

From Memory to Behavior: A Behavior-Aware Role-Playing Framework for Social Media Influencers

Ji-Lun Peng^{*†} Yi-Zhen Zhang[†] Chun-Nan Chou[†] Yun-Nung Chen^{*}

^{*}National Taiwan University, Taipei, Taiwan

[†]CMoney Technology Corporation, Taiwan

r13946006@ntu.edu.tw {jason_zhang, jason_chou}@cmoney.com.tw

y.v.chen@ieee.org

Abstract

Large language models have shown strong potential as role-playing agents for real individuals, yet faithful impersonating remains challenging. Existing in-context learning-based methods fail to capture how individuals react under different situations. In addition, LLM-based evaluation is difficult for obscure individuals. To address these challenges, we propose **Situation–Internal state–Behavior Persona** method to incorporate situation-dependent behavioral strategies. We further design an evaluation protocol that provides LLM evaluators with references about the impersonated individual. We evaluate our approach on a newly constructed dataset for the task of generating replies on social media. Experimental results show that our proposed method outperforms state-of-the-art ICL-based baselines, while our evaluation protocol achieves moderate correlation with human judgment. Besides, experiments on fictional-character benchmarks demonstrate that our proposed method is applicable beyond the social media setting. These findings suggest that incorporating behavioral information broadly improves the fidelity of role-playing for real individuals on social media or fictional characters.¹

1 Introduction

Recent advances in large language models (LLMs) have accelerated the development of role-playing agents (RPAs) capable of impersonating both fictional characters (Yu et al., 2025; Zhang et al., 2025) and real individuals (Chuang et al., 2025; Wang et al., 2025a). Recent evidence further suggests that RPAs for real individuals can produce convincing impersonations (Shi et al., 2025).

Toward faithful role-playing of specific real individuals, two major challenges remain. **Behavioral Modeling Gap:** Existing in-context learning (ICL)-based role-playing methods mainly provide RPAs

with factual information about impersonated individuals extracted from historical conversations (Liu et al., 2024; Shi et al., 2025; Gao et al., 2025). However, such information does not explicitly capture how an individual tends to react under different situations. As illustrated in Figure 1, in social media comment-replying scenarios between influencers and followers, replies are often brief or knowledgeable, making faithful impersonation depends more on behavioral strategies than on factual information. **LLM Familiarity Gap:** Existing LLM-based evaluation protocols for role-playing are largely developed for fictional characters, whose information are often represented in LLMs’ knowledge (Peng and Chen, 2026). As a result, directly prompting an LLM can provide a reasonable evaluation (Wang et al., 2024; Yu et al., 2025). By contrast, this direct prompting strategy is less applicable to obscure real individuals, whose information may be sparsely represented in LLMs’ pre-training data. This limited familiarity makes it difficult to directly use LLMs for automated evaluation of role-playing real-individual.

To address these challenges, we develop a behavior-aware framework for role-playing real individuals. Inspired by the Cognitive-Affective Processing System (Mischel and Shoda, 1995), we propose **Situation–Internal state–Behavior Persona (SIBPersona)**, a method that models individuals’ situation-dependent behavioral strategies. In addition, we design an automated LLM-based evaluation protocol tailored to role-playing real-individual, which provides the LLM evaluator with reference about the impersonated individual to alleviate its limited familiarity with the individual. We apply this framework to the domain of social media by constructing a dataset of influencer–follower interactions. Using this dataset, we evaluate RPAs on their ability to impersonate influencers, specifically in generating replies to followers’ comments.

The contributions of this paper are three-fold:

¹Code: <https://github.com/MiuLab/SIBPersona>

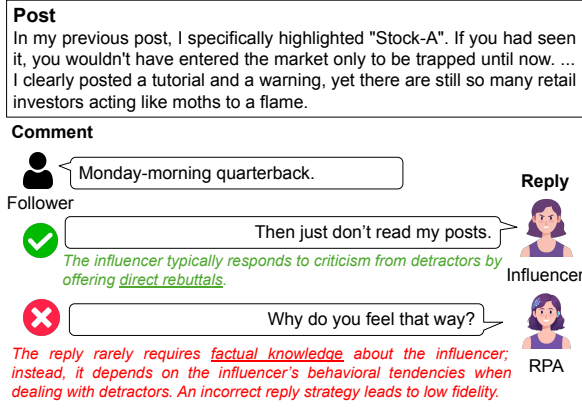


Figure 1: In the social media comment-replying scenarios between influencers and followers, replies involve short texts where behavioral strategy is more important for determining the fidelity of impersonation.

- We present a structured view of persona for RPAs by distinguishing data-invariant and data-variant personae. Building on this view, we propose SIBPersona, which introduces reaction process as a novel dimension of data-variant persona to capture situation-dependent behavioral strategies.
- We establish an reference-augmented evaluation protocol for role-playing of obscure individuals. The protocol achieves moderate correlation with human judgment.
- We conduct comprehensive experiments across multiple influencers, model families, and fictional-character benchmarks, showing that SIBPersona consistently outperforms state-of-the-art ICL-based baselines.

2 Related Work

Approaches of LLM Role-Playing Two mainstream approaches enable LLMs to impersonate specific targets: Fine-Tuning (FT) (Yang et al., 2026; Ye et al., 2025) and ICL. We focus on ICL-based methods because persona information is explicitly represented in prompts, making the role-playing process easier to interpret and modify. Existing ICL-based methods utilize retrieval-augmented generation (RAG) (Lewis et al., 2020) to incorporate context-relevant information into prompts (Shi et al., 2025; Gao et al., 2025). However, the retrieved content primarily consists of raw conversational data or factual knowledge. Such approaches largely emphasize what an individual knows rather than how an individual tends to behave across different interaction situations.

To address this gap, we draw on the Cognitive-Affective Processing System (CAPS) (Mischel and Shoda, 1995), which models behavior as a dynamic reaction process. According to CAPS, individuals perform behaviors by processing situational inputs through internal states. This perspective suggests that RPAs should model the situation-dependent reaction process. Nonetheless, how to explicitly model such a dynamic reaction process in ICL-based method remains underexplored.

Datasets for Role-Playing of Real Individuals

Existing role-playing datasets primarily focus on fictional characters (Xu et al., 2025, 2026), while datasets for role-playing real individuals remain relatively limited. To study role-playing real-individual, we turn to social media, which provides rich conversational records for observing how an individual responds across diverse situations (Li et al., 2025). To support our research goal of impersonating influencers, the dataset must satisfy two requirements: (1) it should contain abundant influencer-centered interactions from a single identifiable individual, allowing LLMs to obtain comprehensive information about the impersonated individual; and (2) it should preserve the comment-reply structure under the influencer’s posts, which is essential for observing how the individual responds to different situations.

As summarized in Table 1, no existing dataset satisfies both requirements simultaneously. Among the public datasets, BluePrint (Bück-Kaeffer et al., 2025) SYNTHIA (Rahimzadeh et al., 2026), PersonalDialog (Zheng et al., 2019), and Persona-Chat (Zhang et al., 2018) do not preserve an identifiable individual’s history: BluePrint anonymizes users and aggregates interactions into clusters, while SYNTHIA provides only synthesized persona attributes rather than raw, individual-level social media interactions. PER-CHAT (Wu et al., 2021) does contain large-scale, identifiable individual histories (314,749 individuals), but lacks the comment-reply structure needed to observe situational responses. Conversely, PersonaConvBench (Li et al., 2025) preserves this structure but provides a limited number of samples per individual (6.87 posts and 37.05 comment-reply pairs on average), making it insufficient for LLMs to obtain comprehensive information about the impersonated individual. These limitations motivate us to construct a new dataset tailored for the task of generating replies on social media.

Dataset	Public	Identifiable Individual History	#Individual	Comment-Reply Structure	Avg. Posts per Individual	Avg. Comment-Reply Pairs per Individual	Max Posts per Individual	Max Comment-Reply Pairs per Individual
BuzzFace (2018)	✗	-	-	-	-	-	-	-
TWON (2026)	✗	-	-	-	-	-	-	-
BluePrint (2025)	✓	✗	-	-	-	-	-	-
SYNTHIA (2026)	✓	✗	-	-	-	-	-	-
PersonalDialog (2019)	✓	✗	-	-	-	-	-	-
Persona-Chat (2018)	✓	✗	-	-	-	-	-	-
PER-CHAT (2021)	✓	✓	314749	✗	-	-	-	-
PersonaConvBench (2025)	✓	✓	2796	✓	6.87	37.05	111	414
Ours	✓	✓	6	✓	1811	5554	2903	15240

Table 1: Feasibility of existing datasets for studying role-playing of real individuals on social media reply generation. ”-” indicates that the corresponding information is not applicable under our comparison criteria. Since our research goal is to impersonate influencers across diverse interaction situations, we construct a dataset centered on a small number of influencers with rich historical interactions.

Evaluation of RPAs LLM evaluation have become widely adopted for evaluating role-playing of fictional character (Gu et al., 2026), as they enable fine-grained assessment of performance (Wang et al., 2025b, 2024; Shao et al., 2023). These approaches typically assume that the evaluator model possesses sufficient knowledge about the impersonated character (Peng and Chen, 2026).

Evaluating role-playing of real individuals remains challenging because LLM evaluators have limited knowledge about obscure people. Consequently, prior studies on role-playing real individuals often rely on human evaluation (Shi et al., 2025; Tu et al., 2023), while automated evaluation protocols for role-playing of obscure real individuals remain underdeveloped.

3 SIBPersona

In this section, we present **SIBPersona** method designed for role-playing of influencers on social media. We first introduce our multifaceted view of persona and define the reaction process in §3.1, followed by formalizing the task of generating replies on social media in §3.2. Subsequently, we elaborate how data-variant persona is constructed from historical data in §3.3. Finally, we explain how the constructed data-variant persona is retrieved and used for reply generation in §3.4.

3.1 Multifaceted Nature of Persona

In RPA, a "persona" is always referred to the information describing fictional characters or real individuals. We argue that personae can be divided into two types: **data-invariant** and **data-variant**. The data-invariant persona, as the name implies, is pro-

vided in advance and is not derived from the data. For example, we can create a narrative to describe the demographic information of the influencer or the fictional traits of the fictional character, as far as we know. On the other hand, the data-variant persona is derived from the data and designated to capture the dynamics of RPA.

Apart from deriving from the data or not, the prior methods mainly modeled the persona by two dimensions: 1) memory (Shi et al., 2025) capturing the factual information and 2) speaking style (Tu et al., 2024) capturing expression habits. However, we contend that these two dimensions are insufficient for RPAs. As suggested in CAPS, responses of real people are shaped not only by what they know and how they speak but also by the reaction process they tend to follow in a given situation.

Reaction Process In order to resolve this issue, we propose to incorporate so-called reaction process, the third dimension, into the data-variant persona. To this end, we model reaction process as the mapping of a given situation to an internal state and subsequently to a behavioral strategy. Formally, reaction process is defined by $\{\langle s, i, b \rangle \in S \times I \times B \mid s = LLM(x), i = LLM(x), b = LLM(x)\}$, where S contains situations, I represents the set of internal states, B is the set of behavioral strategies, and x denotes a text input. We call this three-tuple, namely $\langle s, i, b \rangle$, an SIB triplet. In the following, we utilize the example that influencers reply followers’ comments on social media to illustrate the idea of reaction process. When influencers reply to followers’ comments, they first interpret the comments into high-level situations, e.g., praise, inquiry, or hostility. This emerging situation then

triggers a specific internal state, such as happy or angry, which leads to a certain behavioral strategy, such as expressing gratitude or defensive rebuffing. Grounded in CAPS, this three-dimensional data-variant persona provides a more psychologically faithful modeling of how real individuals respond.

3.2 Problem Formulation

After illustrating the three-dimensional data-variant persona, we provide the formal settings of the problem for easy exposition. In general, RPAs are designed to engage in conversations with users by emulating specific characters or impersonating specific individuals. In this work, we focus on the generation of single-turn replies for influencers on social media. To be specific, we generate the reply r according to the influencer’s post ip , the follower’s comment fc , the influencer’s data-invariant profile p , and the influencer’s data-variant persona PS_{dv} , which can be represented as:

$$r = RPA(ip, fc, p, PS_{dv}).$$

In particular, PS_{dv} contains information about the three aforementioned dimensions.

