TwinVoice: A Multi-dimensional Benchmark Towards Digital Twins via LLM Persona Simulation

Anonymous ACL submission

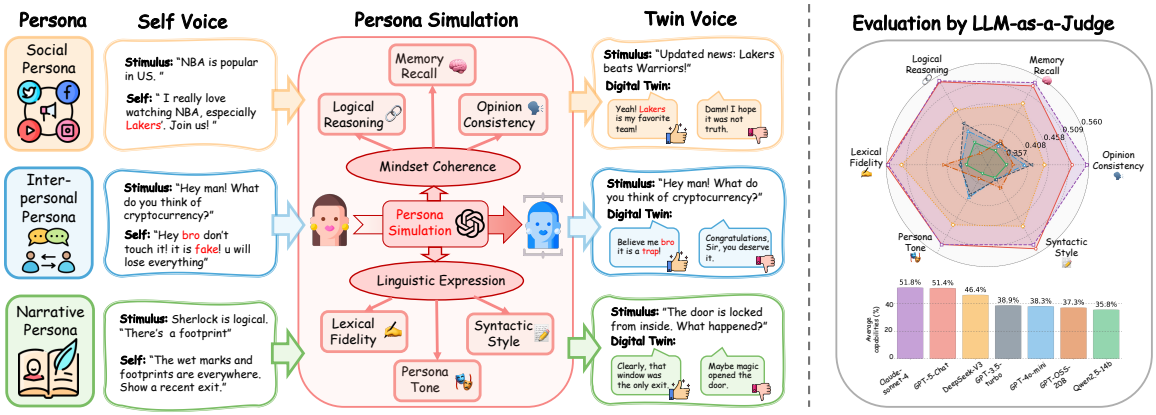


Figure 1: **The conceptual framework of TwinVoice: (Left)** The evaluation is structured across three ***dimensions*** that represent different aspects of a persona: (1) ***Social Persona*** that reflects its behavior on social platforms; ***Interpersonal Persona*** that reflects its private interaction, and ***Narrative Persona*** that corresponds to a fictional scenario. The LLMs are prompted with a person's historical context to simulate their behavior and evaluated by six fundamental ***capabilities***: memory recall, logical reasoning, opinion consistency, lexical fidelity, persona tone, and syntactic style. **(Right)** Experimental results on modern LLMs averaged over three dimensions.

Abstract

Large Language Models (LLMs) are exhibiting emergent human-like abilities and are envisioned as the tool for simulating an individual's communication patterns, behaviors, and personality traits. However, current evaluations of LLM-based persona simulation remain limited: most rely on synthetic dialogues and lack fine-grained analysis of the capability for persona simulation. To address these limitations, we introduce TwinVoice, a comprehensive benchmark for assessing persona simulation across diverse real-world contexts. TwinVoice encompasses three dimensions: Social Persona (public social interactions), Interpersonal Persona (private dialogues), and Narrative Persona (role-based expression). It further decomposes the evaluation into six fundamental capabilities, including opinion consistency, memory recall, logical reasoning, lexical fidelity, persona tone, and syntactic style. Experimental results reveal that while advanced models achieve moderate accuracy in persona simulation, they still fall short of capabilities such as syntactic style and memory recall. Our data, code, and evaluation results are available[1].

[1] https://anonymous.4open.science/r/TwinVoice-B08E

1 Introduction

Large Language Models (LLMs) are rapidly evolving from text generation tools into human-like agents (Bubeck et al., 2023; Wei et al., 2022; Chang et al., 2024). Existing studies have shown that the most advanced LLMs are capable of producing text indistinguishable from human writing (Jones and Bergen, 2025; Jones et al., 2025; Jones and Bergen, 2024). Consequently, the research focus is shifting toward a highly specific challenge:

Can we construct "digital twins" of specific individuals that are indistinguishable from themselves?

To address this challenge, the primary technical path is through LLM-based persona simulation, which replicates a person's unique style of talking, behavior, and personality (Shanahan et al., 2023; Park et al., 2023) based on their historical behavioral data. This technology promises to unlock a series of applications, including highly personalized assistants (Ma et al., 2023; Li et al., 2025a), social simulations (Li et al., 2023; Ran et al., 2025), healthcare (Barricelli et al., 2020), and marketing (Hornik and Rachamim, 2025). Despite growing interest in creating digital twins with

LLM-based persona simulation, its current ability remains unexplored due to the lack of systematic evaluation frameworks (Toubia et al., 2025; Zhou et al., 2025).

Recently, a series of benchmarks have been proposed to evaluate LLM's ability in imitating and predicting human behaviors. For example, BehaviorChain (Li et al., 2025b) evaluates continuous persona-based behavior by requiring models to iteratively predict the next action given a persona profile and history. Similarly, Human Simulacra and PersoBench assess human-likeness and personalized response quality, while other studies probe persona-driven decision making, counterfactual adherence, and large-scale dynamic profiling (Xie et al., 2025; Afzoon et al., 2024; Xu et al., 2024; Kumar et al., 2025; Jiang et al., 2025). However, existing benchmarks face limitations in both their scope and granularity. On the one hand, the predominant reliance on synthetic dialogues (Shen et al., 2023; Tu et al., 2024) prevents benchmarks from capturing the rich expression of human identity across diverse real-world contexts (***Scope Limitation***). On the other hand, current benchmarks often assess persona simulation simply based on an LLM's accuracy in predicting human behavior. This leaves a critical gap in understanding the fundamental capabilities—such as memory, reasoning, and lexical fidelity–that a model is expected to possess for persona simulation (***Granularity Limitation***).

To address those limitations, we introduce **TwinVoice**, a comprehensive benchmark designed for realistic and fine-grained persona evaluation (see Figure 1). For the scope limitation, TwinVoice is grounded across three complementary dimensions in persona simulation: Social Persona, Interpersonal Persona, and Narrative Persona. The Social Persona dimension leverages real-world social media data to evaluate a public-facing identity, while the Interpersonal Persona dimension utilizes multi-session dialogue data to assess a more private, relational self. While these two dimensions are grounded in authentic digital footprints, the Narrative Persona dimension is designed to complement such data with fictional scenarios to test behaviors and narrative consistency in more diverse contexts. For the granularity limitation, TwinVoice shifts from the holistic accuracy-based evaluation to a capability-level assessment. Building on psycholinguistic evidence that language conveys both what people say and how they say it (Pennebaker et al., 2003), we group persona fidelity into Mindset Coherence and Linguistic Expression, comprising six fundamental capabilities. Mindset Coherence assesses the logical and factual consistency of the content, including Opinion Consistency (Zaller, 1992), Memory Recall (Clark and Brennan, 1991), and Logical Reasoning (Kahneman, 2011). Linguistic Expression evaluates the language's stylistic form, encompassing Lexical Fidelity (Mehl et al., 2006; Koppel et al., 2009), Persona Tone (Brown, 1987), and Syntactic Style (Biber, 1995). Based on the above design, TwinVoice further benchmarks LLMs' capabilities on both discriminative-based and generative-based evaluations. The discriminative-based evaluation is based on the accuracy of LLMs on behavior prediction, while the generative-based evaluation is conducted by comparing LLM's output with ground truth via an LLM-as-a-Judge paradigm.

We test a series of state-of-the-art LLMs on TwinVoice and reveal several key insights into current capabilities and limitations in persona simulation with LLMs. On discriminative-based evaluation, GPT-3.5-Turbo averages an accuracy of 47.5%, while advanced models reach 71.2% for GPT-5 and 76.2% for Claude-Sonnet-4 (Anthropic, 2025). In the generative-based evaluation, GPT-5 (OpenAI, 2025) leads with 48.5% judged accuracy and a 2.13 pairwise score, with Claude-Sonnet-4 close at 47.9% and 2.12. Across all evaluations, we observe that performance dispersion across different LLMs is large, indicating high discriminative power of TwinVoice. However, these LLMs still lag behind human performance. Based on a subset under the discriminative-based evaluation, humans' majority vote accuracy achieves 66.0%, which is higher than GPT-5's performance of 60.0%. Across all fine-grained capabilities, we observe that LLMs perform best on Lexical Fidelity and Opinion Consistency and worst on Persona Tone and Memory Recall. This indicates the core limitations of modern LLMs for persona simulation.

Contributions of this work are threefold: **(1)** We introduce TwinVoice, a comprehensive benchmark for evaluating LLM-based persona simulation across multiple real-world scenarios with systematic competency decomposition; **(2)** We develop novel evaluation methodologies and categorize LLM's ability for persona simulation into six fine-grained capabilities; and **(3)** We provide extensive empirical analysis showing the limitations of the most advanced LLMs in person simulation and

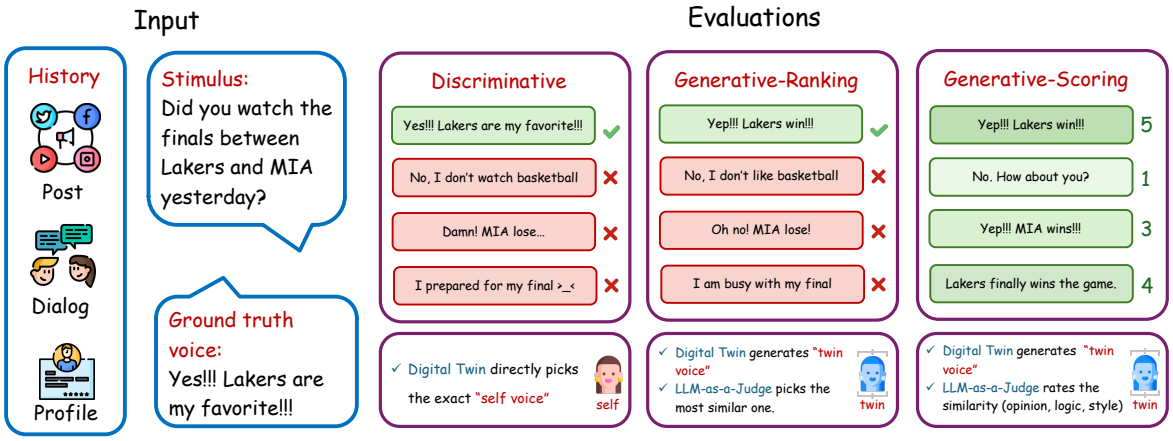


Figure 2: **TwinVoice experiment evaluation overview: Left**: The LLMs are prompted with a specific persona's history and tasked with a stimulus. **Right**: Three evaluation settings: Discriminative: the LLMs answer a multi-choice question, where the correct choice is the ground truth persona behavior. The discriminative-based evaluation performance is assessed by Acc. Generative-Ranking: the LLMs write an output and an LLM-as-Judge selects the best candidate, yielding Acc.(Gen). Generative–Scoring: the LLMs write and the Judge rates the similarity of the output and the ground truth, yielding Score (Gen).

offer crucial insights for advancing personalized AI systems.

