal invariant learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4227–4241.

Wanzheng Zhu and Suma Bhat. 2020. GRUEN for evaluating linguistic quality of generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 94–108, Online. Association for Computational Linguistics.

A Appendix

A.1 Detailed Information Provided to `PersonaTwin` as Persona

In this study, we developed and tested a series of prompts aimed at simulating and understanding the influence of various combinations of demographic, behavioral, and psychological factors on the modeling of group personas. The prompts were meticulously crafted to reflect different configurations of participant characteristics, enabling us to systematically assess the impact of these factors on the accuracy and relevance of the generated responses.

A.1.1 Demographic Information

We included a comprehensive set of demographic variables to capture the foundational characteristics of the participants. The demographic variables tested were:

- **Age:** Ranging from 18 to 99 years.
- **Sex:** Male or Female.
- **Race:** Categories such as White, Black or African American, Asian, Native American or American Indian, Native Hawaiian or Pacific

Islander, Multiracial or Biracial, Other, and Prefer not to answer.

- **Education:** Levels ranging from education lower than college to higher education.

- **Income:** Income brackets ranging from less than $20,000 to $90,000 or more, including options for uncertainty or preference not to answer.

A.1.2 Behavioral Information

To capture participants' habits and lifestyle choices, which could influence their health and psychological state, we included the following behavioral variables:

- **Prescription drug usage:** Number of prescription drugs taken regularly.

- **Primary care physician status:** Whether the participant has a primary care physician.

- **Frequency of visits to primary care physician:** Number of visits in the past two years.

- **Physical activity:** Average hours per week of physical exercise or activity.

- **Eating habits:** Overall healthiness of eating habits.

- **Smoking and alcohol consumption:** Frequency of smoking and drinking.

- **Health consciousness:** Attitudes towards health and preventive measures.

- **Overall health:** Self-assessed overall health.

A.1.3 Psychological Information

Psychological variables were incorporated to explore deeper aspects of the participants' mental states and outlooks. These variables included:

- **Personality traits:** Self-assessment on key personality dimensions (Extraverted, enthusiastic; Agreeable, kind; Dependable, organized;Emotionally stable, calm; Open to experience, imaginative)

A.2 LLM & Data Collection Details

We used the OpenAI API for GPT-4o with top_p set to 1, max_tokens set to 200, min_tokens set to 0, and temperature set to 0.6 (with all other parameters at their default values), and the Replicate API for Llama-3-70b with top_p set to 0.9, max_tokens set to 200, min_tokens set to 0, and temperature set to 0.6.

The data collection protocol for this project was approved by the University of Virginia IRB-SBS under SBS Number 2017014300.

B Additional Experimental Results

B.1 Big Five Personality Trait Estimation

As shown in Table 5, we evaluate PersonaTwin's ability to predict missing Big Five trait scores by reporting mean squared error (MSE) against gold labels. Persona Few-shot consistently outperforms Persona Zero-shot across all five dimensions and approaches the performance of the Persona Oracle, demonstrating the framework's flexibility and accuracy when handling incomplete persona data.

B.2 Text Generation ROUGE Evaluation

Table 3 presents ROUGE-1 and ROUGE-L scores for persona-generated text under each condition. Persona Few-shot yields higher ROUGE scores than both Zero-Shot and Persona Zero-shot across all tasks, confirming that incorporating existing persona dimensions into few-shot prompts improves alignment with reference outputs.

B.3 Downstream Task Evaluation

Table 6 summarizes performance on downstream prediction tasks—MSE, Pearson's r, F1, and AUC—along with percentage lift over the Zero-Shot baseline. Persona Few-shot delivers substantial gains across all metrics (up to 900% lift in Pearson's r), while Persona Zero-shot also outperforms pure Zero-Shot, illustrating the clear downstream benefits of generating text with enriched persona information.