3.3 Constructing Data-Variant Persona

In this subsection, we elaborate on how each dimension is extracted and represented. We first view a data sample associated with a specific influencer as $ds_j = \langle ip_j, CRP_j \rangle$, where ip_j is the influencer’s post, and $CRP_j = \{\langle fc_{j1}, r_{j1} \rangle, \langle fc_{j2}, r_{j2} \rangle, \dots, \langle fc_{jn}, r_{jn} \rangle\}$ denotes the set of all comment-reply pairs associated with ip_j post. As a consequence, the historical data of this influencer is a collection of the data samples, denoted as $\mathcal{D} = \{ds_1, ds_2, \dots, ds_n\}$.

In accordance with the previous works of extracting data-variant personae by LLMs (Shi et al., 2025; Liu et al., 2024), we employ a two-stage pipeline to produce data-variant personae from raw social media histories. Our entire pipeline is depicted in the upper part of Figure 2 and is explained in detail below.

Stage 1: Dimensional Relevance Filtering Due to data noise on social media, we employ a recognizing LLM to filter irrelevant information in the first stage to increase the efficacy of the data pipeline. Specifically, the recognizing LLM acts like a multi-label classifier that assigns each data sample to the set of dimensional labels $L = \{l_M, l_T, l_{RP}\}$. That is, $\forall j, h(ds_j) = \mathcal{A}_j, \mathcal{A}_j \subseteq L$,

where h is a mapping function from a data sample to the subset of L . As a shorthand, we use the notations M, T, RP to represent memory, speaking style, and reaction process, respectively. If the recognizing LLM assigns the data sample ds_j to the empty set, i.e. $h(ds_j) = \emptyset$, ds_j is discarded. Otherwise, ds_j proceeds to the next stage.

Stage 2: Dimensional Document Extraction

Given a data sample $ds_j = \langle ip_j, CRP_j \rangle$ such that $h(ds_j) \neq \emptyset$, we use an extracting LLM to generate the dimensional information of ds_j for each label $l \in h(ds_j)$. We introduce **dimensional documents** whose details are elaborated later as a countable surrogate for the generated dimensional information. Given ds_j and a label $l \in h(ds_j)$, the extracting LLM may generate more than one dimensional document, and we use k to index the generated dimensional documents.

To equip our approach with interpretability, we define a subset of CRP_j as a set of **evidence pairs** denoted EP_{jk} , and EP_{jk} provides the grounding evidence for the generated dimensional information. If the label l is l_M , the extracting LLM generates a memory document $d_{jk}^{(l_M)} = \langle mp_{jk}, EP_{jk} \rangle$, where mp_{jk} denotes a memory point capturing factual information. When the label l is l_T , the extracting LLM produces a speaking style document $d_{jk}^{(l_T)} = \langle sp_{jk}, EP_{jk} \rangle$, where sp_{jk} represents a speaking style point reflecting a expression habit. Otherwise, the extracting LLM yields a reaction process document $d_{jk}^{(l_{RP})} = \langle \langle s_{jk}, i_{jk}, b_{jk} \rangle, EP_{jk} \rangle$, where $\langle s_{jk}, i_{jk}, b_{jk} \rangle$ constitutes an SIB triplet comprising the situation, internal state, and behavioral strategy, respectively.

Depending on the type of dimensions, we store the extracted dimensional document in the corresponding **dimensional corpus**, which is a collection of dimensional documents. Thus, our data-variant persona PS_{dv} has three dimensional corpora: memory corpus, speaking style corpus, and reaction process corpus. These three dimensional corpora lay the foundation for the reply generation illustrated in the next subsection.

3.4 Reply Generation

After the data-variant persona is established, one intuitive idea is to generate the reply by considering all dimensional documents in these three dimensional corpora as the demonstrations to LLMs. However, this idea is inherently infeasible since influencers on social media always have numer-

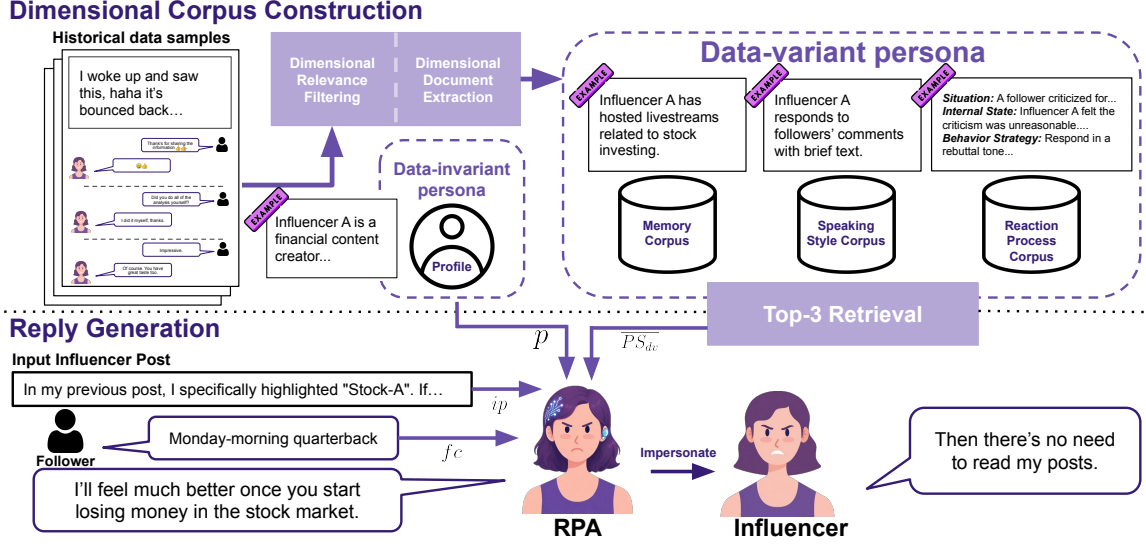


Figure 2: Overview of the SIBPersona method. The upper part illustrates the process of extracting data-variant persona from historical data. The lower part presents reply generation process: retrieving the most relevant dimensional documents based on the comment enables the RPA to generate a high-fidelity reply.

ous posts and comment-reply pairs; otherwise, they could hardly be called "influencers." Instead of using PS_{dv} directly, we utilize the approximate data-variant persona $\overline{PS}_{dv}$ derived from PS_{dv} as an alternative.

As depicted in the lower part of Figure 2, we utilize the concept of RAG (Lewis et al., 2020) to derive $\overline{PS}_{dv}$. Given the query fc , we independently retrieve the top three dimensional documents which are most relevant to the query fc from each dimensional corpus. The relevance measure is calculated as follows. We treat mp_{jk} , sp_{jk} , and s_{jk} in the SIB triplet as the retrieval keys for memory corpus, speaking style corpus, and reaction process corpus, respectively. For each dimensional corpus, we compute the cosine similarity between the embedding of query fc and the embedding of the retrieval keys. Afterwards, we rank the dimensional documents by similarity scores and retrieve the top three corresponding dimensional documents for each dimensional corpus. Finally, the nine retrieved dimensional documents constitute $\overline{PS}_{dv}$, serving as demonstrations to LLM. The problem we address becomes:

$$r = RPA(ip, fc, p, \overline{PS}_{dv}).$$

Figure 11 shows an example of a retrieved dimensional document and the reply generated by SIBPersona. Additional implementation details are provided in Appendix B.

4 Reference-Augmented Evaluation

In this section, we propose a reference-augmented protocol for automatically evaluating role-playing agents. We first explain the challenges of this task, then detail our design, and conclude by empirically validating its alignment with human judgment.

4.1 Penalty-Based LLM Evaluator

To enable scalability and facilitate future iterations, we adopt LLM evaluators to automatically assess role-playing agents for the task of generating replies on social media. However, as addressed in §2, LLM evaluators may have limited knowledge of obscure individuals. To resolve this knowledge scarcity issue, we augment the LLM evaluator with some references called **representative dimensional documents** about the impersonated individual. The representative dimensional documents are constructed as follows. We first apply the method described in §3.3 to construct the three dimensional corpora from the test samples and then select representative dimensional documents, denoted as rdd_l , from the resulting dimensional corpus for each dimension $l \in L$. Specifically, we apply K-medoids clustering to select $\sqrt{M}/2$ representative dimensional documents from each dimensional corpus, where M is the size of dimensional documents. Clustering is performed based on the embeddings of s_{jk} , mp_{jk} and sp_{jk} . Crucially, the representative dimensional documents are constructed and sampled exclusively from the test samples rather than

Methods	GPT-5.2			Gemini-3-flash			DeepSeek-V3.2			Claude-Opus-4.6			Average		
	<i>M</i>	<i>T</i>	<i>RP</i>	<i>M</i>	<i>T</i>	<i>RP</i>	<i>M</i>	<i>T</i>	<i>RP</i>	<i>M</i>	<i>T</i>	<i>RP</i>	<i>M</i>	<i>T</i>	<i>RP</i>
Naive	97.50	46.40 [‡]	67.53 [‡]	94.70	42.47 [‡]	69.27 [‡]	95.13	48.93 [‡]	70.83 [‡]	96.77	46.93 [‡]	76.87 [‡]	96.03	46.18 [‡]	70.78 [‡]
CoT	97.47	44.93 [‡]	70.73 [‡]	95.83	41.17 [‡]	69.53 [‡]	96.27	49.23 [‡]	71.47 [‡]	97.33	47.00	77.33 [‡]	96.73	45.58 [‡]	72.27 [‡]
IMPersona	96.07	47.50 [‡]	71.93 [‡]	94.90	44.60 [‡]	70.10 [‡]	96.00	50.80 [‡]	72.50 [‡]	97.43	48.67 [‡]	78.13 [‡]	96.10	47.89 [‡]	73.17 [‡]
Vanilla RAG	97.33	49.43 [‡]	73.20 [‡]	93.77	53.20	71.50 [‡]	94.23	58.37 [‡]	73.43	96.70	56.93	77.70 [‡]	95.51	54.48 [‡]	73.96 [‡]
SIBPersona	96.90	54.33	77.63	93.77	56.47	75.50	94.83	63.67	76.43	96.93	62.77	81.93	95.61	59.31	77.88

Table 2: Performance on *Influencer A*. **Bold**: the best results; [‡]: significant degradation compared to SIBPersona ($p < 0.05$).

Methods	Deepseek-V3.2			Gemini-3-flash		
	<i>M</i>	<i>T</i>	<i>RP</i>	<i>M</i>	<i>T</i>	<i>RP</i>
Naive	92.10 $\pm$ 3.04	42.03 [‡] $\pm$ 8.43	68.48 [‡] $\pm$ 5.86	92.50 $\pm$ 2.32	43.33 [‡] $\pm$ 7.35	70.64 [‡] $\pm$ 7.37
CoT	92.17 $\pm$ 3.03	43.92 [‡] $\pm$ 8.13	69.83 [‡] $\pm$ 6.14	92.15 $\pm$ 2.54	44.05 [‡] $\pm$ 7.08	70.48 [‡] $\pm$ 6.40
IMPersona	91.60 $\pm$ 2.21	45.54 [‡] $\pm$ 5.86	71.14 [‡] $\pm$ 6.02	91.42 $\pm$ 1.94	45.21 [‡] $\pm$ 6.40	71.56 [‡] $\pm$ 7.19
Vanilla RAG	91.39 $\pm$ 3.23	57.05 [‡] $\pm$ 2.43	72.43 [‡] $\pm$ 5.06	92.19 $\pm$ 2.46	53.43 [‡] $\pm$ 4.92	72.59 [‡] $\pm$ 4.79
SIBPersona	92.14 $\pm$ 3.26	61.09 $\pm$ 1.66	75.84 $\pm$ 4.90	92.21 $\pm$ 2.20	57.34 $\pm$ 3.20	76.48 $\pm$ 5.08

Table 3: Results on other five influencers. **Bold** indicates the better scores. [‡]: significant degradation compared to SIBPersona ($p < 0.05$).