2 Task Formulation

2.1 Problem Definition

TwinVoice evaluates LLMs' ability to simulate human personas through a unified task paradigm that captures the essence of digital twin functionality. Formally, we define the persona simulation task as follows:

Given a persona's historical data $\mathcal{H} = \{h_1, h_2, \ldots, h_n\}$ and a current stimulus s, the history is instantiated per dimension (Social, Interpersonal, or Narrative) as social posts, multi-session conversations, or narrative materials, respectively. The objective is to generate a response r that maximally approximates the ground truth response r^* that the original persona would produce in stimulus s, which can be formulated as an optimization problem:

$$r^* = \arg\max_r P(r|\mathcal{H}, s, \theta_{\text{persona}}), \qquad (1)$$

where θ_{persona} represents the latent persona characteristics learned from historical data $\mathcal{H}$. The evaluation objective is to assess the extent to which an LLM M can approximate this optimal response:

$$\text{Score} = f_{\text{sim}}(M(\mathcal{H}, s), r^*), \qquad (2)$$

where f_{sim} denotes a similarity function that measures persona consistency across multiple dimensions.

TwinVoice instantiates this general framework across three dimensions, each defined by its history source and interaction stimulus:

Across all three settings, we adopt a capability-centric evaluation rather than a single holistic score. The decomposition and scoring criteria are detailed in Section 3.2.

2.2 Evaluation Methodology

We combine both discriminative and generative evaluations to enhance the evaluation effectiveness. The discriminative-based evaluation is based on a single-choice question as stimuli. On the other hand, the stimuli of the generative-based evaluation are open-ended questions, and the LLM's response is evaluated via an LLM-as-a-Judge.

2.2.1 Discriminative Evaluation

The discriminative evaluation transforms the generation task into a single-answer multiple-choice selection problem. For each test instance (s, r^*), we construct a candidate set $\mathcal{C} = \{r^*, r_1, r_2, r_3\}$ where r^* is the ground truth response and $\{r_1, r_2, r_3\}$ are distractors. The evaluated LLM must select the most persona-consistent response from the shuffled candidate set.

The construction of distractors varies across dimensions to ensure realistic evaluation scenarios:

- **Social Persona:** Distractors are sampled from authentic responses by other users to similar posts, preserving topical relevance while introducing stylistic and opinion variations.

Persona Dimensions

Social Persona. In this dimension, a user's historical social media posts and comments are used to construct $\mathcal{H}^{\text{social}} = \{h_1^{(\text{social})}, h_2^{(\text{social})}, \ldots, h_m^{(\text{social})}\}$, and the current stimulus s is a new post. The challenge lies in maintaining stylistic consistency and opinion alignment in public discourse.

Interpersonal Persona. Here, multi-session conversational history is used to construct $\mathcal{H}^{\text{inter}} = \{h_1^{(\text{inter})}, h_2^{(\text{inter})}, \ldots, h_k^{(\text{inter})}\}$ where each $h_i^{(\text{inter})}$ represents a dialogue session. The current stimulus s is a new utterance from a conversation partner, requiring the model to generate contextually appropriate responses while maintaining conversational authenticity and memory-grounded consistency.

Narrative Persona. In this dimension, character background information and behavioral records are used to construct $\mathcal{H}^{\text{narra}} = \{h_1^{(\text{narra})}, h_2^{(\text{narra})}, \ldots, h_l^{(\text{narra})}\}$ where each $h_i^{(\text{narra})}$ denotes either background information or a prior action. The stimulus s describes a narrative scenario requiring character reaction, testing the model's ability to maintain role-based expression fidelity.

- **Interpersonal Persona:** Distractors are selected from real conversational responses in similar contexts, maintaining conversational appropriateness while differing in personal characteristics.

- **Narrative Persona:** Distractors are generated using advanced LLMs with alternative character interpretations, ensuring narrative coherence while diverging from the target persona's behavioral patterns.

The performance of discriminative evaluation is measured by the LLM's accuracy:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[M(\mathcal{H}_i, s_i) = r_i^*], \quad (3)$$

where N is the total number of test instances and $\mathbf{1}[\cdot]$ is the indicator function.

2.2.2 Generative Evaluation

Real-world digital twin applications require open-ended generation capabilities beyond the discriminative-based evaluation. Hence, we test LLMs' ability in a generative-based evaluation by employing LLM-as-a-Judge (Gu et al., 2024; Ye et al., 2025) to assess response quality along multiple dimensions.

The generative-based evaluation consists of two distinct judging approaches:

Generative-Ranking. The LLM-as-a-Judge identifies the most consistent response from a candidate set containing the generated response and the same distractors used in discriminative evaluation. Then the LLM's performance is evaluated based on the accuracy of whether its generated response is ranked as the most consistent among all candidates.

Generative-Scoring. The LLM-as-a-Judge rates generated responses against ground truth using structured evaluation criteria. Given a stimulus s, generated response r_{gen}, and ground truth r^*, the LLM-as-a-Judge assigns a score on a 1–5 scale based on three key dimensions: opinion consistency, logical coherence, and stylistic fidelity. The scoring rubric emphasizes faithful persona replication, with higher scores awarded to responses that demonstrate comprehensive alignment across all dimensions.

The generative evaluation score is computed as:

$$\text{Score}_{\text{gen}} = \frac{1}{N}\sum_{i=1}^{N} \text{Judge}(r_{\text{gen},i}, r_i^*, s_i), \quad (4)$$

where $\text{Judge}(\cdot)$ represents the LLM-as-a-Judge for the generative-scoring or generative-ranking setups. In this paper, we implement the LLM-as-a-Judge with GPT-5. We conduct additional experiments to demonstrate that the proposed LLM-as-a-Judge aligns with human judgment and exhibits minimal self-bias when evaluating its own outputs.

3 Benchmark Construction

3.1 Data Pre-processing

We construct the dataset with different protocols for three dimensions. Dataset statistics are presented in Table 1.

Social Persona. We construct the social persona benchmark based on the PChatbot Chinese microblog corpus (Qian et al., 2021). To ensure in-

Table 1: Dataset statistics across three dimensions. Each instance corresponds to a unique persona (#Users = #Instances). Avg = average; Gen = generative; Disc = discriminative. Token counts include instruction templates.

Dimension	Instances	Avg history turns	Avg prompt tokens (Disc)	Avg prompt tokens (Gen)
Social Persona	2000	15.0	1371.1	1215.2
Interpersonal Persona	2500	30.0	1163.5	1139.4
Narrative Persona	1187	15.7	934.3	910.7

stance quality, we selected users with rich histories (avg. reply length > 10 chars; Type-Token Ratio not in the bottom 20th percentile) and unambiguous choices (option cosine similarity < 0.95). We ranked all samples based on the similarity between the true response and the nearest distractor. Then, the 2,000 samples ranked at the bottom were selected to ensure the discriminability of the constructed candidates.

Interpersonal Persona. The dimension of interpersonal persona is constructed based on the Pushshift Telegram corpus (Baumgartner et al., 2020), which contains personalized dialogue sessions in different channels. We applied a multi-stage filtering strategy to distill a high-quality message set from 438,975 raw messages. We first selected active users engaged in ≥ 3 channels and have submitted > 500 total messages. We retained only the top 10% most informative utterances (removing lengths < 5 tokens), and applied semantic deduplication (threshold 0.90), yielding 6,150 messages. Finally, we extracted 2,500 multilingual tasks (including several languages like EN, RU, ES, PT) and prompted GPT-5 to generate the distractors. We incorporated users' cross-channel history as memory to evaluate consistency across contexts.

Narrative Persona. We selected eight novels from the Project Gutenberg corpus (Project Gutenberg, 1971–) to test the model's ability to mimic the speaking styles of the given characters. From these novels, we extracted 1,187 speech segments covering more than 50 characters. We first segmented novels into short, indexed chunks, and from each chunk we extracted at most one utterance together with its context. We then matched the speakers to the list of main characters, whose profiles contained their personality traits, goals, motivations, and utterance histories. Finally, we combined these speech segments with relevant character profiles and constructed the test stimuli based on the segment content and the character profiles. The distractors to the ground truth response are generated by selecting from the other available speech segments.

3.2 Capability Decomposition

Guided by psycholinguistic evidence that language conveys both what people say (content) and how they say it (style) (Pennebaker et al., 2003), we coarsely group persona fidelity into two complementary dimensions: *mindset coherence* and *linguistic expression*. This view is consistent with stable individual differences in language documented across psychology and linguistics and their computational operationalizations (Costa and McCrae, 1992; Biber, 1991; Stamatatos, 2009; Neuman, 2016; Li et al., 2016). We instantiate these via **six fundamental capabilities**: mindset coherence comprises Opinion Consistency (Zaller, 1992), Memory Recall (Clark and Brennan, 1991), and Logical Reasoning (Kahneman, 2011), whereas linguistic expression comprises Lexical Fidelity (Mehl et al., 2006; Koppel et al., 2009), Persona Tone (Brown, 1987), and Syntactic Style (Biber, 1995).

We employ a prompt-aligned rubric: for each instance, annotators choose exactly one primary capability and independently assess all six capabilities as true or false under strict criteria. Capabilities are non-orthogonal by design, so a data sample can reflect multiple capabilities. Full instructions, criteria, and prompt excerpts appear in Appendix B, with seed examples and the JSON output format for reproducibility.

4 Experiments

4.1 Overall Results

We present the main results in Table 2, results of fine-grained capabilities in Figure 3, and text-similarity metrics in Appendix G. From Table 2, **we observe that state-of-the-art models, notably GPT-5-Chat and Claude-Sonnet-4, lead the performance for both discriminative and generative-based evaluation.** This indicates that the most advanced models usually achieve better performance. However, the accuracy of the generative-based evaluation remains lower than that of the discriminative-based evaluation. This

Table 2: **Benchmark results for Digital Twin models:** We evaluate models using three distinct metrics: **Acc. (%)** is the accuracy of the discriminative evaluation. **Acc. (Gen) (%)** is the accuracy where a generative model's output is evaluated via multiple choice questions by an LLM-as-a-Judge. **Score (Gen)** is a pairwise comparison score against the ground truth for generative outputs by an LLM-as-a-Judge. Higher values indicate better performance. The best result and the second best result are in **Bold** and underlined, respectively.