	MSE Scores for Big Five Trait Estimation				
Model	Extraverted	Agreeable	Conscientious	Stable	Open
		GPT-4o			
Persona Oracle	2.1415	1.2388	1.5742	1.7049	1.4314
Few-shot Oracle	3.0926	1.2983	1.5987	2.1525	1.3573
Persona Few-shot	2.1528	1.2666	1.5834	1.6430	1.4392
Persona Zero-shot	2.5386	1.5271	2.0106	2.0165	2.0232
		Llama-3-70b			
Persona Oracle	1.8920	1.9153	2.0244	1.7264	1.6746
Few-shot Oracle	3.1451	1.7787	3.5147	2.9706	1.7339
Persona Few-shot	1.9303	1.8510	2.0556	1.6242	1.6657
Persona Zero-shot	2.5028	1.5366	2.2742	1.9523	1.6144

Table 5: MSE for Big Five personality trait estimation (lower is better).

Condition	**Model**	MSE	Lift	Pearson's r	Lift	F1	Lift	AUC	Lift
Persona Few-Shot	GPT-4o	0.36	23.4%	0.27	800.0%	0.61	134.6%	0.63	23.5%
	Llama-3-70b	0.35	25.5%	0.30	900.0%	0.64	128.6%	0.65	27.5%
Persona Zero-Shot	GPT-4o	0.43	8.5%	0.12	300.0%	0.44	69.2%	0.56	9.8%
	Llama-3-70b	0.47	0.0%	0.10	233.3%	0.47	67.9%	0.55	7.8%
Zero-Shot	GPT-4o	0.47	–	0.03	–	0.26	–	0.51	–
	Llama-3-70b	0.47	–	0.03	–	0.28	–	0.51	–

Table 6: Downstream task metrics and lift over zero-shot baseline.

Stage	Component	Description
Input Data	Demographics	Age=25, Gender=Male, Race="Black or African American", Education="College graduate", Income="$20,000-$34,999"
	Behavioral	HealthImportance=3/5, PreventionBelief=2/5, SelfCareValue=3/5
	Psychological	Extraversion=5/5, Agreeableness=4/5, EmotionalStability=5/5
Template Application	Template_dem	"You are 25 years old, male, of Black or African American descent. You have a college degree and an annual income of $20,000-$34,999."
	Template_beh	"You find it moderately important to live in the best possible health. You think that maintaining a healthy lifestyle may or may not guarantee lifelong health."
	Template_psy	"You strongly agree that you are extraverted and enthusiastic. You agree that you are agreeable and kind. You strongly agree that you are emotionally stable and calm."
Initial Generation	System Prompt	The concatenated templates form the system prompt (P) for the LLM, generating the initial digital twin (T_0).
Conversation Integration	Health Literacy	**Q**$_1$**:** "Please describe to what degree you can obtain and understand health information for decisions." **R**$_1$**:** "When I visit a doctor I try to get as much information that is needed for my health... I tend to ask a lot of questions." *Updates T_0 to include information-seeking behavior*
	Trust Assessment	**Q**$_2$**:** "Please explain why you trust or distrust your primary care physician." **R**$_2$**:** "Sometimes I think they take things out of control because everyone's body is different..." *Updates T_1 to reflect medication skepticism*
	Anxiety Assessment	**Q**$_3$**:** "What makes you feel anxious when visiting the doctor's office?" **R**$_3$**:** "To find out what is wrong with me and sometimes I don't want to hear the truth..." *Updates T_2 to include contextual anxiety*
	Health Knowledge	**Q**$_4$**:** "Describe an experience demonstrating your knowledge of health issues." **R**$_4$**:** "I have asthma which often has me rush to the doctor for check ups..." *Updates T_3 to include chronic condition management*
Final State	Medical History	The final digital twin T_4 incorporates asthma as a chronic condition
	Healthcare Attitudes	Information-seeking but skeptical of medical interventions
	Emotional Responses	Contextualized anxiety about potential diagnoses

Table 7: An Example of Multi-Tiered Template Functions in PersonaTwin. The table demonstrates how raw input data is transformed through template functions and conversation integration to create an evolving digital twin.