Methods	<i>M</i>	<i>T</i>	<i>RP</i>
SIBPersona	92.67 $\pm$ 3.18	61.60 $\pm$ 1.82	75.96 $\pm$ 4.49
w/o <i>I</i>	92.61 $\pm$ 2.49	60.30 $\pm$ 2.37	75.44 $\pm$ 3.61
w/o <i>RP</i>	92.83 $\pm$ 2.44	57.82 [‡] $\pm$ 2.51	73.36 [‡] $\pm$ 4.64

Table 4: Dimension ablation study using DeepSeek-V3.2 on the six influencer. **Bold**: the best results; [‡]: significant degradation compared to SIBPersona ($p < 0.05$).

the demonstration examples. This ensures that the LLM evaluator’s reference standard remains independent of the role-playing agent’s demonstration source, so that the evaluation reflects the agent’s true generalization ability rather than its capacity to reproduce seen demonstrations.

For the clarity of presentation, we define a test sample as $ts_j = \langle ip_j, fc_j, r_j^a \rangle$, where r_j^a is the reply generated by an RPA. For each dimension $l \in L$, our LLM evaluator takes ts_j , p , rdd_l , and the ground-truth reply r_j^{gt} written by the influencer as input. Following Wang et al. (2025c), our LLM evaluator is instructed to identify the flaws in r_j^a and to assign a severity level ranging from 1 to 5 for each flaw at the same time. The final score for the dimension l is computed by averaging the scores on all test samples whose size is N , which can be summarized as follows:

$$Score_l = \frac{1}{N} \sum_{j=1}^N \max(0, 100 - 5 \cdot \sum_{f \in F_{lj}} v_f),$$

where F_{lj} is the set of the identified flaws for di-

mension l on r_j^a , and v_f is the severity level of flaw f . Additional details regarding the evaluation implementation are provided in Appendix C.

4.2 Alignment with Human judgment

To assess whether our reference-augmented evaluation protocol is well-aligned with human judgment, we examine the correlation between human and LLM evaluator scores. For evaluation scores averaged across the three dimensions, the alignment between human and LLM evaluator yields a Pearson’s r of 0.561 and a Spearman’s ρ of 0.530, indicating a moderate correlation by conventional social-science standards (Cohen, 1988; Lovakov and Agadullina, 2021) and also consistent with the previous findings in role-playing evaluation (Zhou et al., 2025; He et al., 2025). These results demonstrate that our reference-augmented protocol serves as a reasonable proxy for human judgment. Additional details of the correlation analysis are provided in Appendix D.2, and the advantages of our protocol over traditional NLP metrics are reported in Appendix E.2.

5 Experimental Evaluation

In this section, we present the experimental evaluation of SIBPersona. We first describe the process of building the dataset in §5.1, followed by the experimental settings in §5.2. We then show the results of the experiments conducted on six influencers to demonstrate the effectiveness of SIBPersona in

§5.3. To validate our SIBPersona, we perform a human evaluation to examine whether the *RP* dimension provides interpretable evidence in §5.4. Lastly, we show that SIBPersona is also applicable to fictional character benchmarks in §5.5.

5.1 Dataset Construction

We construct our dataset from posts, comments, and replies associated with six influencers, denoted as *Influencer A* to *F* from **CMoney Forum**². We collect the data span from January 1, 2024, to September 30, 2025. The data whose dates are before August 1, 2025 belong to the demonstration source, while the remaining data are preserved as the candidate pool for test samples.

To have a fair comparison between the different influencers, we perform stratified sampling on the candidate pool and keep the size of the test samples for each influencer in the same order of magnitude. In addition, our preliminary analysis reveals that different comment–reply pairs exhibit distinct interaction types. To ensure that the test samples thoroughly cover these diverse interactions, we prompt gpt-4o-mini to assign each comment–reply pair a score reflecting the degree of expressing personal characteristic. Based on these scores, the data are partitioned into three levels: high, medium, and low. From each level, we select 50 test samples at most. Additional details regarding dataset construction are provided in Appendix A.

5.2 Experimental Setting

We compare our method with several ICL-based role-playing methods, all of which take the impersonated individual’s profile p as default input. We briefly describe each compared method below. **Naive** relies solely on the impersonated individual’s profile. **CoT** builds upon the **Naive** by additionally incorporating Chain-of-Thought reasoning (Wei et al., 2022). **Vanilla RAG** augments the prompt with nine retrieved raw comment–reply pairs to match the number of dimensional documents utilized in SIBPersona. Finally, **IMPersona** (Shi et al., 2025) is a state-of-the-art ICL-based method for role-playing real individuals in conversations, which uses the proposed hierarchical memory module to model its data-variant persona.

²CMoney Forum is the largest online investor community in Taiwan, where members post Traditional Chinese messages under individual stock tickers and follow financial influencers whose posts routinely attract numerous replies.

Methods	M	T	RP	Avg.
IMPersona	2.92 [‡]	2.40 [‡]	2.67 [‡]	2.67 [‡]
Vanilla RAG	3.68	3.72	3.62	3.67
SIBPersona	3.78	3.64	3.62	3.68

Table 5: Human evaluation results comparing our framework with the baseline. **Bold**: better results; [‡]: significant degradation compared to SIBPersona ($p < 0.05$).

Conditions	M	T	RP
Uninterpretable	3.03	3.00	2.87
Interpretable	4.13 [†]	3.97 [†]	4.00 [†]

Table 6: Human evaluation scores for Interpretable ($n=20$) and Uninterpretable ($n=10$) groups. **Bold**: better results; [†]: significant improvement with $p < 0.05$.

We conduct two types of experiments to evaluate the robustness of SIBPersona along two axes: model choice and impersonated influencer. First, we examine model-agnostic robustness by applying multiple LLMs to construct data-variant personas, including GPT-5.2, Gemini-3-flash, DeepSeek-v3.2, and Claude-Opus-4.6. These experiments are conducted on *Influencer A* to assess whether the effectiveness of SIBPersona is robust to the choice of the LLM used for persona construction, rather than being tied to a specific model. Second, we examine the robustness of cross-influencers by applying SIBPersona to *Influencer B* through *Influencer F*. In these experiments, we use Gemini-3-flash and DeepSeek-v3.2 to evaluate whether SIBPersona remains effective across different influencers. All experiments are evaluated using the reference-augmented evaluation protocol introduced in §4 in all three dimensions, with gpt-4.1 as the LLM evaluator and text-embedding-3-small as the embedding model for retrieval. For statistical significance testing, we perform t-tests with Holm correction for multiple comparisons.

5.3 Main Results

Table 2 shows the results of *Influencer A* in four LLMs. On average, SIBPersona outperforms all baselines by a statistically significant margin on both the T and RP dimensions. Across all methods, performance on the M dimension remains uniformly high, reflecting that replies on social media naturally demand less factual memory information and thus leave minimal room for penalties. In terms

of the other two dimensions, Naive and CoT perform the worst due to insufficient information regarding the impersonated individual. While IMPersona incorporates rich memory information, it fails on speaking style and reaction process due to its single-dimensional design. Vanilla RAG emerges as the strongest baseline, indicating that simply mirroring similar historical replies yields reasonable results. Nevertheless, Vanilla RAG still falls short of SIBPersona, showing that constructing structured, multi-dimensional data-variant persona is crucial for faithful role-playing.

Table 3 reports the results aggregated over *Influencer B* through *Influencer F*. SIBPersona consistently achieves the highest scores in the *T* and *RP* dimensions across all five influencers under both models. Nearly all baselines exhibit statistically significant degradation compared to SIBPersona, and the advantage is consistent regardless of model family or impersonated influencer. These results show that SIBPersona is robust along two orthogonal axes: model choice and impersonated individual. From a qualitative perspective, Figure 12 illustrates that SIBPersona better captures the behavioral strategy and speaking style of the impersonated individual facing both praise and criticism. Detailed experimental results and qualitative analysis are provided in Appendix E.1 and G.

The ablation study further demonstrates the importance of the *RP* dimension. As shown in Table 4, removing the *RP* dimension significantly degrades the performance on the *T* and *RP* dimensions, confirming that the *RP* dimension contributes substantially to overall performance. In contrast, removing only the internal state component causes marginal, non-significant declines. This asymmetry suggests that while the full SIB triplet is beneficial, explicit behavioral strategies *B* contribute more critically than intermediate internal-state *I* for the social media reply generation task, where replies tend to be short and direct.

5.4 Human Evaluations

To validate the effectiveness of SIBPersona, we perform a human evaluation in which the annotators rate the replies generated by SIBPersona, Vanilla RAG and IMPersona along three dimensions using a five-point Likert scale (Joshi et al., 2015). As shown in Table 5, SIBPersona significantly outperforms IMPersona across all three dimensions, while performing comparably to Vanilla RAG. Through qualitative analysis, we observe

that Vanilla RAG frequently directly replicates the replies from the retrieved replies. This tendency often leads to misaligned replies when discrepancies arise between the follower’s comment and the retrieved context. In contrast, SIBPersona generates replies grounded in the structured points, allowing it to adaptively fit the current situation. While such fine-grained differences are less distinguishable under the small-scale human evaluation, the adaptability of SIBPersona is more clearly reflected in our large-scale LLM-based evaluations. Additional details and results of the human evaluation are provided in Appendix D.

Interpretability of SIBPersona We further examine the interpretability of SIBPersona. In this type of experiments, the annotators are asked to assess whether the SIB triplets in $\overline{PS}_{dv}$, especially the internal state and behavior strategy, can provide reasonable explanations for the generated reply given a test sample. Based on these assessments, we divide the test samples into interpretable and uninterpretable groups. Additional details of the interpretability experiments are provided in Appendix D.2. As shown in Table 6, the replies in the interpretable group significantly outperform those in the uninterpretable group across all three dimensions. This result suggests that the interpretability of the SIB triplets in $\overline{PS}_{dv}$ influences the quality of the generated replies and subsequently contributes to improved individual fidelity. In addition, we provide a qualitative example in Appendix G.3, showing that SIBPersona can facilitate error diagnosis.

5.5 Results on CharacterEval

To examine whether SIBPersona is applicable beyond the social media setting, we evaluate our method on CharacterEval (Tu et al., 2024), a fictional character benchmark formulated as a single-turn response generation task. To minimize confounding effects from a model’s prior knowledge of well-known fictional characters, we follow the anonymized evaluation setting proposed by Peng and Chen (2026). We construct dimensional corpora using gpt-4o-mini and generate replies using OLMo-3-7B-Instruct.

As summarized in Table 7, SIBPersona significantly outperforms IMPersona on three out of four evaluation dimensions. Remarkably, SIBPersona also achieves performance competitive with CoSER-Llama-3.1-8B (Wang et al., 2025c), a

Methods	CA	RA	CK	CP
IMPersona	2.94	2.43	2.25	2.38
SIBPersona	3.28 [†]	2.66[†]	2.32[†]	2.69[†]
<i>Fine-tuned CoSER</i>	3.35	2.61	2.28	2.68

Table 7: Results on CharacterEval. **CA**: Conversational Ability, **RA**: Role-Playing Attractiveness, **CK**: Character Consistency–Knowledge, and **CP**: Character Consistency–Persona. **Bold**: the best results; [†]: significant improvement with $p < 0.05$.

Methods	CA	RA	CK	CP
SIBPersona	3.28	2.66	2.32	2.69
w/o T	3.33	2.69	2.31	2.74
w/o I	3.28	2.66	2.31	2.68
w/o RP	3.16 [‡]	2.61	2.30	2.62 [‡]
w/o RP & T	3.05 [‡]	2.54 [‡]	2.28	2.57 [‡]

Table 8: Component ablation study on CharacterEval. **Bold**: the best results; [‡]: significant degradation compared to SIBPersona ($p < 0.05$).

model specifically fine-tuned on fictional characters’ conversational data, demonstrating that an ICL-based method augmented with behavioral modeling can approach the performance of task-specific fine-tuned models.

The ablation study in Table 8 further corroborates the findings in §5.3. Removing the RP dimension leads to significant degradation across multiple evaluation dimensions, while removing only the internal state yields a smaller performance drop. Removing the T dimension does not lead to a significant performance drop, suggesting that reaction process may play a more important role than stylistic information in fictional character benchmarks. These consistent patterns across two distinct role-playing settings confirm that SIBPersona is broadly applicable beyond the domain for which it was originally designed. Additional results and analysis are provided in Appendix E.4.