Model / Tasks		Social			Interpersonal			Narrative			Average		
		Acc. (%)	Acc. (Gen)(%)	Score (Gen)	Acc. (%)	Acc. (Gen)(%)	Score (Gen)	Acc. (%)	Acc. (Gen)(%)	Score (Gen)	Acc. (%)	Acc. (Gen)(%)	Score (Gen)
LLM	GPT-3.5-Turbo	34.9	26.0	2.57	41.2	40.1	1.53	66.3	46.2	1.98	47.5	37.4	2.03
	Qwen2.5-14B	36.2	30.1	2.56	49.6	42.0	1.56	60.5	44.6	1.68	48.8	38.9	1.93
	GPT-4o-mini	35.3	26.9	2.61	39.2	41.3	1.50	63.1	46.5	1.91	45.9	38.2	2.01
	GPT-OSS-20B	39.1	24.1	2.39	63.3	46.0	1.47	43.9	48.0	1.77	48.8	39.4	1.88
	DeepSeek-V3	42.6	34.1	**2.77**	70.0	52.7	1.51	81.0	48.6	1.90	64.5	45.1	2.06
	GPT-5-Chat	46.9	**38.7**	2.73	77.4	**54.0**	1.63	89.4	52.9	**2.03**	71.2	**48.5**	**2.13**
	Claude-Sonnet-4	**53.9**	37.5	2.67	**84.4**	52.9	**1.67**	**90.2**	**53.4**	2.02	**76.2**	47.9	2.12

demonstrates that open-ended generation is much more challenging for modern LLMs. Overall, the results point to remaining gaps in persona tone realization and in recalling and using persona-specific details during generation.

4.2 Capability-wise Analysis

Figure 3 details capability-level performance, aggregating discriminative accuracy with generative ranking and scoring. From Figure 3, we have the following observations:

First, the performance of different LLMs is broadly aligned across capabilities: LLMs that lead on one capability tend to lead elsewhere. Second, LLMs score highest on *Lexical Fidelity* and *Opinion Consistency*, and lowest on *Persona Tone* and *Memory Recall*. This demonstrates that current LLMs remain inadequate in tasks that require memory and tone simulation. Third, individual models show distinct comparative advantages; for example, DeepSeek-V3 approaches GPT-5 on *Lexical Fidelity* despite trailing on others. **In summary, current LLMs share some common limitations in capabilities that are required for persona simulation, whereas different models exhibit distinct strengths across various capabilities.**

4.3 Generative Evaluation

4.3.1 LLM-as-a-Judge: Scoring and Ranking

We evaluate generative outputs via Judge ranking (Acc.(Gen)) and 1–5 scoring (Score(Gen)) (Table 2; prompts in Appendix A).

Key results are as follows: GPT-5-Chat attains the strongest aggregate generative performance (Acc.(Gen) 48.5%, Score(Gen) 2.13), closely followed by Claude-Sonnet-4 (47.9%, 2.12). DeepSeek-V3 is competitive and achieves the best Score(Gen) on the Social Persona dimension (2.77), despite trailing the leaders on other dimensions. **Generative performance is systematically lower than discriminative accuracy, underscoring the difficulty of free-form generation.**

Table 3: Agreement of GPT-5 as a Judge against human annotations and inter-annotator reliability.

Task	GPT-5 vs. Human	Inter-annotator Reliability
Ranking (four choice)	0.646^{κ}	0.673^{κ}
Scoring (one to five)	0.591^{ρ}	0.605^{ρ}

Note: κ is Fleiss's kappa for categorical labels and ρ is Spearman correlation for ordinal scores. Sample size is 50.

4.3.2 Reliability and Bias Analysis of the Judge

We validate the LLM-as-a-Judge methodology with a human study. Three expert annotators evaluated a stratified sample of 50 items per judging mode (ranking and scoring), following our instruction set (Appendix E). Annotators worked independently and were blinded to each other's labels.

Agreement between GPT-5-as-a-Judge and humans is reported in Table 3 and is comparable to human inter-annotator reliability: for ranking, Fleiss's κ is 0.646 (GPT-5 vs. human) versus 0.673 (human–human); for scoring, Spearman's ρ is 0.591 (GPT-5 vs. human) versus 0.605 (human–human). **These results indicate that the proposed benchmark achieves a high human inter-annotator agreement, and the LLM-as-a-Judge provides reliable annotations aligned with humans.**

Judge Bias Analysis. To address potential self-preference bias, we conduct a cross-evaluation using Claude-Sonnet-4 as an alternative judge (Table 4). Results are consistent across evaluators:

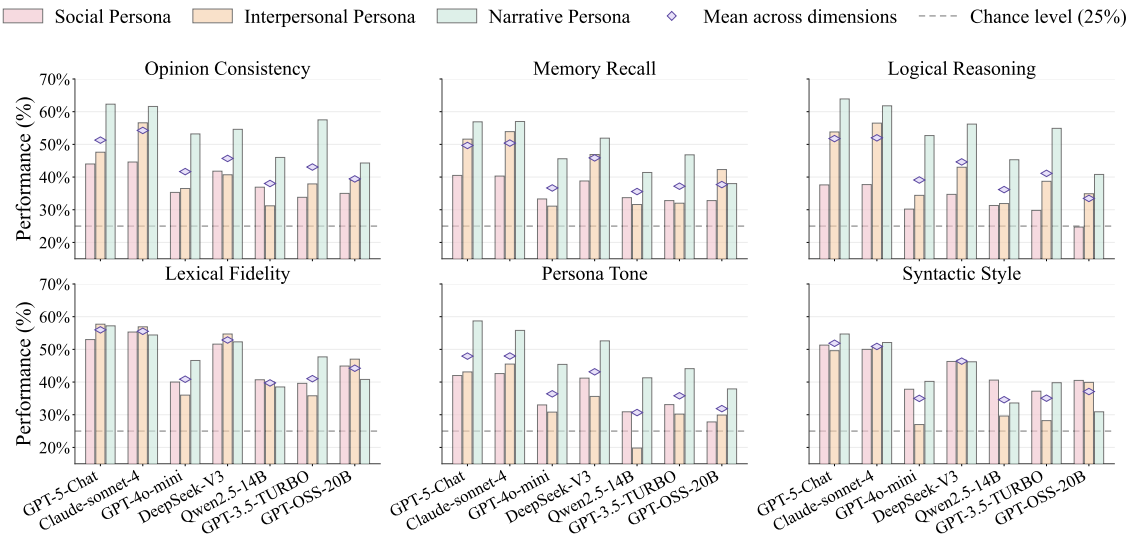


Figure 3: **Performance across six capabilities.** Each panel shows one capability. For each LLM, bars give scores on the three dimensions—Social, Interpersonal, and Narrative. Purple diamonds indicate the mean across the three dimensions for that model. The y-axis is the average over the three evaluation protocols: discriminative, generative ranking, and generative scoring. The gray dashed line denotes the chance level (25%).

Table 4: **Cross-Judge Evaluation. Acc.** (Gen)(%) and **Score** (Gen)(1–5) on three dimensions (Social, Interpersonal, Narrative) with different LLM-as-a-Judges.

Persona	Evaluated LLMs	Judge: GPT-5		Judge: Claude	
		Acc. (Gen)(%)	Score (Gen)	Acc. (Gen)(%)	Score (Gen)
Social Persona	GPT-5-Chat	38.7	2.73	37.4	2.34
	Claude-Sonnet-4	37.5	2.67	36.1	2.26
Interpersonal Persona	GPT-5-Chat	54.0	1.63	52.5	1.44
	Claude-Sonnet-4	52.9	1.67	51.9	1.50
Narrative Persona	GPT-5-Chat	52.9	2.03	49.2	1.74
	Claude-Sonnet-4	53.4	2.02	49.9	1.77

both judges favor GPT-5-Chat in *Social* and *Interpersonal* dimensions, while crucially preferring Claude-Sonnet-4 in the *Narrative* dimension (e.g., 53.4% vs 52.9% by GPT-5 judge). Although Claude is systematically stricter (yielding lower scores), the relative hierarchy remains invariant. **This demonstrates that bias associated with self-judgment does not affect the relative performance comparison between GPT-5-Chat and Claude-Sonnet-4.**

4.3.3 Text Similarity Metrics

To provide an objective measurement, the generative-based evaluation is also measured by standard text similarity metrics—BLEU-1, METEOR, and BERT-Score—and reports results in Appendix G. These metrics primarily reflect lexical overlap and local paraphrase rather than opinion alignment, reasoning trajectories, or persona tone. Averaged over the three dimensions, Claude-Sonnet-4 attains the best BERT-Score (76.90) and METEOR (18.24), while GPT-5-Chat achieves the best BLEU-1 (19.13). **The resulting ranking of different LLMs is broadly consistent with our evaluation based on LLM-as-a-Judge, offering cross-validation.**

4.4 Human vs. Model Performance

We benchmark human performance on the Social Persona discriminative task. Three expert annotators labeled a stratified set of 50 items following our guidelines (Appendix E). Given the task's reliance on long contexts and implicit cues, we note that human performance is a reference rather than a strict upper bound.

Table 5 compares models to human baselines. GPT-5-Chat reaches 0.60 accuracy, which is lower than the human mean of 0.64 and the majority-vote aggregate of 0.66. Agreement with humans is high but short of human–human reliability: Fleiss's κ is 0.634 for model vs. human and 0.690 for inter-annotator agreement.

These results indicate that state-of-the-art LLMs still lag behind human performance. Given that humans are still imperfect players in persona simulation, we believe that human performance is only a practical reference and can be further approached with the advance of LLM capbilities.

Summary of Findings. Across three persona di-

Table 5: We compare GPT-5 accuracy against human's average and majority Vote on the discriminative-based evaluation($N = 50$).

Accuracy			Agreement (κ)	
GPT-5	**Human (Avg)**	**Human (Vote)**	**Model-Hum.**	**Inter-Ann.**
0.60	0.64	0.66	0.634	0.690

Note: Human (Avg) is the annotator average. (Vote) is the aggregated majority vote. Agreement uses Fleiss's κ.

mensions and two task formulations, strong models (GPT-5-Chat, Claude-Sonnet-4) lead consistently, yet free-form persona simulation remains notably harder than multiple-choice selection. Capability analysis pinpoints style control and memory recall as primary bottlenecks, while lexical fidelity and opinion consistency are comparatively robust. The benchmark can effectively test LLMs in the generative-based evaluation with LLM-as-a-Judge, with evidence including its alignment with human judgments, and complementary validation based on text-similarity metrics. Across settings, results exhibit substantial variance between models without evident ceiling effects. There remains clear headroom in three areas: maintaining persona coherence over extended contexts and across sessions, producing a persona-consistent tone, and recalling and using persona-specific facts during generation.

5 Related Work

5.1 Personalized Agents and Digital Twins

The construction of digital twins, virtual replicas of specific individuals, is an emerging challenge in AI (Shanahan et al., 2023; Park et al., 2023). Originating in engineering as counterparts to physical systems (Grieves and Vickers, 2017), the concept now extends to AI agents that capture a person's communication style, preferences, and personality. Recent efforts have operationalized this vision across diverse domains. Examples include reviving anime characters (Li et al., 2023), simulating agent societies from novels (Ran et al., 2025), and evaluating impersonation of writing styles and memories (Shi et al., 2025). Applications have been explored in healthcare (Barricelli et al., 2020), marketing (Hornik and Rachamim, 2025), and through industry systems like SecondMe (Shang et al., 2024) for lifelong personal modeling. While these human-centered digital twins promise highly personalized chatbots (Ma et al., 2023; Li et al., 2025a) and ubiquitous computing applications (Fast et al., 2016), prior research has often focused narrowly on style imitation, overlooking the broader competencies required for authentic persona simulation.

5.2 Datasets, Benchmarks, and Evaluation for Persona Simulation

Progress in this area depends on high-quality datasets and benchmarks. Recent resources have begun to fill this gap, offering diverse evaluation protocols. Benchmarks have been developed from large-scale surveys of human traits (Toubia et al., 2025; Chen et al., 2025), persona-based behavior chains (Li et al., 2025b), psychology-guided agent evaluations (Xie et al., 2025), persona-driven decision-making tasks (Afzoon et al., 2024; Xu et al., 2024), and multi-party dialogue role identification (Zhou et al., 2025). More recent work explores challenging settings like counterfactual simulation (Kumar et al., 2025) and dynamic user profiling (Jiang et al., 2025).

Despite this growing landscape, evaluations remain fragmented and often rely on synthetic data, limiting their ecological validity. This highlights the need for a unified framework to advance digital twin research rigorously. Our TwinVoice benchmark addresses these limitations by leveraging real-world social media, conversational, and fictional data to provide authentic and systematic evaluation across multiple persona dimensions.

6 Conclusion

This paper addresses the evaluation of LLM-based persona simulation by introducing **TwinVoice**. Built on real-world and fictional data from three dimensions, TwinVoice aims at testing LLMs' ability in persona simulation by decomposing it into six capabilities of mindset coherence and linguistic expression. Our extensive evaluation of state-of-the-art models reveals a crucial gap: while leading models like GPT-5-Chat and Claude-Sonnet-4 show improved accuracy over their predecessors, their performance still falls significantly short of human-level capabilities. We also find that LLMs are adept at mimicking surface-level linguistic styles, they consistently fail to maintain long-term consistency, particularly in memory recall and opinion stability. By establishing the first fine-grained baselines in this domain, TwinVoice not only exposes the key limitations of current models but also provides a clear roadmap towards personalized AI and digital twins built with LLMs.

Limitations

TwinVoice currently spans three dimensions and five languages: Social (Chinese), Interpersonal (English, Spanish, Portuguese, Russian), and Narrative (English). Despite this breadth, language balance within each dimension remains imperfect, and phenomena such as code-switching and dialectal variation are underrepresented. Future releases will expand per-dimension language coverage and diversify domains where consented and de-identified data are available.

Ethics Statement

We follow standard ethical guidelines for dataset usage, evaluation, and model deployment. All datasets used in this paper are publicly available under their original licenses, and we removed personally identifiable information (PII) where applicable. No human subjects experiments were conducted beyond voluntary annotation; annotators (if any) received fair compensation and provided informed consent. We prohibit misuse of our benchmark and models for profiling or harmful decision making about individuals. Third-party models/APIs used in our experiments comply with their terms of service. Upon acceptance, we will release our code, prompts, and evaluation scripts with a research license and a model card detailing limitations and appropriate use.

References

Saleh Afzoon, Usman Naseem, Amin Beheshti, and Zahra Jamali. 2024. Persobench: Benchmarking personalized response generation in large language models. *arXiv preprint arXiv:2410.03198*.

Anthropic. 2025. Introducing claude 4. URL: https://www.anthropic.com/news/claude-4. Official announcement of the Claude 4 model family, including Opus 4 and Sonnet 4.

Barbara Rita Barricelli, Elena Casiraghi, Jessica Gliozzo, Alessandro Petrini, and Stefano Valtolina. 2020. Human digital twin for fitness management. *IEEE Access*, 8:26637–26664.

Jason Baumgartner, Savvas Zannettou, Megan Squire, and Jeremy Blackburn. 2020. The pushshift telegram dataset. *Preprint*, arXiv:2001.08438.

Douglas Biber. 1991. *Variation across speech and writing*. Cambridge university press.

Douglas Biber. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press.

Penelope Brown. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, and 1 others. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. 2025. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint (Anthtropic technical report)*.

Herbert H Clark and Susan E Brennan. 1991. Grounding in communication.

Paul T Costa and Robert R McCrae. 1992. Normal personality assessment in clinical practice: The neo personality inventory. *Psychological assessment*, 4(1):5.

Ethan Fast, William McGrath, Pranav Rajpurkar, and Michael S Bernstein. 2016. Augur: Mining human behaviors from fiction to power interactive systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 237–247.

Michael Grieves and John Vickers. 2017. Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Jacob Hornik and Matti Rachamim. 2025. Ai-enabled consumer digital twins as a platform for research aimed at enhancing customer experience. *Management Review Quarterly*.

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

Cameron Jones and Ben Bergen. 2024. Does gpt-4 pass the turing test? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5183–5210.

Cameron R Jones and Benjamin K Bergen. 2025. Large language models pass the turing test. *arXiv preprint arXiv:2503.23674*.

Cameron Robert Jones, Ishika Rathi, Sydney Taylor, and Benjamin K Bergen. 2025. People cannot distinguish gpt-4 from a human in a turing test. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 1615–1639.

Daniel Kahneman. 2011. *Thinking, fast and slow*. macmillan.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology*, 60(1):9–26.

Sai Adith Senthil Kumar, Hao Yan, Saipavan Perepa, Murong Yue, and Ziyu Yao. 2025. Can llms simulate personas with reversed performance? a benchmark for counterfactual instruction following. *Preprint*, arXiv:2504.06460.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song Yan, Haosheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. 2023. Chatharuhi: Reviving anime character in reality via large language model. *Preprint*, arXiv:2308.09597.

Hao Li, Chenghao Yang, An Zhang, Yang Deng, Xiang Wang, and Tat-Seng Chua. 2025a. Hello again! llm-powered personalized agent for long-term dialogue. *Preprint*, arXiv:2406.05925.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*.

Rui Li, Heming Xia, Xinfeng Yuan, Qingxiu Dong, Lei Sha, Wenjie Li, and Zhifang Sui. 2025b. How far are llms from being our digital twins? a benchmark for persona-based behavior chain simulation. *Preprint*, arXiv:2502.14642.

Xiao Ma, Swaroop Mishra, Ariel Liu, Sophie Su, Jilin Chen, Chinmay Kulkarni, Heng-Tze Cheng, Quoc Le, and Ed Chi. 2023. Beyond chatbots: Explorellm for structured thoughts and personalized model responses. *Preprint*, arXiv:2312.00763.

Matthias R Mehl, Samuel D Gosling, and James W Pennebaker. 2006. Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life. *Journal of personality and social psychology*, 90(5):862.

Yair Neuman. 2016. *Computational personality analysis: Introduction, practical applications and novel directions*. Springer.

OpenAI. 2025. Introducing gpt-5. URL: https://openai.com/index/introducing-gpt-5/. Official announcement of the GPT-5 model, a unified system with built-in reasoning capabilities.

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.

James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577.

Project Gutenberg. 1971–. Project gutenberg. https://www.gutenberg.org.

Hongjin Qian, Xiaohe Li, Hanxun Zhong, Yu Guo, Yueyuan Ma, Yutao Zhu, Zhanliang Liu, Zhicheng Dou, and Ji-Rong Wen. 2021. Pchatbot: a large-scale dataset for personalized chatbot. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2470–2477.

Yiting Ran, Xintao Wang, Tian Qiu, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2025. Bookworld: From novels to interactive agent societies for creative story generation. *Preprint*, arXiv:2504.14538.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

Jingbo Shang, Zai Zheng, Jiale Wei, Xiang Ying, Felix Tao, and Mindverse Team. 2024. Ai-native memory: A pathway from llms towards agi. *Preprint*, arXiv:2406.18312.

Tianhao Shen, Sun Li, Quan Tu, and Deyi Xiong. 2023. Roleeval: A bilingual role evaluation benchmark for large language models. *arXiv preprint arXiv:2312.16132*.

Quan Shi, Carlos E. Jimenez, Stephen Dong, Brian Seo, Caden Yao, Adam Kelch, and Karthik Narasimhan. 2025. Impersona: Evaluating individual level lm impersonation. *Preprint*, arXiv:2504.04332.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.

Olivier Toubia, George Z. Gui, Tianyi Peng, Daniel J. Merlau, Ang Li, and Haozhe Chen. 2025. Twin-2k-500: A dataset for building digital twins of over 2,000 people based on their answers to over 500 questions. *Preprint*, arXiv:2505.17479.

Quan Tu, Shilong Fan, Zihang Tian, and Rui Yan. 2024. Charactereval: A chinese benchmark for role-playing conversational agent evaluation. *arXiv preprint arXiv:2401.01275*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, and

1 others. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Qiujie Xie, Qiming Feng, Tianqi Zhang, Qingqiu Li, Linyi Yang, Yuejie Zhang, Rui Feng, Liang He, Shang Gao, and Yue Zhang. 2025. Human simulacra: Benchmarking the personification of large language models. In *The Thirteenth International Conference on Learning Representations*.

Rui Xu, Xintao Wang, Jiangjie Chen, Siyu Yuan, Xinfeng Yuan, Jiaqing Liang, Zulong Chen, Xiaoqing Dong, and Yanghua Xiao. 2024. Character is destiny: Can role-playing language agents make persona-driven decisions? *Preprint*, arXiv:2404.12138.