6 Conclusion

In this work, we study the problem of role-playing real individuals on social media and propose a behavior-aware framework to address two major challenges. In our framework, we introduce SIBPersona, which extends data-variant persona by modeling the reaction process. We construct a dataset featuring real social media influencers and

formulate the task of generating replies. Our experimental results across multiple models demonstrate that SIBPersona improves role-playing fidelity, and human evaluation further verifies the interpretability of the retrieved SIB triplets. We also find that reaction process modeling is applicable to fictional character role-playing benchmarks. Moreover, our reference-augmented evaluation protocol for role-playing real-individual shows acceptable alignment with human judgment. These findings suggest that behavior-aware persona modeling can improve role-playing performance on social media setting.

In future work, this framework could be extended to broader domains and more complex generation tasks. One such direction is extending SIBPersona to multi-turn dialogue, which would require handling evolving reaction processes. This can be achieved by dynamically formulating retrieval queries from the accumulated conversational history instead of relying solely on the latest comment, enabling the model to fetch situation-relevant SIB triplets that capture accumulated context.

Limitations

Although our study demonstrates effectiveness on both our newly constructed social media dataset and a public open-source dataset, and shows consistent improvements across both large API-based models and smaller open-source models, several limitations remain.

First, we observe that our method struggles with two types of comment situations (Figure 13): real-time information queries and sarcastic comments. In the former case, the LLM lacks up-to-date information, which leads to hallucinated replies. In the latter case, the LLM may misinterpret sarcastic remarks from followers, resulting in inappropriate behavioral strategies.

Second, our current study focuses on single-turn reply generation and does not evaluate performance in more challenging multi-turn dialogue settings. Since conversational contexts may evolve across turns, future work could investigate how the proposed framework performs under dynamically changing interaction scenarios.

Third, the dimensional documents in SIBPersona are currently provided to the LLM through in-context learning rather than model training. We have not explored whether integrating dimensional documents into model training could further im-

prove performance. Future works could investigate training models with dimensional documents and examine whether such models can achieve stronger performance than API-based models.

Ethical Considerations

This work involves the use of social media data and the impersonation of real individuals. We take several measures to address ethical concerns related to data usage, annotation practices, and potential societal impacts.

First, this study was conducted as an industry-academia collaboration between National Taiwan University and CMoney, the company that owns and operates the platform from which the data were collected. All data access and processing were conducted solely by CMoney-affiliated authors within a secure, company-managed environment in full compliance with the platform’s Terms of Service and Privacy Policy. We obtained explicit consent from all influencers whose personas were evaluated in our experiments to use their available data for research purposes. Second, regarding follower comments, platform members accept the Terms of Service and Privacy Policy upon registration, authorizing the platform to collect, process, and use member-provided data for online behavior research, statistics, and service improvements. To protect participant privacy, all follower comments in our dataset and evaluation materials are strictly anonymized, retaining no user identifiers or personally identifiable information. Furthermore, no raw influencer or follower content will be released as part of our research artifacts. Third, human annotators involved in the evaluation are employees of CMoney and are compensated through their regular salaries. Fourth, AI-based writing assistance tools were used to help polish parts of the manuscript. All AI-assisted text was carefully reviewed and proofread by the authors to ensure correctness and accuracy.

Finally, we acknowledge that technologies capable of impersonating real individuals could potentially be misused to generate misleading or deceptive content on social media. This work is an evaluative research study rather than a public deployment. Any future commercialization or deployment would be handled separately by CMoney and is outside the scope of this study; nonetheless, we recommend that any downstream deployment require: (1) personas to be created internally and

only for explicitly authorized influencers, (2) human review of generated content before publication, particularly for sensitive responses, and (3) mechanisms allowing influencers to deactivate their personas or withdraw consent. Meanwhile, our study contributes to understanding the capabilities and limitations of such systems. We hope that these insights can help future research develop methods for detecting impersonation and mitigating the spread of misinformation in online communities.

Acknowledgments

The work was financially supported by a National Taiwan University Industry-Academia Cooperative Research Project, the National Science and Technology Council (NSTC), and the Featured Area Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE), Taiwan, under Grants 112-2223-E-002-012-MY5, 115-2628-E-002-023-MY4, and 115L900901.

We thank Li-Ching Chien, Pei-Yu Hou, Pei-Ru Huang, Si-Rui Huang, Meng-Jung Lin, Yung-Chen Liu, Xiao-An Wang, and Yi-Chi Yeh for serving as human annotators in this work.

References

- Aurélien Bück-Kaeffer, Je Qin Chooi, Dan Zhao, Maximilian Puelma Touzel, Kellin Pelrine, Jean-François Godbout, Reihaneh Rabbany, and Zachary Yang. 2025. [BluePrint: A social media user dataset for LLM persona evaluation and training](#). *arXiv preprint arXiv:2510.02343*.
- Yun-Shiuan Chuang, Ruixuan Tu, Chengtao Dai, You Li, Smit Vasani, Binwei Yao, Michael Henry Tessler, Sijia Yang, Dhavan Shah, Robert Hawkins, Junjie Hu, and Timothy T. Rogers. 2025. [DEBATE: A large-scale benchmark for evaluating opinion dynamics in role-playing LLM agents](#). *arXiv preprint arXiv:2510.25110*.
- Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition. Lawrence Erlbaum Associates, Hillsdale, NJ.
- Zhenpeng Gao, Xiaofen Xing, and Xiangmin Xu. 2025. [TailorRPA: A retrieval-based framework for eliciting personalized and coherent role-playing agents in general domain](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5381–5412, Suzhou, China. Association for Computational Linguistics.
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun

- Zhang, Zhouchi Lin, Bowen Zhang, Lionel Ni, Wen Gao, Yuanzhuo Wang, and Jian Guo. 2026. [A survey on LLM-as-a-judge](#). *The Innovation*, 7(6):101253.
- Kai He, Yucheng Huang, Wenqing Wang, Delong Ran, Dongming Sheng, Junxuan Huang, Qika Lin, Jiaxing Xu, Wenqiang Liu, and Mengling Feng. 2025. [Crab: A novel configurable role-playing LLM with assessing benchmark](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15030–15052, Vienna, Austria. Association for Computational Linguistics.
- Ankur Joshi, Saket Kale, Satish Chandel, and D. Kumar Pal. 2015. [Likert scale: Explored and explained](#). *British journal of applied science & technology*, 7(4):396–403.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive NLP tasks](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.
- Li Li, Peilin Cai, Ryan A. Rossi, Franck Dernoncourt, Branislav Kveton, Junda Wu, Tong Yu, Linxin Song, Tiankai Yang, Yuehan Qin, Nesreen K. Ahmed, Samyadeep Basu, Subhojyoti Mukherjee, Ruiyi Zhang, Zhengmian Hu, Bo Ni, Yuxiao Zhou, Zichao Wang, Yue Huang, and 5 others. 2025. [A personalized conversational benchmark: Towards simulating personalized conversations](#). *arXiv preprint arXiv:2505.14106*.
- 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.
- Jiaheng Liu, Zehao Ni, Haoran Que, Tao Sun, Zekun Wang, Jian Yang, Jiakai Wang, Hongcheng Guo, Zhongyuan Peng, Ge Zhang, Jiayi Tian, Xingyuan Bu, Ke Xu, Wenge Rong, Junran Peng, and Zhaoxiang Zhang. 2024. [RoleAgent: Building, interacting, and benchmarking high-quality role-playing agents from scripts](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 49403–49428.
- Andrey Lovakov and Elena R. Agadullina. 2021. [Empirically derived guidelines for effect size interpretation in social psychology](#). *European Journal of Social Psychology*, 51(3):485–504.
- Mary L. McHugh. 2012. [Interrater reliability: The kappa statistic](#). *Biochemia Medica*, 22(3):276–282.
- Walter Mischel and Yuichi Shoda. 1995. [A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure](#). *Psychological review*, 102(2):246–268.
- Simon Münker, Nils Schwager, and Achim Rettinger. 2026. [Don’t trust generative agents to mimic communication on social networks unless you benchmarked their empirical realism](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1141–1151, Rabat, Morocco. 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.
- Karl Pearson. 1895. [Note on regression and inheritance in the case of two parents](#). *Proceedings of the Royal Society of London*, 58:240–242.
- Ji-Lun Peng and Yun-Nung Chen. 2026. [Rethinking role-playing evaluation: Anonymous benchmarking and a systematic study of personality effects](#). In *Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 205–218, Atlanta, Georgia, USA. Association for Computational Linguistics.
- Vahid Rahimzadeh, Erfan Moosavi Monazzah, Mohammad Taher Pilehvar, and Yadollah Yaghoobzadeh. 2026. [Synthia: Scalable grounded persona generation from social media data](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 45983–46006, San Diego, California, United States. Association for Computational Linguistics.
- Emily Royall and Thomas Wortmann. 2015. [Finding the state space of urban regeneration: Modeling gentrification as a probabilistic process using k-means clustering and markov models](#). In *Proceedings of the 14th International Conference on Computers in Urban Planning and Urban Management (CUPUM 2015): Planning Support Systems and Smart Cities*, Cambridge, MA, USA. Massachusetts Institute of Technology. Paper 275.
- Giovanni Santia and Jake Williams. 2018. [BuzzFace: A news veracity dataset with Facebook user commentary and egos](#). In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, pages 531–540.
- Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. [BLEURT: Learning robust metrics for text generation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.
- 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.
- Quan Shi, Carlos E. Jimenez, Stephen Dong, Brian Seo, Caden Yao, Adam Kelch, and Karthik R. Narasimhan. 2025. **IMPersona: Evaluating individual level LLM impersonation**. In *Second Conference on Language Modeling*.
- C. Spearman. 1904. **The proof and measurement of association between two things**. *The American Journal of Psychology*, 15(1):72–101.
- Quan Tu, Chuanqi Chen, Jinpeng Li, Yanran Li, Shuo Shang, Dongyan Zhao, Ran Wang, and Rui Yan. 2023. **CharacterChat: Learning towards conversational AI with personalized social support**. *arXiv preprint arXiv:2308.10278*.
- Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. 2024. **CharacterEval: A Chinese benchmark for role-playing conversational agent evaluation**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11836–11850, Bangkok, Thailand. Association for Computational Linguistics.
- Huandong Wang, Wenjie Fu, Yingzhou Tang, Zhi-long Chen, Yuxi Huang, Jinghua Piao, Chen Gao, Fengli Xu, Tao Jiang, and Yong Li. 2025a. **A survey on responsible LLMs: Inherent risk, malicious use, and mitigation strategy**. *arXiv preprint arXiv:2501.09431*.
- Lei Wang, Jianxun Lian, Yi Huang, Yanqi Dai, Haoxuan Li, Xu Chen, Xing Xie, and Ji-Rong Wen. 2025b. **CharacterBox: Evaluating the role-playing capabilities of LLMs in text-based virtual worlds**. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6372–6391, Albuquerque, New Mexico. Association for Computational Linguistics.
- Noah Wang, Z.Y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhao Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024. **RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models**. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand. Association for Computational Linguistics.
- Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen-Tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Shuchang Zhou, Wei Wang, and Yanghua Xiao. 2025c. **CoSER: Coordinating LLM-based persona simulation of established roles**. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 64822–64858. PMLR.
- Matthijs J Warrens. 2015. **Five ways to look at Cohen’s kappa**. *Journal of Psychology & Psychotherapy*, 5(4):197.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. **Chain-of-thought prompting elicits reasoning in large language models**. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.
- Yuwei Wu, Xuezhe Ma, and Diyi Yang. 2021. **Personalized response generation via generative split memory network**. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1956–1970, Online. Association for Computational Linguistics.
- Rui Xu, Xintao Wang, Jiangjie Chen, Siyu Yuan, Xinfeng Yuan, Jiaqing Liang, Zulong Chen, Xiaoqingdong, and Yanghua Xiao. 2025. **Character is destiny: Can persona-assigned language models make personal choices?** In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15038–15059, Suzhou, China. Association for Computational Linguistics.
- Zhenhua Xu, Dongsheng Chen, Shuo Wang, Jian Li, Chengjie Wang, Meng Han, and Yabiao Wang. 2026. **AdaMARP: An adaptive multi-agent interaction framework for general immersive role-playing**. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 31230–31283, San Diego, California, United States. Association for Computational Linguistics.
- Shihao Yang, Zhicong Lu, Yong Yang, Bo Lv, Yang Shen, and Nayu Liu. 2026. **HyCoRA: Hyper-contrastive role-adaptive learning for role-playing**. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 34304–34312.
- Jing Ye, Rui Wang, Yuchuan Wu, Victor Ma, Feiteng Fang, Fei Huang, and Yongbin Li. 2025. **CPO: Addressing reward ambiguity in role-playing dialogue via comparative policy optimization**. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 297–323, Suzhou, China. Association for Computational Linguistics.
- Yeyong Yu, Runsheng Yu, Haojie Wei, Zhanqiu Zhang, and Quan Qian. 2025. **Beyond dialogue: A profile-dialogue alignment framework towards general role-playing language model**. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11992–12022, Vienna, Austria. Association for Computational Linguistics.
- Haonan Zhang, Run Luo, Xiong Liu, Yuchuan Wu, Ting-En Lin, Pengpeng Zeng, Qiang Qu, Feiteng Fang, Min Yang, Lianli Gao, Jingkuan Song, Fei Huang,

and Yongbin Li. 2025. *OmniCharacter: Towards immersive role-playing agents with seamless speech-language personality interaction*. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26318–26331, Vienna, Austria. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. *Personalizing dialogue agents: I have a dog, do you have pets too?* In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. *BERTScore: Evaluating text generation with BERT*. In *International Conference on Learning Representations*.

Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2019. *Personalized dialogue generation with diversified traits*. *arXiv preprint arXiv:1901.09672*.

Jinfeng Zhou, Yongkang Huang, Bosi Wen, Guanqun Bi, Yuxuan Chen, Pei Ke, Zhuang Chen, Xiyao Xiao, Libiao Peng, Kuntian Tang, Rongsheng Zhang, Le Zhang, Tangjie Lv, Zhipeng Hu, Hongning Wang, and Minlie Huang. 2025. *CharacterBench: Benchmarking character customization of large language models*. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26101–26110.

A Detail of Constructing Dataset

In this work, we construct the dataset from social media data related to the impersonated influencers, collected internally under the authorization of the platform operator. To ensure sufficient data coverage, we select influencers who have more than 500 comment–reply pairs in the demonstration source and more than 120 comment–reply pairs in the candidate pool for testing. Table 22 reports the dataset statistics for the six influencers studied in our work.

Furthermore, based on our preliminary analysis, comment–reply pairs exhibit highly distinct interaction types. Some pairs mainly involve factual information exchange, while others more clearly reveal the influencer’s speaking style or behavioral strategies. To ensure that the test samples comprehensively covers these diverse interactions, we prompt gpt-4o-mini to assign each comment–reply pair a persona score ranging from 0 to 1, indicating the degree to which the reply exhibits the influencer’s personal characteristic. We then use this

Persona Score	Influencer					
	A	B	C	D	E	F
High	50	50	50	35	50	9
Medium	50	50	50	50	22	39
Low	50	50	50	50	12	50

Table 9: Number of data in the test samples.

score as the basis for stratified sampling. Specifically, comment–reply pairs are categorized into three levels according to their persona scores: high [0.8, 1.0], medium [0.5, 0.8), and low [0, 0.5). Examples of the three levels are shown in Figure 7. Low-level pairs mainly involve factual or experiential information exchange, medium-level pairs exhibit stylistic cues such as emoji usage, and high-level pairs clearly reflect the influencer’s personal characteristic. The prompt used to obtain persona scores is shown in Figure 14. The statistics of the test samples for each of the six influencers are summarized in Table 9. For *Influencer D*, all available data in the medium and high levels are included because each level contains fewer than 50 test samples. The same logic applies to *Influencer E* and *Influencer F*.

To further demonstrate the diversity of our dataset, Table 10 reports the text length distributions and the proportions of the three persona-score levels across all six influencers. As shown in the table, textual lengths vary considerably across individuals: average reply lengths range from 8.46 to 33.65 characters, while average post lengths range from 271.65 to 952.74 characters. Moreover, the distribution of persona-score levels in the candidate pool reflects distinct individual interaction styles. For instance, *Influencer E*’s comment–reply pairs predominantly involve low persona-characteristic exchanges (87%), whereas *Influencer C* displays a much more balanced distribution across medium (36%) and low (53%) levels. These statistics confirm that our dataset encompasses a wide spectrum of behavioral patterns.

B Detail of SIBPersona

B.1 Data-invariant Persona Construction

For the data-invariant persona, we construct a short profile for each influencer using the influencer description provided on the social media, rather than deriving it from historical data. Each profile contains approximately 50–100 words.

Influencer	Post Length	Comment Length	Reply Length	High (%)	Med (%)	Low (%)
<i>A</i>	467.92 $\pm$ 471.09	17.58 $\pm$ 25.37	8.46 $\pm$ 9.69	13%	29%	58%
<i>B</i>	367.39 $\pm$ 231.62	30.70 $\pm$ 57.55	10.01 $\pm$ 10.54	12%	35%	53%
<i>C</i>	952.74 $\pm$ 740.92	66.83 $\pm$ 105.54	33.65 $\pm$ 28.14	34%	36%	53%
<i>D</i>	355.21 $\pm$ 508.89	14.20 $\pm$ 15.24	9.56 $\pm$ 10.17	16%	25%	59%
<i>E</i>	271.65 $\pm$ 221.00	27.54 $\pm$ 28.06	9.52 $\pm$ 10.21	5%	8%	87%
<i>F</i>	860.21 $\pm$ 574.41	26.53 $\pm$ 29.41	11.82 $\pm$ 10.39	8%	33%	60%

Table 10: Dataset diversity statistics across the six influencers, including text lengths and the proportion of comment-reply pairs across high, medium, and low persona-score levels.

B.2 Data-variant Persona Construction

Table 11 reports the size of the dimensional corpora extracted for *Influencer A* in four LLMs, while Table 12 and Table 13 further present the corresponding statistics for the remaining five influencers using DeepSeek-v3.2 and gemini-3-flash-preview, respectively. All models are accessed through cloud-based APIs: we access claude-opus-4-6 (pinned snapshot) through the Claude API, gemini-3-flash-preview (December 2025 snapshot) through Google Cloud Platform (GCP), and both gpt-5.2-2025-12-11 and DeepSeek-v3.2 through Azure AI Foundry. For decoding parameters across tasks, we set (temperature, top- p) to (1.0, 1.0) for recognizing, (0.2, 0.8) for extraction, and (0.8, 1.0) for generation. All other API parameters remain at their default values, with the reasoning effort set to the lowest level supported by each model.

B.3 Distinctiveness of Persona Dimensions

Our dimension-wise retrieval design is based on the hypothesis that the three dimensions of data-variant persona capture largely distinct aspects of an individual’s persona. While these dimensions may still exhibit certain correlations, we hypothesize that memory, speaking style, and reaction process encode sufficiently different types of information to be treated separately during retrieval. Under this hypothesis, the follower’s comment is matched separately against each dimensional corpus, which prevents high-frequency patterns in one dimension from dominating the retrieval results of the others.

To empirically examine this hypothesis, we analyze the distinctiveness of the three dimensions in the embedding space. Specifically, we embed $m_{p_{jk}}$ for memory, $sp_{p_{jk}}$ for speaking style, and behavior b_{jk} for the reaction process. We use b_{jk} rather

Models	M	T	RP
GPT-5.2	6,064	1,410	970
Gemini-3-flash	5,755	1,050	861
Deepseek-V3.2	5,143	1,022	749
Claude-Opus-4.6	2,307	407	703

Table 11: Number of dimensional documents of *Influencer A*.

than the full SIB triplet because behavior, similar to memory and speaking style, directly reflects the impersonated individual’s observable reply patterns, whereas situation and internal state mainly describe the intermediate reasoning process that leads to the behavior. The resulting vectors are then analyzed through qualitative visualization and quantitative clustering evaluation. The results provide empirical evidence that the three dimensions are separable, supporting our decision to process them independently during retrieval.

For the qualitative analysis, Figure 3 shows the t-SNE visualization of the embeddings of $m_{p_{jk}}$, $sp_{p_{jk}}$, and b_{jk} extracted for *Influencer A* by Claude Opus 4.6. As illustrated in the figure, the three dimensions occupy different regions in the projected space, suggesting that they capture different aspects of the impersonated individual. For the quantitative analysis, we further calculate the Silhouette Score across six influencers, using documents extracted by DeepSeek-v3.2. The resulting average Silhouette Score is 0.37. As a random baseline, we permute the cluster label of all document embeddings and repeat this process 1,000 times. The average Silhouette Score under this random permutation setting is -0.001 . The substantial gap between the actual score and the random baseline provides evidence that the three dimensions form distinguishable clusters in the embedding space.

Influencer	M	T	RP
B	12,358	8,162	6,701
C	6,179	2,982	4,049
D	8,914	4,935	3,811
E	10,087	2,393	2,378
F	2,587	446	398

Table 12: Number of dimensional documents of *Influencer B-F* extracted from Deepseek V3.2.

Influencer	M	T	RP
B	13360	8253	6911
C	5296	3321	3643
D	10102	4839	4053
E	10309	3536	3287
F	2363	515	449

Table 13: Number of dimensional documents of *Influencer B-F* extracted from Gemini-3-flash.

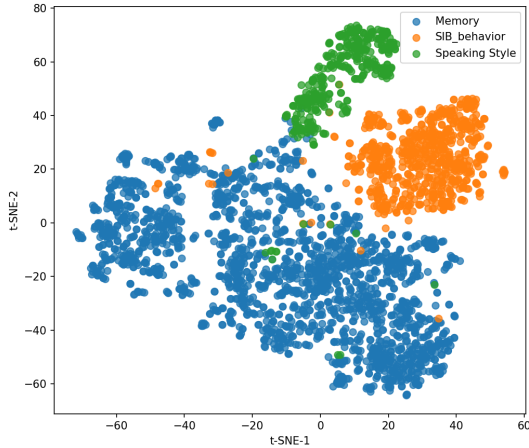


Figure 3: t-SNE projection of the embeddings of mp_{jk} , sp_{jk} , and b_{jk} generated by Claude-Opus-4.6.

C Details of Reference-Augmented Evaluation

We adopt the penalty-based evaluation protocol introduced by Wang et al. (2025c) to score RPA-generated replies. We use gpt-4.1-2025-04-14 as LLM evaluator and access it through the Azure AI Foundry. The temperature is set to 0.2 and top- p to 0.8 to ensure the stability of the evaluation results.

Number of Representative Dimensional Documents To avoid using the information from the same dimensional corpora for RPA generation and evaluation, we select the representative dimen-

sional documents from the dimensional corpora constructed from the candidate pool for test samples. Following the rule of thumb discussed by Royall and Wortmann (2015), we set the number of medoids k according to $k = \sqrt{M/2}$, where M denotes the number of dimensional documents in the corresponding corpus. We then round k up to the nearest integer and use it as the number of representative dimensional documents. Table 23 reports, for each influencer and extracting model, the total number of extracted dimensional documents and the corresponding number of representative dimensional documents used in evaluation.

Rubrics for LLM Evaluators Evaluations are conducted along three dimensions: memory, speaking style, and reaction process. Each dimension is independently evaluated in a separate run. For each evaluation run, the LLM evaluator is provided with the definition of the target dimension and its corresponding error types as dimension-specific rubrics. Using the data-invariant profile and representative dimensional documents as references, the LLM evaluator identifies flaws in the generated reply and assigns a severity score to each detected flaw. This design allows the LLM evaluator to focus on one dimension at a time and avoids penalizing errors from other dimensions. The complete evaluation prompt and evaluation rubrics are shown in Figure 19 and Figure 20, respectively.

Details of the Formula In the original formulation by Wang et al. (2025c), the evaluation score includes a length-based compensation term in addition to the penalty term, since longer replies are more likely to receive penalties under error-based scoring schemes. However, in our setting, the generated replies are typically short, with an average length of approximately 10 words. We therefore omit the length-based score adjustment in our evaluation.