Ziyi Ye, Xiangsheng Li, Qiuchi Li, Qingyao Ai, Yujia Zhou, Wei Shen, Dong Yan, and Yiqun Liu. 2025. Learning llm-as-a-judge for preference alignment. In *The Thirteenth International Conference on Learning Representations*.

John Zaller. 1992. *The nature and origins of mass opinion*. Cambridge university press.

Lingfeng Zhou, Jialing Zhang, Jin Gao, Mohan Jiang, and Dequan Wang. 2025. Personaeval: Are llm evaluators human enough to judge role-play? *Preprint*, arXiv:2508.10014.

A Evaluation Protocols and Full Prompts

This appendix details our evaluation protocols and the full instruction templates used across multiple data forms, including public social interactions, interpersonal messaging, and narrative dialogue. We adopt a unified instruction design and provide template variants for different data shapes when needed. Unless otherwise noted, the LLM-as-a-Judge component is instantiated with GPT-5.

A.1 Scope and Alignment with Competencies

Our evaluation comprises (1) discriminative multiple-choice selection and (2) generative evaluation, including persona imitation (free-form generation) and LLM-as-a-Judge assessment via ranking and scoring. The judge scoring rubric is organized along three pillars—Opinion Consistency, Logical & Factual Fidelity, and Stylistic Similarity—which align with the six fundamental capabilities defined in the main text. We offer equivalent template variants per evaluation mode to fit different data shapes; metrics and scoring criteria remain identical across variants.

A.2 Unifying Instructions and Placeholders

We use a single instruction family per evaluation mode. Differences are limited to how inputs are presented. We standardize placeholders as follows:

- `{history}`: persona-establishing prior content by the same user or character.
- `{context}`: the situation/post/message/scene the user or character is responding to (replacing earlier `{anchor}` or `{anchor_post}`).
- `{ground_truth_reply}` or `{groundtruth_response}`: the human-written reply.
- `{lmut_reply}` or `{generated_content}`: the model-generated reply to be evaluated.

A.3 Discriminative Evaluation (Multiple-Choice Selection)

We employ three distinct instruction templates for the discriminative task. The **Canonical Prompt** for general social media interactions is presented in Table 7. For the interpersonal dimension, we use the **Messaging Variant** shown in Table 8. Finally, for the narrative dimension, we utilize a specialized **Distractor Generation Prompt** (Table 9).

A.4 Generative Evaluation: Persona Imitation

The **Canonical Prompt** for Social Persona is shown in Table 10. For Interpersonal Persona, we use a **JSON-structured variant** (Table 11). For the Narrative Persona, the **Digital Twin Line Generation Prompt** (Table 12) enforces constraints.

A.5 LLM-as-a-Judge: Scoring-Based Evaluation

The unified **Scoring Prompt** is shown in Table 15.

A.6 LLM-as-a-Judge: Ranking-Based Evaluation

The **Canonical Ranking Prompt** (Table 13) requires the model to output a choice and a brief reasoning. For Narrative, we use a **Letter-Only Prompt** (Table 14).

B Capability Annotation Prompts and Labeling Protocol

We annotate each example to identify which capability a model must primarily exercise to replicate a user's reply, while also recording the presence of all six capabilities. To ensure consistency and structured outputs, we employ an expert LLM (GPT-5, temperature 0) using the standardized instruction set detailed in Table 16.

C Capability Distinguishing Case Studies

This section presents case studies that illustrate how our six capabilities appear in practice. The examples are drawn from our public social persona corpus. For readability we show faithful translations and only the key slices. If any discrepancy arises, the original Chinese dataset is authoritative. Explanatory remarks appear outside the boxes. Inside each box, ····· marks omitted portions of longer cases.

C.1 Opinion Consistency

The user maintains a specific stance across contexts, namely choosing shows based on a favorite actor and praising acting skill. The new reply preserves this granular stance rather than defaulting to generic positivity.

Case 1: Opinion Consistency (user 527222)

Context. ····· "Tonight is the finale. Xiang Qian returns to the seaside house where he once lived in hard times, surely full of feelings.

Seeing Alisa in this moment is so beautiful, hope they both have a good life." ·····
Key History. ····· "I watched this show for Huang Zitao, I think his acting is great." ·····
Ground Truth Reply. ····· "I watched it for Liu Tao, her acting is really getting better and better." ·····

Why this shows Opinion Consistency: The historical pattern is watch for a specific actor and praise that acting. The ground truth reply mirrors the same stance toward another named actor, preserving topic granularity and evaluative angle.

C.2 Memory Recall

The reply uses a nickname that is not introduced in the immediate context, presupposing shared knowledge from prior interactions. Understanding the line fully requires recalling who that nickname refers to.

Case 2: Memory Recall (user 205470)

Context. ····· "Met a teacher who is a high level LEGO player, buys LEGO by the sack." .....
Key History. ····· "When Dan jie builds LEGO she looks like a serious kid, always supporting Dan jie." .....
Ground Truth Reply. ····· "When she plays LEGO her eyes light up, still that adorable Wang Sansui." .....

Why this shows Memory Recall: The affectionate nickname Wang Sansui is not grounded in the current context and relies on earlier persona knowledge to resolve the reference.

C.3 Logical Reasoning

The user's pattern is Observation then Deduction. In history, a physical observation supports an inference. The reply replicates this approach by citing scene features to argue against an assumption.

Case 3: Logical Reasoning (user 369593)

Context. ····· "Do an ice drifting video. If it is not minus twenty or thirty degrees, do not show off." .....
Key History. ····· "There is no snow on the roof opposite, which shows the heat inside that house is considerable." .....
Ground Truth Reply. ····· "This river channel is quite narrow and there is a road next to it, so it probably did not fall in from drifting on the ice." .....

Why this shows Logical Reasoning: The reply marshals concrete observations (narrow channel, road nearby) to support a causal judgment, matching the user's habit of evidence based inference.

C.4 Lexical Fidelity

A personal catchphrase recurs across contexts. The reply deploys the same idiosyncratic exclamation seen in history, signaling a learned lexical signature.

Case 4: Lexical Fidelity (user 45899)

Context. ····· "Emirates Bling777 plane is encrusted with Swarovski crystals, the joy of the rich is beyond imagination." .....
Key History. ····· "OMG, for this kind of dog, give me a dozen and it is not too many." .....
Ground Truth Reply. ····· "OMG, this, this, it is full of diamonds?! Maybe one will drop off for me." .....

Why this shows Lexical Fidelity: The same colloquial exclamation equivalent to OMG appears in both history and reply, demonstrating consistent, user-specific lexical choice.

C.5 Persona Tone

The user favors playful hyperbole and adoring expressions that are nonliteral. The reply echoes that tone with a different bodily metaphor, preserving the same stylistic stance.

Case 5: Persona Tone (user 270844)

Context. ····· "Group stage, Hai Lu's acting is on point, those long legs are eye catching. Did not expect such solid dance foundation, the high kicks are captivating." .....
Key History. ····· "Listening made my ears pregnant, you all should listen, it is super good. Hope my male god keeps getting better. Could you be my boyfriend, so shy." .....
Ground Truth Reply. ····· "Hai Lu, your long legs had me staring at them the whole time, haha, my nose is about to bleed." .....

Why this shows Persona Tone: Both history and

reply use exuberant, nonliteral bodily metaphors (ears pregnant, nosebleed) as playful, adoring exaggerations that define the user's persona.

C.6 Syntactic Style

Beyond words and tone, the user's structure features stacked, breathless exclamations with intensifiers. The reply reproduces that sentence shape.

Case 6: Syntactic Style (user 108194)

Context. ····· "Sci fi fans, gather up. The film The Wandering Earth is set for Lunar New Year, a concept poster has been released." ·····
Key History. ····· "Wow wow wow, I am truly so excited inside, really looking forward to it, hahaha." ·····
Ground Truth Reply. ····· "Wow wow wow, look closely, this poster design really has such a vibe, you could call it outstanding. This kind of movie theme is especially attractive, must support." ·····

Why this shows Syntactic Style: The reply stacks short, exclamatory clauses with intensifiers and colloquial particles, recreating the user's distinctive, breathless rhythm observed in history.

D Radar Charts across Three Dimensions

We present capability-wise radar charts for the three persona dimensions: Social Persona, Interpersonal Persona, and Narrative Persona. For each dimension, we report four evaluation configurations: (i) Combined Average (aggregated across protocols), (ii) Discriminative (multiple-choice selection), (iii) Generative Ranking (LLM-as-a-Judge; Acc.(Gen)), and (iv) Generative Scoring (LLM-as-a-Judge; Score(Gen), 1–5). Each radar covers six capabilities: Opinion Consistency, Memory Recall, Logical Reasoning, Lexical Fidelity, Persona Tone, and Syntactic Style. The results are presented in Figure 4,Figure 5,Figure 6,Figure 7,Figure 8, and Figure 9.

E Human Annotation Guidelines

E.1 Task Background and Objectives

This study aims to evaluate the performance of Large Language Models (LLMs) as judges in digital twin tasks. To validate the reliability of model judgments, we need human annotators to independently annotate selected data to establish a trustworthy benchmark.

The annotation task consists of three subtasks corresponding to different evaluation modes: discriminative tasks, generative ranking tasks, and generative scoring tasks. Each annotator will annotate the same 100 data samples to ensure consistency and comparability in evaluation.

Important Note: All provided content (anchor posts, reply history, choices) is in Chinese. You should analyze and understand the content within the Chinese language context, but your reasoning and annotations should be provided in English when specified.

E.2 Discriminative Task Annotation

E.2.1 Task Description

In the discriminative task, you need to act as a specific social media user, becoming their digital twin. Based on the given conversation history and anchor post, select the most appropriate reply from four candidates that best matches the user's personal style and language habits.

E.2.2 LLM Prompt (Use the Same Evaluation Standard)

The LLM uses the following prompt for this task. Please follow the same reasoning approach:

> *Your task is to act as a specific social media user, becoming their digital twin. Note: All provided text (history, post, choices) is in Chinese. You must analyze the user's style directly within the Chinese language context.*
>
> Based on the user's reply history, think and respond with their mindset, tone, and style.
>
> Your reply history: (Note: "AnchorPost" is another user's post, and "UserReply" is your own reply.)
>
> Now, you see a new post: [anchor post]
>
> Below are 4 candidate replies. Which one is most likely something you would say?
>
> Please respond by explaining your choice from the user's perspective using "I".

E.2.3 Evaluation Criteria

- **Style Consistency**: Does the reply maintain consistency with the user's language style demonstrated in conversation history?

- **Tone Matching**: Does the reply's tone (formal/informal, humorous/serious, etc.) match the user's characteristics?
- **Vocabulary Usage**: Are the vocabulary choices and expressions consistent with the user's habits?
- **Logical Coherence**: Is the reply content logically related to the anchor post and historical context?

E.2.4 Additional Human Guidance

- Carefully read through the entire conversation history to understand the user's communication patterns
- Pay attention to recurring phrases, greeting patterns, and emotional expressions
- Consider the user's typical response length and level of detail
- Think from the user's perspective: "If I were this user, which response would I most likely choose?"

E.2.5 Annotation Method

Please fill in the option number (0, 1, 2, or 3) that you consider most appropriate in the `human_choice` field, corresponding to the index position in the choices array.

E.3 Generative Ranking Task Annotation

E.3.1 Task Description

In the generative ranking task, you need to identify which candidate reply is most similar to a reference reply in terms of style, tone, vocabulary, sentiment, and topic.

E.3.2 LLM Prompt (Use the Same Evaluation Standard)

The LLM uses the following prompt for this task:

> You are an expert evaluator of writing style. Your task is to compare several candidate replies against a known "Reference Reply" written by a specific user.
>
> Your goal is to identify which candidate is the most similar to the reference in terms of style, tone, vocabulary, sentiment, and topic.
>
> Now, determine which single candidate is the closest match to the Reference Reply. The reasoning should be concise, limited to 2-3 sentences, focusing on the stylistic similarities.

E.3.3 Evaluation Criteria

- **Style Similarity**: Lexical choices, sentence structure, formality level
- **Tone Matching**: Emotional tone, attitude, and mood
- **Vocabulary Consistency**: Use of similar words, phrases, or expressions
- **Sentiment Alignment**: Overall emotional orientation and sentiment
- **Topic Relevance**: Relevance and approach to the main topic

E.3.4 Additional Human Guidance

- Focus on stylistic elements rather than factual content
- Look for subtle language patterns and preferences
- Consider both what is said and how it is said
- Compare the "voice" and "personality" reflected in each candidate

E.3.5 Annotation Method

Please fill in the letter (A, B, C, or D) of the option you consider best matching in the `human_choice` field.

E.4 Generative Scoring Task Annotation

E.4.1 Task Description

In the generative scoring task, you need to assess how well a generated reply replicates a ground truth reply, providing a score from 1-5 based on comprehensive evaluation criteria.

E.4.2 LLM Prompt (Use the Same Evaluation Standard)

The LLM uses the following detailed evaluation framework:

> *You are a meticulous and objective evaluator for a digital twin benchmark. Your task is to assess how well a 'Generated Reply' replicates a 'Ground Truth Reply' for a given social media post.*
>
> The evaluation rests on three key pillars:

1. **Opinion Consistency**: Does the Generated Reply express the exact same core opinion, stance, and sentiment as the Ground Truth?
2. **Logical & Factual Fidelity**: Is the Generated Reply based on the same reasoning and facts as the Ground Truth?
3. **Stylistic Similarity**: How closely does the Generated Reply match the Ground Truth in terms of lexical, tone, and syntactic elements?

E.4.3 Scoring Rubric (1-5 Scale)

- **5 - Perfect Replication**: Perfect match across all three pillars. Feels like a natural, alternative expression from the same user.
- **4 - High Fidelity**: Opinion and Logic/Factual pillars are perfectly matched. Only minor, subtle differences in Style.
- **3 - Core Alignment, Detail Loss**: Core opinion is consistent, but noticeable loss of detail in Logic or Style pillars.
- **2 - Partial Relevance, Major Deviation**: Major failure in at least one of the three pillars.
- **1 - Irrelevant or Contradictory**: Almost nothing in common with the Ground Truth or expresses contradictory opinion.

E.4.4 Additional Human Guidance

- First identify the core opinion/stance in the ground truth reply
- Check if the generated reply maintains the same logical flow and reasoning
- Evaluate stylistic elements: word choice, sentence length, formality, emotional tone
- Consider the reply as a whole - would it serve as an acceptable substitute?
- Be objective and consistent across all annotations

E.4.5 Annotation Method

Please fill in your score (1, 2, 3, 4, or 5) in the `human_score` field.

E.5 General Guidelines and Notes

E.5.1 Quality Assurance

- Read all conversation history carefully to understand the user's communication patterns
- Maintain objectivity and consistency throughout the annotation process
- Avoid letting personal preferences influence your judgment
- Each data sample should be annotated independently
- When facing difficult decisions, choose the relatively best option
- Double-check for missing annotations or format errors after completion

E.5.2 Language Considerations

- All content is in Chinese - analyze within the Chinese language context
- Pay attention to Chinese-specific expressions, internet slang, and cultural references
- Consider Chinese punctuation and writing conventions
- Understand the social media context and communication norms

F Use of Large Language Models

F.1 Scope of Use

LLMs assisted with (i) prompt drafting and refinement, (ii) minor code refactoring suggestions, (iii) generating synthetic evaluation items (e.g., distractor options and candidate responses), and (iv) light copy-editing of non-technical prose. LLMs did *not* originate novel claims, conduct final analyses, or decide conclusions; all substantive results are author-verified.

F.2 Models and Access

We used the following LLMs via API/local inference: **GPT-5-Chat** (OpenAI), **Claude-Sonnet-4** (Anthropic), **DeepSeek-V3** (DeepSeek), **GPT-4o-mini** (OpenAI), **GPT-3.5-Turbo** (OpenAI), **GPT-OSS-20B** (Open-source community), **Qwen2.5-14B** (Alibaba / Qwen Team). **Access window:** 06/2025–09/2025.

F.3 Human Oversight

All LLM outputs were screened by the authors; items entering quantitative evaluation were validated via deterministic scripts or double review.

F.4 Reproducibility

We include the full evaluation prompts and protocols, the 1–5 scoring rubric, the textual recipes for constructing multiple-choice questions, the data filtering thresholds per dimension, dataset sizes/statistics, and the evaluation equations and metrics. These disclosures are sufficient to re-implement our evaluation.

F.5 Data Privacy and Safety

Only public data were processed; no PII or sensitive user data were sent to third-party services. We complied with provider Terms of Service and applied toxicity/safety filters where applicable.

F.6 Limitations

LLM outputs may reflect training-data biases or hallucinations. We mitigated these via rule-based validators and manual review; residual errors may remain.

G Text Similarity Metrics

For completeness, we additionally evaluate performance using standard text similarity metrics: BLEU-1, METEOR, and BERT-Score. As detailed in Table 6, Claude-Sonnet-4 achieves the highest average BERT-Score (76.90) and METEOR (18.24), while GPT-5-Chat leads in BLEU-1 (19.13). Although these metrics prioritize lexical overlap and local paraphrase over high-level reasoning, the observed performance hierarchy aligns with our judge-based evaluation, providing complementary evidence to the main findings.

Table 6: Objective metrics for Digital Twin models. We evaluate the generative outputs against the ground truth using three distinct metrics. **BLEU-1** ↑ measures unigram precision. **METEOR** ↑ considers precision, recall, and synonymy. **BERT-Score** ↑ measures semantic similarity using contextual embeddings. Higher values are better for all metrics. **Bold** numbers denote the best result and underlined numbers denote the second best in each column.

	Model / Tasks	Social			Interpersonal			Narrative			Average		
		BLEU-1 ↑	METEOR ↑	BERT-Score ↑	BLEU-1 ↑	METEOR ↑	BERT-Score ↑	BLEU-1 ↑	METEOR ↑	BERT-Score ↑	BLEU-1 ↑	METEOR ↑	BERT-Score ↑
LLM	GPT-3.5-Turbo	16.03	15.50	62.96	24.76	22.52	81.54	12.06	12.86	84.10	17.62	16.96	76.20
	Qwen2.5-14B	17.68	15.38	63.25	26.09	23.76	81.57	11.67	11.92	83.99	18.48	17.02	76.27
	GPT-4o-mini	15.94	15.19	62.89	23.48	21.38	81.26	**12.50**	**13.34**	84.13	17.31	16.64	76.09
	GPT-OSS-20B	14.55	12.87	61.90	20.67	19.20	81.17	10.81	10.59	84.36	15.34	14.22	75.81
	DeepSeek-V3	16.85	15.49	63.25	26.86	25.21	82.65	11.11	11.58	84.12	18.27	17.43	76.67
	GPT-5-Chat	18.67	14.09	63.26	**27.18**	**25.30**	**82.67**	11.54	11.59	84.27	**19.13**	16.99	76.73
	Claude-Sonnet-4	**18.68**	**18.14**	**64.19**	25.22	23.45	82.14	12.38	13.12	**84.37**	18.76	**18.24**	**76.90**

Table 7: The canonical instruction prompt for the discriminative multiple-choice selection task (General).

Discriminative Selection Prompt (General)

Your task is to act as a specific social media user, becoming their digital twin. Note: All provided text (history, context, choices) is in the original language of the data. You must analyze the user's style directly within that language.
Based on the user's reply history, think and respond with their mindset, tone, and style.
Your reply history: (Note: "Context" is another user's post/message, and "UserReply" is your own reply.) history
Now, you see a new context message: "context"
Below are 4 candidate replies. Which one is most likely something you would say?
A. a B. b C. c D. d
Please respond in the following JSON format. In the "reasoning" field, use the first-person perspective ("I") to explain your choice.

```
{
  "predicted_comment": "A",
  "reasoning": "Explain, from my perspective as the user, why I would choose
     this option."
}
```

Table 8: The instruction prompt for discriminative selection adapted for the interpersonal messaging dimension.

Discriminative Selection Prompt (Messaging Variant)

You are given a user's reply history and 4 candidate replies to a context message. Only one of the replies was actually written by this user. The other three were written by different users replying to the same context message. Your task is to choose the most likely reply written by the same user, based on writing style, tone, and expression habits. Focus on how the user typically speaks, their phrasing, and how they respond emotionally or humorously.
User's Historical Conversations: history
Current Context Message: "context"
Candidate Replies: A. a B. b C. c D. d
Please respond in the following JSON format:

```
{
  "predicted_comment": "A",
  "reasoning": "Explain why this option best matches the user's style."
}
```

Table 9: The prompt used for generating distractor options in the Narrative dimension.