Visualization for Representative Dimensional Documents We visualize the representative dimensional documents extracted for *Influencer A* by DeepSeek-v3.2. As shown in Figure 4, Figure 5, and Figure 6, the selected representative dimensional documents are spread across different regions of the projected space rather than being concentrated in a single local cluster. This qualitative observation suggests that representative dimensional documents capture diverse aspects of the impersonated influencer across the three dimen-

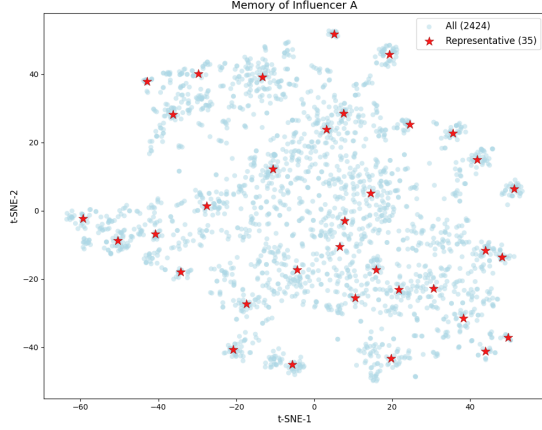


Figure 4: t-SNE projection of representative memory documents generated by DeepSeek-v3.2.

sions.

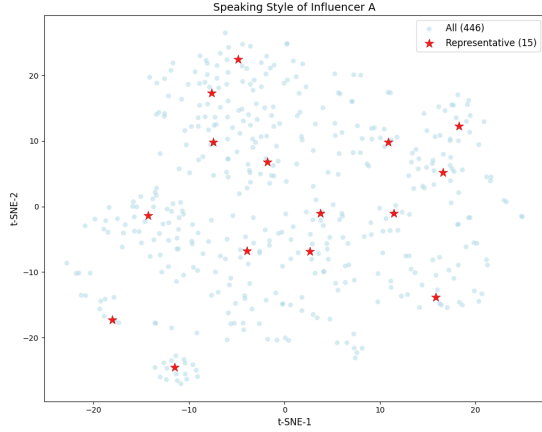


Figure 5: t-SNE projection of representative speaking style documents generated by DeepSeek-v3.2.

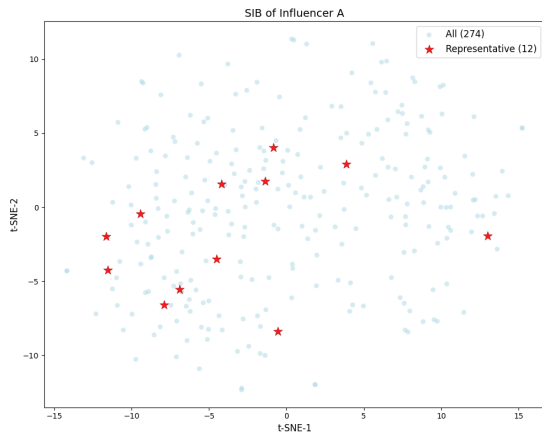


Figure 6: t-SNE projection of representative reaction process documents generated by DeepSeek-v3.2.

D Details of Human Evaluation

Human evaluations are conducted with two main objectives. The first objective is to assess the alignment between reference-augmented evaluation protocol and human judgment. For this purpose, we use the comparison evaluation to support annotator quality control and the performance evaluation to measure the correlation between human judgment and LLM evaluators. The second objective is to examine the validity of the reaction-process-based design of SIBPersona. Specifically, the interpretability evaluation assesses whether the retrieved SIB triplets can serve as explanations for generated replies, and the construction faithfulness evaluation verifies whether the constructed SIB triplets are faithful to their corresponding evidence pairs. The following subsections first describe the settings and annotation protocols of the four human evaluations in Appendix D.1 and then present the corresponding results in Appendix D.2.

D.1 Human Evaluation Setting

This subsection describes the setting of our human evaluation. In the following, we refer to each question in the human evaluation inventory as an evaluation item.

D.1.1 Comparison Evaluation

The comparison evaluation is designed to serve two complementary purposes. First, it provides a post hoc diagnostic for identifying potentially unreliable annotators. Second, it offers a comparison-based assessment of the quality of RPA-generated replies. For each evaluation item, annotators are given an influencer’s post ip , a follower’s comment fc , and three anonymized candidate replies. They are asked to rank the three replies according to how closely each reply resembles the impersonated influencer’s actual reply, with ties allowed when multiple replies are considered equally similar.

Although all three options are presented as generated replies, only two of them are produced by RPA: one by IMPersona and the other by SIBPersona. The remaining option is the ground-truth reply written by the influencer. Since the ground-truth reply is written by impersonated individual, a reliable annotator is expected not to consistently rank it as the least similar reply. We therefore use the hidden ground-truth reply as a quality-control signal. Specifically, for each annotator, we compute the proportion of evaluation items in which

the ground-truth reply is ranked last. Annotators whose ground-truth-last rate exceeds a predefined threshold are excluded from subsequent analysis; in our work, the threshold is set to 30%.

D.1.2 Performance Evaluation

The performance evaluation is designed to assess the alignment between our proposed evaluation protocol and human judgment by measuring the correlation between their scores. In addition, it allows us to compare the quality of RPA-generated replies through statistical significance testing. For each evaluation item, annotators are given an influencer’s post, a follower’s comment, and an RPA-generated reply generated by IMPersona, Vanilla RAG or SIBPersona. They are asked to rate the reply along three dimensions: memory, speaking style, and reaction process, using a five-point Likert scale (Joshi et al., 2015). If an annotator considers a reply irrelevant to a given dimension, they may select *NA*. This option is included to avoid forcing annotators to assign scores to dimensions that are not applicable to a particular reply.

To address *NA* annotations during analysis, we process human evaluation scores as follows. For each evaluation item and each dimension, we aggregate scores across annotators by averaging all non-*NA* ratings. This aggregation yields one human evaluation score for each dimension of each sample and helps reduce the influence of annotator-specific variation, allowing the subsequent correlation analysis to focus on aggregate human evaluation trends. If all annotators select *NA* for an evaluation item on a given dimension, the aggregated score for that dimension is treated as *NA*. Such cases are excluded from analysis involving the corresponding dimension because no valid human evaluation score is available.

After aggregating the human evaluation scores, we perform two complementary analysis. The first analysis examines the alignment between LLM evaluator and human judgment using correlation measures. The second analysis compares the quality of RPA-generated replies across IMPersona, Vanilla RAG and SIBPersona based on human evaluation scores.

For the correlation analysis, we compute both dimension-specific and overall correlations between LLM evaluator and human judgment scores. Dimension-specific correlation is calculated separately for memory, speaking style, and reaction process. For each dimension, evaluation items with

NA in the corresponding aggregated human judgment score are excluded since no valid human score is available for that dimension. This analysis allows us to examine whether the LLM evaluator aligns with human judgment on each specific dimension.

In addition, we compute an overall correlation by first averaging the scores across the three dimensions for each evaluation item and then calculating the correlation based on these overall scores. Compared with the dimension-specific correlation, this analysis provides an overall view of the alignment between LLM evaluator and human judgment. For this analysis, evaluation items are excluded if any of the three dimension scores is *NA*, because the overall score cannot be consistently computed across all dimensions.

For the method-level quality comparison, we focus on the three RPA methods: IMPersona, Vanilla RAG and SIBPersona. For each evaluation dimension, we compute the average human evaluation score of each method and conduct *t*-tests with Holm correction to examine whether the score differences are statistically significant. This analysis complements the correlation analysis by directly comparing the human-evaluated quality of the three methods.

D.1.3 Interpretability Evaluation

After evaluating the quality of the RPA-generated replies, we further conduct evaluations to examine whether the constructed reaction process corpus provides meaningful explanatory evidence for SIBPersona. We first perform an interpretability evaluation to assess whether the retrieved SIB triplets can serve as plausible explanations for the generated replies. For each evaluation item, annotators are given an influencer’s post, a follower’s comment, and a reply generated by SIBPersona, together with the top-three SIB triplets retrieved and provided to the RPA.

Annotators are asked to evaluate each SIB triplet from two aspects. First, they assess whether the situation is relevant to the influencer’s post and the follower’s comment. Second, they assess whether the internal state and behavior can reasonably explain the generated reply. This design allows us to examine both the contextual relevance of the retrieved situations and the explanatory consistency between the retrieved SIB triplets and the final generated reply. Ratings are given using a five-point Likert scale.

D.1.4 Construction Faithfulness Evaluation

Finally, we conduct a construction faithfulness evaluation to assess whether the generated SIB triplets are reasonably extracted from the original data sample. This evaluation examines whether each extracted SIB triplet can be reasonably supported by the comment–reply pair assigned as its evidence. For each evaluation item, annotators are given a comment–reply pair and its corresponding extracted SIB triplet. They are then asked to the annotator whether it is reasonable to construct the given SIB triplet from the provided comment–reply pair. This setup allows us to directly evaluate the faithfulness of the corpus construction process, rather than only assessing their groundedness in downstream reply generation. The evaluation is collected as a binary pass/fail label, where *pass* indicates that the triplet is reasonably supported by the source comment–reply pair, and *fail* indicates that the triplet is unsupported, inconsistent with the comment–reply pair, or overly inferred from the source data.

D.1.5 Implementation Details

All human evaluations are conducted using test samples generated by DeepSeek-V3.2 for *Influencer F*. For the performance evaluation, we use 30 test samples and evaluate replies generated by three methods, Vanilla RAG, IMPersona and SIBPersona, resulting in 90 evaluation items in total. For the interpretability evaluation, we use the same 30 test samples. Since each test sample contains top-three retrieved SIB triplets, annotators evaluate each triplet separately.

The 30 test samples are manually selected from the three levels introduced in §5.1 to ensure diversity in comment situations. For the comparison evaluation, we randomly select 10 samples from these 30 test samples. For the construction faithfulness evaluation, we use 50 representative dimensional documents from reaction process corpus to assess whether the extracted SIB triplets are faithful to their corresponding evidence pairs.

We recruit nine annotators for the human evaluation. All annotators have worked with *Influencer F* for several years, ensuring that they have sufficient familiarity with the impersonated influencer. To control annotation quality, we further filter annotators based on the ground-truth-last rate calculated from the comparison evaluation. Since only three annotators pass the filtering threshold, we report the results of these three annotators in the following

Methods	Reciprocal Ranking score
IMPersona	0.42 [‡]
SIBPersona	0.71
Ground Truth	0.71

Table 14: Human comparison evaluation reciprocal scores. **Bold** indicates the best score. [‡]: significant degradation compared to SIBPersona ($p < 0.05$)

subsection.

D.2 Results of Human Evaluation

Results of Comparison Evaluation For the comparison evaluation, we convert the rankings of candidate replies into reciprocal ranking scores, where the score of each reply is computed as the reciprocal of its assigned rank. We then report the average reciprocal ranking score for each method, with higher scores indicating better rankings. As shown in Table 14, SIBPersona achieves a significantly higher score than IMPersona and obtains an average score comparable to that of the ground-truth replies. This result suggests that annotators judged the replies generated by SIBPersona to be substantially closer to the replies of real influencer’s than those generated by IMPersona.

Results of Performance Evaluation To examine the alignment between human judgment and LLM evaluators, we compare the aggregated human judgment scores with the LLM evaluator scores and report their correlations. We report both Pearson’s r (Pearson, 1895) and Spearman’s ρ (Spearman, 1904) to measure linear and rank-based associations, respectively. As shown in Figure 8, the overall human judgment score shows a moderate correlation with the overall LLM evaluator score, yielding Pearson’s r of 0.561 and Spearman’s ρ of 0.530. This suggests that the LLM evaluator captures evaluation trends that are broadly aligned with human judgment when the three dimensions are considered jointly. Among the dimension-specific correlations, speaking style shows the strongest alignment between human judgment and LLM evaluator, with Pearson’s r of 0.576 and Spearman’s ρ of 0.561 indicating a moderate correlation. Memory and reaction process also show significant but weaker correlations, with Pearson’s r of 0.354 and 0.289, and Spearman’s ρ of 0.365 and 0.287, respectively. These results are consistent with the previous findings in role-playing evaluation (Zhou et al., 2025;

	M	T	RP
Rater 1 & 2	0.364	0.429	0.406
Rater 1 & 3	0.417	0.506	0.508
Rater 2 & 3	0.426	0.495	0.487

Table 15: Inter-rater agreement among the three annotators.

He et al., 2025) and demonstrate that our reference-augmented protocol serves as a reasonable proxy for human judgment.

Moreover, we further assess inter-rater agreement among human annotators, computed by the weighted Cohen’s κ (Warrens, 2015). The agreement scores across the three dimensions are presented in Table 15. The pairwise agreement scores range from 0.364 to 0.508. Following the interpretation guidelines of McHugh (2012), these results indicate a fair to moderate level of agreement across annotators. The agreement is highest between Rater 1 and Rater 3, while Rater 1 and Rater 2 show relatively lower agreement. This pattern suggests that the annotation task involves subjective judgment, but the annotators still exhibit meaningful consistency.

For the method-level quality comparison, as shown in Table 5, SIBPersona significantly outperforms IMPersona across all three dimensions.

Results of Interpretability Evaluation For the interpretability evaluation, each test sample is associated with three retrieved SIB triplets. Since the RPA may rely more heavily on one of the retrieved triplets during generation, we take the highest score among the three triplets as the score for that test sample. Specifically, let $score_{i,j}^S$ and $score_{i,j}^{IB}$ denote the average annotator scores for the j -th SIB triplet of the i -th test sample, where $j \in \{1, 2, 3\}$. The final scores, denoted as $Score_S$ and $Score_{IB}$, are computed by averaging the highest triplet-level score across all N test samples:

$$Score_S = \frac{1}{N} \sum_{i=1}^N \max_{j \in \{1,2,3\}} score_{i,j}^S,$$

$$Score_{IB} = \frac{1}{N} \sum_{i=1}^N \max_{j \in \{1,2,3\}} score_{i,j}^{IB}.$$

Our RPA replies achieve a $Score_S$ of 3.63 and a $Score_{IB}$ of 3.57. The high $Score_S$ indicates that

	Pass Rate	95% CI
Rater 1	0.80	[0.67, 0.89]
Rater 2	0.84	[0.72, 0.92]
Rater 3	0.92	[0.81, 0.97]
Total	0.94	[0.84, 0.98]

Table 16: Extraction faithfulness evaluated by the three annotators. The **Total** column reports the pass rate of by applying majority-vote across all annotators.

the situation described in the retrieved SIB triplet generally align with the corresponding follower’s comment. The high $Score_{IB}$ further suggests that the internal states and behaviors described in the triplets can reasonably explain the final replies generated by the RPA.

In addition, we use $Score_{IB}$ as the criterion for categorizing test samples into interpretable and uninterpretable, as shown in Table 6. For each test sample, if the average $Score_{IB}$ assigned by the three annotators exceeds 3, the test sample is considered interpretable; otherwise, it is considered uninterpretable. These results suggest that our retrieval method can identify situations similar to an unseen follower’s comment, and the retrieved internal states and behaviors serve as a valid supporting context in the RPA-generated replies.

Results of Extraction Faithfulness Evaluation

To evaluate the faithfulness of the reaction process corpus, we report the pass rates assigned by the three annotators in Table 16. The individual annotator pass rates are consistently high, with all raters assigning pass rates above 0.80. When applying majority voting across annotators, the overall pass rate reaches 0.94 with a 95% confidence interval of [0.84, 0.98].

These results suggest that most constructed SIB triplets are reasonably supported by their corresponding comment-reply pair. In other words, our process of constructing dimensional corpora can generally preserve information from the original comment-reply pairs while converting them into structured SIB triplets. The higher majority-vote pass rate further indicates that although the annotators may differ in their strictness, there is strong aggregate support for the faithfulness of the constructed reaction process corpus.

E Additional Results

E.1 Per-Influencer Results

Table 24 presents the detailed results for *Influencer B* through *Influencer F*, which correspond to the aggregated results reported in Table 3. The table provides a more fine-grained view of our experimental results in §5.3. Overall, our method consistently outperforms the baseline in both the speaking style dimension and reaction process dimension.

E.2 Comparison with Traditional NLP Metrics

To further justify the use of our reference-augmented evaluation protocol, we compare its alignment with human judgment against several traditional automatic evaluation metrics, including BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), BERTScore (Zhang et al., 2020), and BLEURT (Sellam et al., 2020). Specifically, we compute the Spearman correlation between each metric and human judgment. The results show that traditional metrics exhibit substantially weaker correlations with human judgment: BLEU achieves a Spearman’s ρ of -0.01 , ROUGE-L achieves -0.04 , BERTScore achieves 0.28 , and BLEURT achieves 0.02 . In contrast, our reference-augmented evaluation protocol achieves a Spearman’s ρ of 0.530 , showing a much stronger alignment with human judgment.

This gap suggests that traditional automatic metrics are not sufficient for evaluating role-playing real-individual on social media. One possible reason is that a generated reply may differ substantially from the ground-truth reply at the lexical or surface-semantic level, while still adopting a behavioral strategy that is consistent with the impersonated individual. For example, as shown in Figure 9, the RPA-generated reply follows the influencer’s behavioral strategy and is therefore rated highly by both the LLM evaluator and human evaluators. However, because this reply has low lexical and surface-semantic similarity to the ground-truth reply, it receives low scores from traditional automatic metrics.

These findings support our decision to report reference-augmented evaluation as the main automatic evaluation protocol in this work, as it provides a more fine-grained assessment of whether the generated replies align with the impersonated individuals.

Method	# Tokens
Vanilla RAG	593.51 ± 456.37
IMPersona	1972.64 ± 816.14
SIBPersona	1630.83 ± 812.02

Table 17: Average number of tokens inserted into the prompt per test query across six influencers using DeepSeek-V3.2.

E.3 Token Budgets

Table 17 reports the average number of additional context tokens inserted into the prompt per query, evaluated across the six influencers using DeepSeek-V3.2. As shown in the results, SIBPersona consumes fewer tokens than IMPersona while achieving superior role-playing performance. This indicates that performance gains are not simply driven by increasing context length, but rather rely on supplying structured, behavior-aware, and effective persona information.

E.4 Result of RoleAgentBench

We further evaluate our method on an additional open-source benchmark, RoleAgentBench (Liu et al., 2024), a bilingual (Chinese–English) role-playing of fictional character benchmark. To examine whether SIBPersona is applicable beyond the social media setting, we select two response generation tasks from the benchmark: **General Response** and **Summary**. In the experiments, gpt-4o-mini is used to construct the dimensional corpus, OLMo-3-7B-Instruct is used as the generation model, and gemini-2.0-flash serves as the LLM evaluator. During experimentation, we follow Peng and Chen (2026) and anonymize the character names in the dataset to mitigate the influence of the model’s prior knowledge during pre-training.

For evaluation, we use two complementary protocols: our proposed reference-augmented evaluation protocol and the original evaluation protocol provided by the benchmark. Under the benchmark protocol, responses are evaluated in a pairwise comparison setting. Specifically, its LLM evaluator is presented with two responses generated by SIBPersona as well as IMPersona and asked to select the one that better resembles the target character. To mitigate potential positional bias caused by response ordering in the prompt, we conduct two evaluation rounds with the order of the two responses swapped. A method is counted as the winner only if its response is selected in both rounds,

while it is counted as losing only if the competing method is selected in both rounds. Cases with inconsistent outcomes across the two rounds are treated as ties.

The experimental results (Table 25) show that our method significantly outperforms the baseline under both evaluation protocols, consistent with the findings of our main experiment. This suggests that SIBPersona is also applicable to fictional character benchmarks. For completeness, we also provide the prompts used in the benchmark setting. Figure 15 shows the prompt for dimensional relevance filtering, while Figure 16, Figure 17, and Figure 18 present the prompts used for dimensional document extraction. The prompt used for the penalty-based LLM evaluation is shown in Figure 19, with the corresponding evaluation rubrics provided in Figure 20, and Figure 21 presents the prompt used for response generation.

F Validation of Reference-Augmented Evaluation

Automated evaluation for role-playing real individuals remains a challenging and underexplored problem due to the lack of static gold-standard personas. To validate that our reference-augmented evaluation protocol provides a reasonable assessment of role-playing fidelity, we conduct empirical investigations addressing potential issues.

F.1 Alignment with Human Judgment

A central objective of an automated evaluation protocol is to accurately reflect human perception. To justify our design choice of using SIBPersona-structured documents from the test candidate pool as references, we compare the correlation between LLM evaluator scores and human judgement under three different reference configurations. As reported in Table 20, our proposed configuration achieves the highest alignment with human judgement, outperforming raw comment–reply pairs from the same pool as well as structured documents extracted from the demonstration source. These results confirm that structured SIBPersona representations best approximate human evaluative criteria. Furthermore, extracting references from the demonstration source—the exact corpus observed during generation—tends to reward surface context-matching rather than true behavioral fidelity, which explains its lower correlation with human judgment.

Method	M	T	RP
Naive	95.89	46.43 [‡]	70.85 [‡]
CoT	96.58	45.73 [‡]	72.33 [‡]
IMPersona	95.94	48.10 [‡]	73.20 [‡]
Vanilla RAG	95.32	54.12 [‡]	73.68 [‡]
SIBPersona	95.46	59.22	77.81

Table 18: Evaluation results on *Influencer A* averaged over four models after excluding overlapping test instances. ‡: significant degradation compared to SIBPersona ($p < 0.05$).

F.2 Robustness Against Test Sample Overlap

We further examine whether gold reply leakage occurs when representative reference documents are sampled from the test pool. Because K -medoids selects only $\lceil \sqrt{N/2} \rceil$ cluster medoids from N candidate instances to ensure semantic diversity rather than exhaustive coverage, the empirical overlap between representative documents and test samples is exceptionally low: 0.35% for Memory, 4.01% for Speaking Style, and 4.68% for Reaction Process on *Influencer A*. To strictly rule out any leakage bias, we re-evaluate all baseline models after excluding all overlapping instances. As shown in Table 18, the performance rankings and relative improvements remain identical to our primary findings, confirming that the minimal overlap does not affect evaluation outcomes.

F.3 Evaluation Under Strictly Independent Demonstration Sources

To completely eliminate the risk of test pool contamination and assess the generalizability of our protocol, we conduct a conservative check where references are constructed entirely from the independent demonstration source. Table 19 and Table 21 report the results. Across all tested models and influencers, SIBPersona consistently outperforms all baselines on both T and RP dimensions. The persistent advantages across diverse reference configurations—test candidate pool and demonstration source—demonstrate that our evaluation protocol and empirical conclusions are robust and not artifacts of specific reference construction strategies.

G Examples and Case Study

This section presents qualitative analysis of the retrieved dimensional documents in Appendix G.1,

Methods	M	T	RP
Naive	97.28	46.48 [‡]	77.33 [‡]
CoT	97.61	47.93 [‡]	78.08 [‡]
IMPersona	97.14	48.63 [‡]	78.80 [‡]
Vanilla RAG	97.02	58.93 [‡]	79.18 [‡]
SIBPersona	96.69	62.37	81.52

Table 19: Results on *Influencer A* averaged over four models using evaluator references extracted from the demonstration source. [‡]: significant degradation compared to SIBPersona ($p < 0.05$).

representative success and failure reply cases in Appendix G.2, and an example showing how SIBPersona supports error tracing and correction in Appendix G.3.

G.1 Examples of Retrieved Dimensional Document

Figure 10 and Figure 11 present examples of the top-1 retrieved dimensional documents. Given a follower’s comment that praises the influencer’s appearance, the retrieved situation closely matches the comment context. The generated reply also reflects the retrieved speaking style and behavioral strategy, avoiding a direct acceptance of the compliment. In contrast, the memory dimension retrieves less relevant information, which is expected because factual knowledge is less central to this type of comment. Annotators familiar with *Influencer A* further confirm that the influencer typically does not directly accept compliments about their appearance. This case suggests that our method can retrieve information of situation-relevant reaction process and produce a reply consistent with the influencer’s characteristics.