Distractor Writer Prompt (Narrative Variant)

You are a precise persona-grounded writer. Given one TARGET speaker (whose original utterance is the correct answer) and THREE OTHER characters, write EXACTLY THREE distractor lines that those other characters would plausibly say in this context.
Return ONLY this JSON:

```
{
  "distractors":[
    {"text":"...", "by":"<OtherCharacterName>"},
    {"text":"...", "by":"<OtherCharacterName>"},
    {"text":"...", "by":"<OtherCharacterName>"}
  ]
}
```

Context (narration BEFORE anyone speaks): `"""{context_text}"""`
TARGET (do NOT imitate in distractors): - name: {target_name} ... (details omitted)
THREE OTHER characters (write one distractor for each; must sound like them): 1) name: {o1_name} ...
Rules (STRICT):

- **Context fit**: Each distractor must be logically possible...
- **Persona fit**: Each distractor must match the specified...
- **Safety checks**: If any distractor contradicts the context...

Output ONLY the JSON object described above.

Table 10: The canonical prompt for generative persona imitation (Social Persona).

Generative Persona Imitation Prompt (General)

You are acting as a digital twin of a specific social media user. Your task is to analyze the user's posting history to understand their personality, tone, vocabulary, and style. All provided text (history, context) is in the original language of the data. You must analyze and respond in that language.
Here is the user's posting history: (Note: "Context" is a post/message by someone else, and "UserReply" is the user's own reply to it.) — history_text —
Now, you must imitate this user's persona perfectly and write a new reply to the following message. Respond ONLY with the text of the reply. Do not add any extra explanations, greetings, or surrounding text.
Message to reply to: "context"

Table 11: The generative prompt for the messaging dimension (JSON Output).

LMUT Prompt (Messaging Variant, JSON Output)

You are acting as a digital twin of a specific messaging app user. Your task is to analyze the user's messaging history to understand their personality, tone, vocabulary, and style.
Here is the user's messaging history: — history_text —
Now, you must imitate this user's persona perfectly and write a new reply to the following message. Please include your response in the following JSON format:

```
{"generated_content": "your reply text here"}
```

Message to reply to: "context"

Table 12: The generative prompt for the narrative dimension (Dimension 3).

Digital Twin Line Generation (Narrative Variant)

You are the digital twin of the TARGET speaker in a literary dialogue dataset. Your job: write ONE new reply that this TARGET would plausibly say in the exact scene below...
Inputs: TARGET speaker: {speaker}; Scene context: `"""{context}"""`; (Opt) History: {history_block}
Hard requirements (STRICT)

1. Language & Era: Match the book's tone/era...
2. Persona Fit: Keep the TARGET's typical formality...
3. Output format: Return ONLY a JSON object: `{ "generated_content": "<single line>" }`

Table 13: The prompt used by the LLM-as-a-Judge to rank candidate replies.

Judge Ranking Prompt (General)

You are an expert evaluator of writing style. Your task is to compare several candidate replies against a known "Reference Reply" written by a specific user. Your goal is to identify which candidate is the most similar to the reference in terms of **style, tone, vocabulary, sentiment, and topic**.
This is the Reference Reply (the ground truth): — ground_truth_reply —
These are the **Candidate Replies**: candidate_replies_text
Now, determine which single candidate is the closest match to the Reference Reply. You MUST respond ONLY with a JSON object in the following format.

```
{
  "choice": "The letter (A/B/C/D)",
  "reasoning": "A brief explanation..."
}
```

Table 14: The letter-only prompt used for ranking in the narrative dimension.

MAP Prompt (Narrative Variant, Letter Only)

You are a strict classifier. Output ONLY a single letter (A/B/C/D). Choose the option that best matches the style, tone, vocabulary, and stance of the Generated Reply.
[Options] A. {A} B. {B} C. {C} D. {D}
[Generated Reply] {pred}
Output exactly one letter: A, B, C, or D.

Table 15: The prompt for the LLM-as-a-Judge scoring task, including the scoring rubric.

Judge Scoring Prompt (All Variants)

You are a meticulous and objective evaluator for a digital twin benchmark... The evaluation rests on three key pillars:

1. **Opinion Consistency**
2. **Logical & Factual Fidelity**
3. **Stylistic Similarity**

— SCORING RUBRIC (1–5 Scale):

- **5: Perfect Replication**: Perfect match across all three pillars...
- **4: High Fidelity**: Opinion/Logic match, minor style diff...
- **3: Core Alignment**: Core opinion consistent, detail loss...
- **2: Partial Relevance**: Major failure in one pillar...
- **1: Irrelevant**: Contradictory or unrelated...

—
YOUR TASK: Respond ONLY with a JSON object:

```
{
  "analysis": { ... },
  "final_score": "1-5",
  "final_justification": "..."
}
```

Context Message: "{context}"
Ground Truth Reply: "{ground_truth_reply}"
Generated Reply: "{lmut_reply}"

Table 16: The canonical annotation prompt used to label the six fundamental capabilities and identify the primary capability.

Capability Annotation Prompt (Canonical)

 # ROLE AND GOAL
You are an expert linguistic and persona analyst. Your task is to analyze user data to identify the core capabilities a generative model would need to successfully create a "digital twin" of the user. You will be given a user's conversational history, a new context they are replying to, and their actual response ("groundtruth").
 # INPUT DATA STRUCTURE
You will receive a JSON object with: `context` (situation), `groundtruth_response` (actual reply), and `history` (past posts).
 # CORE TASK: CAPABILITY ANNOTATION
Part 1 (Mandatory): Identify the single "primary_capability" (the best-fit label).
Part 2 (Detail): Evaluate all six capabilities (C1–C6), marking "true" or "false" based on strict criteria.
 # CAPABILITY DEFINITIONS AND CRITERIA (Evaluate Independently)
C1: Opinion Consistency

- *Core Question*: Does this response require explicitly reaffirming a specific, previously-stated opinion?
- *Label "true" if*: The response expresses a clear opinion that directly reinforces one from history.
- *Primary if*: The core purpose is to state a known, consistent opinion.

C2: Memory Recall

- *Core Question*: Does the response rely on shared context or information from history?
- *Label "true" if*: Makes reference to past events/info not in the current context.
- *Primary if*: The response would be confusing or lose meaning without knowledge of history.

C3: Logical Reasoning

- *Core Question*: Does this response provide a justification or explanation for a claim?
- *Label "true" if*: Contains rationale (e.g., "because", "since") matching the user's pattern.
- *Primary if*: The response structure is clearly "Claim + Justification".

C4: Lexical Fidelity

- *Core Question*: Does this response use a creative, personal, and repeated signature word/phrase?
- *Label "true" if*: Uses an idiosyncratic word/phrase/emoji repeated in history.
- *Primary if*: The most noticeable feature is the signature word/phrase.

C5: Persona Tone

- *Core Question*: Does the response use a specific, non-literal tone (like sarcasm or deep irony)?
- *Label "true" if*: History shows a pattern of this tone AND the response is a clear instance of it.
- *Primary if*: The meaning is inverted or altered by a clear, persona-defining tone.

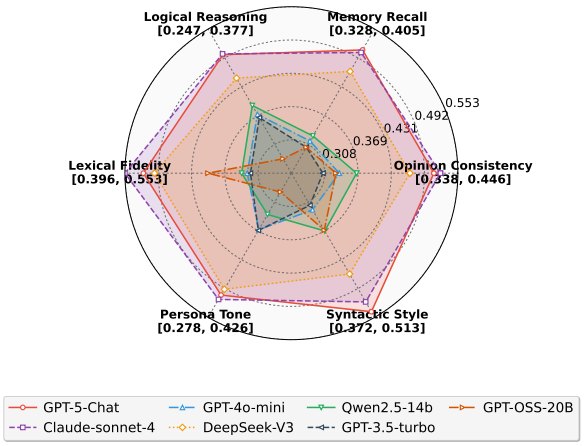
C6: Syntactic Style

- *Core Question*: Does this response use a distinctive, repeated structural pattern?
- *Label "true" if*: Uses a clear, repeated, non-standard stylistic pattern (e.g., fragments).
- *Primary if*: The response is very simple and defined by a structural quirk.

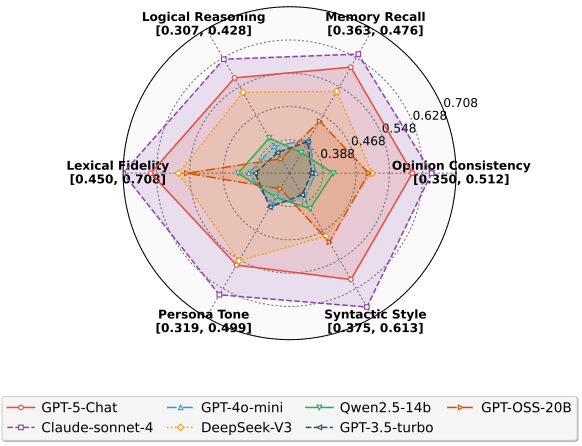
 # INSTRUCTIONS & OUTPUT FORMAT

1. **Step 1**: Determine "primary_capability" (give equal consideration to all capabilities first).
2. **Step 2**: Evaluate all six capabilities (assign "true"/"false" with brief justification).
3. **Step 3**: Output a single JSON object:

```
{
  "primary_capability": "Name_Of_The_Single_Best_Fit_Capability",
  "all_evaluations": {
    "Opinion_Consistency": { "label": false, "reasoning": "..." },
    "Memory_Recall": { "label": false, "reasoning": "..." },
    "Logical_Reasoning": { "label": false, "reasoning": "..." },
    "Lexical_Fidelity": { "label": false, "reasoning": "..." },
    "Persona_Tone": { "label": false, "reasoning": "..." },
    "Syntactic_Style": { "label": false, "reasoning": "..." }
  }
}
```

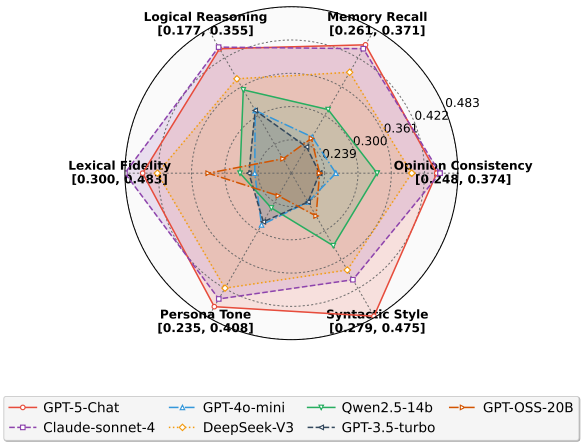


(a) Combined Average

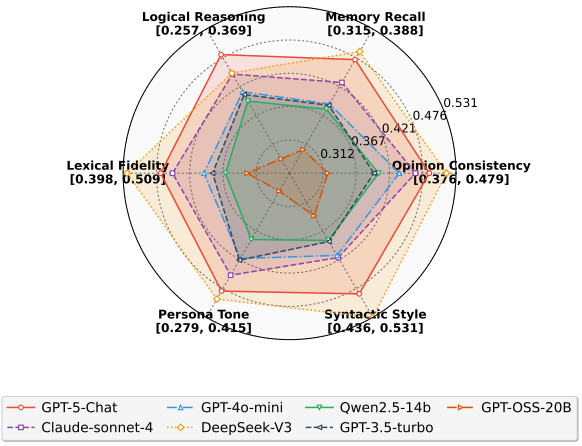


(b) Discriminative

Figure 4: Dimension 1 (Social Persona): (a) Combined Average radar over six capabilities (all labeled capabilities). Aggregates across discriminative and generative protocols; higher is better along each spoke. (b) Discriminative evaluation radar (accuracy-based) across six capabilities (all labeled capabilities). Shows multiple-choice persona matching performance; higher is better.

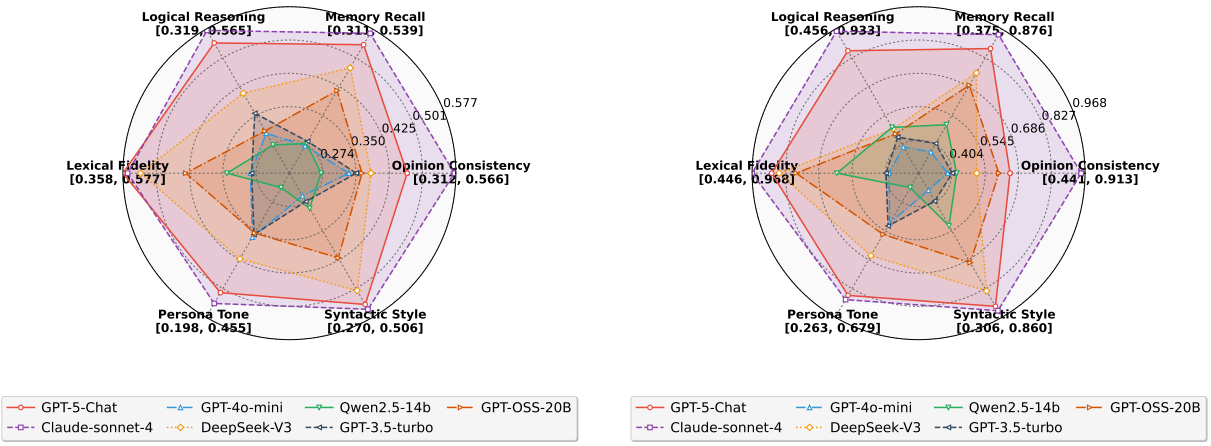


(a) Generative Ranking



(b) Generative Scoring

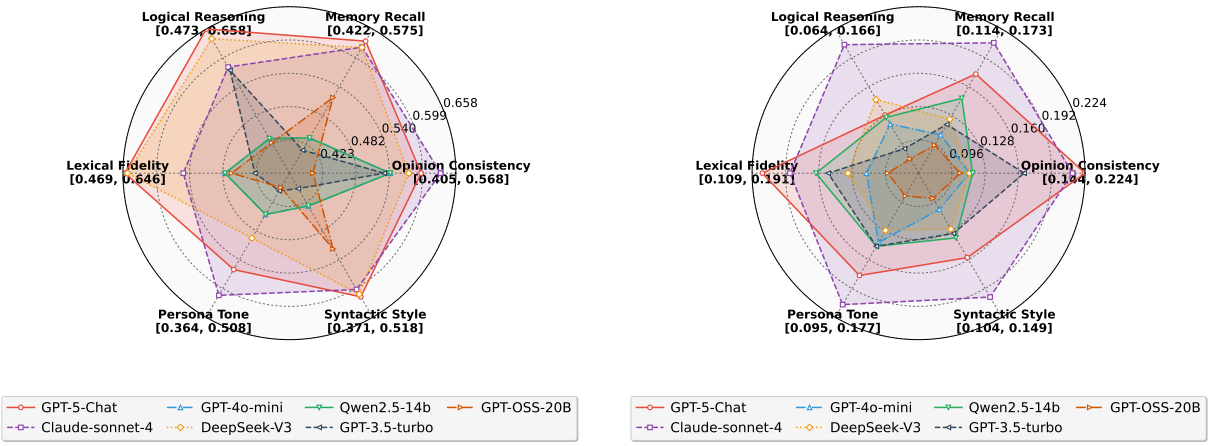
Figure 5: Dimension 1 (Social Persona): (a) Generative Ranking radar (LLM-as-a-Judge, Acc.(Gen)) across six capabilities (all labeled capabilities). Reflects relative imitation quality; higher is better. (b) Generative Scoring radar (LLM-as-a-Judge, Score(Gen), 1–5) across six capabilities (all labeled capabilities). Captures absolute similarity to the ground truth; higher is better.



(a) Combined Average

(b) Discriminative

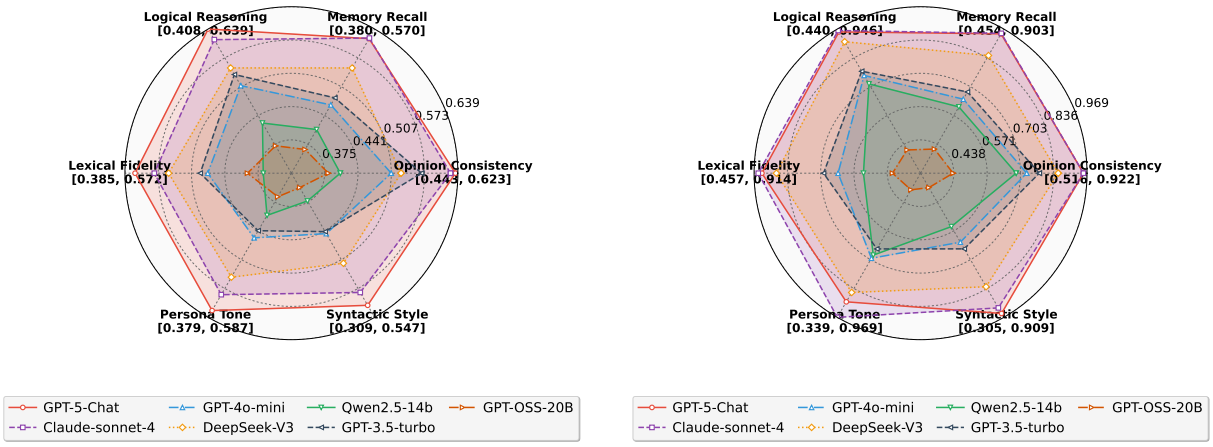
Figure 6: Interpersonal Persona: (a) Combined Average radar over six capabilities (all labeled capabilities). Aggregates across discriminative and generative protocols; higher is better along each spoke. (b) Discriminative evaluation radar (accuracy-based) across six capabilities (all labeled capabilities). Shows multiple-choice persona matching performance; higher is better.



(a) Generative Ranking

(b) Generative Scoring

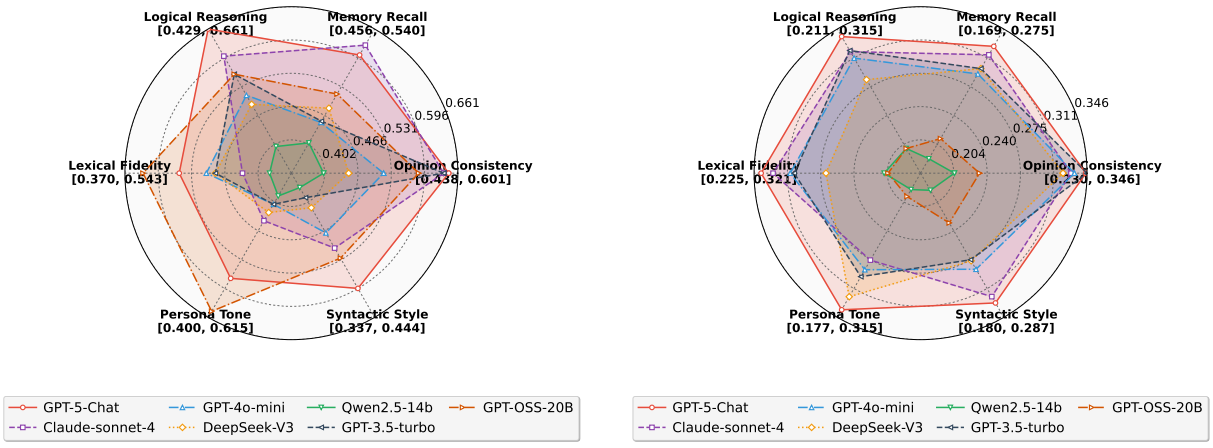
Figure 7: Interpersonal Persona: (a) Generative Ranking radar (LLM-as-a-Judge, Acc.(Gen)) across six capabilities (all labeled capabilities). Reflects relative imitation quality; higher is better. (b) Generative Scoring radar (LLM-as-a-Judge, Score(Gen), 1–5) across six capabilities (all labeled capabilities). Captures absolute similarity to the ground truth; higher is better.



(a) Combined Average

(b) Discriminative

Figure 8: Narrative Persona: (a) Combined Average radar over six capabilities (all labeled capabilities). Aggregates across discriminative and generative protocols; higher is better along each spoke. (b) Discriminative evaluation radar (accuracy-based) across six capabilities (all labeled capabilities). Shows multiple-choice persona matching performance; higher is better.



(a) Generative Ranking

(b) Generative Scoring

Figure 9: Narrative Persona: (a) Generative Ranking radar (LLM-as-a-Judge, Acc.(Gen)) across six capabilities (all labeled capabilities). Reflects relative imitation quality; higher is better. (b) Generative Scoring radar (LLM-as-a-Judge, Score(Gen), 1–5) across six capabilities (all labeled capabilities). Captures absolute similarity to the ground truth; higher is better.