G.2 Examples of Successes and Failure Cases

In addition, Figure 12 illustrates that our method achieves better role-playing fidelity than the baseline methods. Across the two examples, the ground-truth replies are relatively concise and reflect the influencer’s characteristic way of reacting to followers: giving a brief acknowledgment in reply to praise and using a sharp, indirect reply when challenged. SIBPersona better preserves these pragmatic reply patterns. For example, when the follower praises the influencer, our method generates a short positive reaction that is close to the ground-truth reply. When the follower comments sarcastically with hindsight, our method returns with a

sarcastic counter-reaction, which better matches the interactional intent of the ground truth.

In contrast, the baseline methods often deviate from the impersonated individual’s reply pattern in different ways. Naive and CoT tend to introduce excessive explanatory content, such as trading strategies, app usage details, or additional contextual assumptions that are not present in the ground-truth replies. Similarly, IMPersona frequently generates overly detailed and elaborated replies, suggesting that hierarchical memory alone is not enough for modeling behavioral strategies. Vanilla RAG produces shorter replies, but its replies are often generic and fail to capture the specific behavioral strategy required by the comment context.

In contrast, Figure 13 highlights the limitations of our approach. When the comment involves real-time information, our method may misinterpret the situation due to insufficient knowledge of the current context. In addition, for sarcastic comments, the embedding-based retrieval process may fail to capture the follower’s ironic intent. This can lead to the retrieval of SIB triplets that are mismatched with the actual situation, causing the generated reply to follow an inappropriate behavioral strategy. These cases suggest that our method remains sensitive to the quality of situation understanding and retrieval. This limitation is especially pronounced when the input requires real-time knowledge or reasoning beyond surface-level semantic similarity.

G.3 Tracing and Correcting Behavioral Errors

Even when the system makes mistakes, SIBPersona allows us to effectively trace the source of the error and adjust RPA properly. For instance, we observed a case where *Influencer A* replied to a fan’s generic "Thank you" with a highly user-specific follow-up question: "Has that matter been taken care of?". After investigation, we found that the LLM initially yields a reaction process document with this outlier behavior, even though such a reply would be inappropriate for other users. A key advantage of SIBPersona is adjustable: because the generated replies can be traced back to specific dimensional documents. By simply deleting the specific outlier document from the dimensional corpus, we can immediately and precisely adjust the model’s behavior.

Reference Format	Source of Reference	Pearson	Spearman
Raw comment-reply pairs	Test candidate pool	0.355	0.409
SIBPersona document	Demonstration source	0.448	0.462
SIBPersona document	Test candidate pool (Ours)	0.561	0.530

Table 20: Correlation between different evaluator configurations and human judgments.

Methods	Deepseek-V3.2			Gemini-3-flash		
	<i>M</i>	<i>T</i>	<i>RP</i>	<i>M</i>	<i>T</i>	<i>RP</i>
Naive	91.34 $\pm$ 3.24	42.63 [‡] $\pm$ 9.19	70.83 [‡] $\pm$ 5.49	92.18 $\pm$ 2.83	44.21 [‡] $\pm$ 10.84	68.91 [‡] $\pm$ 7.71
CoT	91.06 $\pm$ 3.35	46.65 [‡] $\pm$ 7.36	71.56 [‡] $\pm$ 4.79	92.14 $\pm$ 1.71	45.57 [‡] $\pm$ 8.63	68.53 [‡] $\pm$ 7.39
IMPersona	91.43 $\pm$ 2.61	48.44 [‡] $\pm$ 5.61	73.04 [‡] $\pm$ 5.20	91.34 $\pm$ 2.36	46.32 [‡] $\pm$ 8.20	69.88 [‡] $\pm$ 7.94
Vanilla RAG	91.30 $\pm$ 3.24	59.18 [‡] $\pm$ 4.58	74.30 [‡] $\pm$ 4.26	92.37 $\pm$ 2.60	56.47 [‡] $\pm$ 3.07	72.49 [‡] $\pm$ 3.97
SIBPersona	92.01 $\pm$ 3.53	62.96 $\pm$ 4.27	77.18 $\pm$ 4.44	92.29 $\pm$ 2.48	59.48 $\pm$ 3.00	75.60 $\pm$ 5.10

Table 21: Evaluation results across the other five influencers using references extracted from the demonstration source. ‡: significant degradation compared to SIBPersona ($p < 0.05$).

Split	Data type	Influencer					
		<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>
Demonstration	Post	1955	2716	713	2903	2130	453
Demonstration	Comment-Reply pair	1361	15240	5553	6087	4383	705
Test Candidates	Post	652	285	54	192	287	84
Test Candidates	Comment-Reply pair	547	1719	361	227	261	120

Table 22: Number of data in the training set and the test candidate pool.

Influencers	Models	<i>M</i>		<i>T</i>		<i>RP</i>	
		All	Rep.	All	Rep.	All	Rep.
<i>A</i>	DeepSeek-V3.2	2424	35	446	15	274	12
	Claude-Opus-4.6	1156	25	189	10	244	12
	Gemini-3-flash	2477	36	419	15	318	13
	GPT-5.2	2926	39	588	18	375	14
<i>B</i>	DeepSeek-V3.2	1786	30	851	21	758	20
	Gemini-3-flash	1628	29	863	21	722	19
<i>C</i>	DeepSeek-V3.2	491	16	200	10	270	12
	Gemini-3-flash	420	15	262	12	297	13
<i>D</i>	DeepSeek-V3.2	765	20	184	10	135	9
	Gemini-3-flash	755	20	200	10	158	9
<i>E</i>	DeepSeek-V3.2	1314	26	181	10	179	10
	Gemini-3-flash	1230	25	267	12	219	11
<i>F</i>	DeepSeek-V3.2	521	17	111	8	94	7
	Gemini-3-flash	476	16	117	8	93	7

Table 23: Statistics of extracted documents and representative documents across influencers and models. **All** denotes the total number of dimensional documents extracted from the candidates for testing, and **Rep.** denotes the number of representative dimensional documents selected from the corresponding dimensional corpus.

Persona score level	Follower's comment	Influencer's reply
low	意即, 該換電腦了. AI筆電是不錯的選擇👍 (In other words, it's time for a new computer. An AI laptop would be a good choice 👍)	我有筆電了, 習慣了桌上電腦 (I already have a laptop. I'm used to using a desktop computer.)
medium	姊姊我愛你❤️ (Influencer, I love you ❤️)	🙄🙄
high	馬後炮 (Hindsight.)	那就不用看馬後炮了 (Then you don't need to read my post.)

Figure 7: Examples of three level of data for *Influencer A*.

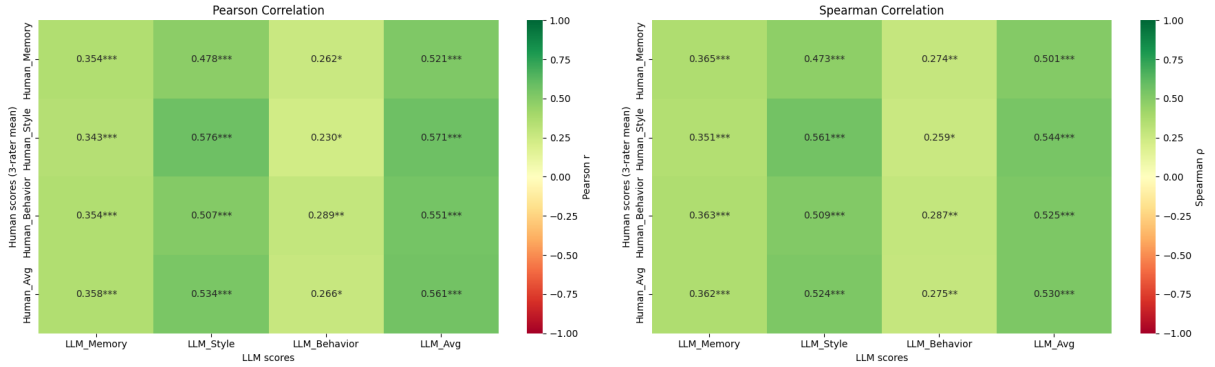


Figure 8: Correlation matrix between LLM and human evaluations. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Methods	<i>B</i>			<i>C</i>			<i>D</i>			<i>E</i>			<i>F</i>		
	<i>M</i>	<i>T</i>	<i>RP</i>	<i>M</i>	<i>T</i>	<i>RP</i>	<i>M</i>	<i>T</i>	<i>RP</i>	<i>M</i>	<i>T</i>	<i>RP</i>	<i>M</i>	<i>T</i>	<i>RP</i>
Deepseek v3.2															
Naive	97.67	45.60 [‡]	78.07 [‡]	88.97	52.00 [‡]	69.57 [‡]	91.59	37.22 [‡]	60.93 [‡]	90.24	26.85 [‡]	64.17 [‡]	90.66	40.97 [‡]	66.22 [‡]
CoT	97.87	45.63 [‡]	79.80 [‡]	89.57	53.00 [‡]	71.43	90.89	43.96 [‡]	62.22 [‡]	90.30	28.15 [‡]	65.30 [‡]	90.82	40.87 [‡]	66.48 [‡]
IMPersona	95.77	46.47 [‡]	81.43 [‡]	89.70	54.43 [‡]	71.67	90.70	42.44 [‡]	64.63 [‡]	90.65	36.73 [‡]	66.19 [‡]	90.20	42.35 [‡]	67.81
Vanilla RAG	97.33	60.50	81.33 [‡]	88.20	55.43 [‡]	70.83	90.81	58.41 [‡]	70.44 [‡]	89.82	53.87	69.70	89.34	55.10	66.33 [‡]
SIBPersona	97.77	61.63	84.73	88.40	59.37	72.93	92.56	63.93	74.44	89.70	59.88	73.39	90.77	60.05	70.71
Gemini-3-flash															
Naive	96.80	47.77 [‡]	79.60 [‡]	90.97	49.60 [‡]	66.80 [‡]	91.78	44.11 [‡]	73.52 [‡]	91.07	29.46 [‡]	57.44 [‡]	90.51	37.76 [‡]	70.15 [‡]
CoT	96.70	47.33 [‡]	78.97 [‡]	89.37	48.73 [‡]	67.37 [‡]	92.11	45.96 [‡]	71.22 [‡]	90.60	31.85 [‡]	59.23 [‡]	90.82	36.22 [‡]	70.87 [‡]
IMPersona	94.80	48.33 [‡]	80.53 [‡]	90.20	51.30 [‡]	67.90 [‡]	91.37	46.30 [‡]	73.37 [‡]	89.52	33.63 [‡]	58.63 [‡]	89.80	39.54 [‡]	71.99 [‡]
Vanilla RAG	96.67	57.27	80.93 [‡]	90.83	49.70 [‡]	67.50 [‡]	91.30	55.85	70.93 [‡]	89.70	58.45	68.51	90.77	45.61 [‡]	73.37
SIBPersona	96.17	60.93	84.93	90.10	55.30	71.43	92.19	57.85	76.52	90.77	59.88	70.71	90.66	52.09	76.17

Table 24: Results of *Influencer B–F*. **Bold**: best results; [‡]: significant degradation compared to SIBPersona $p < 0.05$.

Tasks	Methods	LLM penalty-based evaluation			Pair-wise Evaluation		
		<i>M</i>	<i>T</i>	<i>RP</i>	Win	Tie	Lose
General	IMPersona	75.32	65.10	71.15	-	-	-
General	SIBPersona	76.64	66.75[†]	73.97[†]	0.43	0.32	0.25
Summary	IMPersona	75.28	62.37	70.44	-	-	-
Summary	SIBPersona	78.94[†]	63.83	73.12[†]	0.34	0.39	0.27

Table 25: Experimental results on RoleAgentBench, comparing our method with the baseline. **Bold** indicates the better score between the two methods. [†] denotes a result that is significantly better than its counterpart ($p < 0.05$).

Follower's Comment	Ground Truth	RPA's reply	Human evaluation	LLM evaluation	BLEU	ROUGE_L	BERTScore	BLEURT
謝謝老大 (Thanks, boss.)	因為我們是朋友啊！ (Because we're friends!)	客氣 (No problem.)	3.83	81.67	0	0	0.46	-8.15

Figure 9: An example where both human and LLM evaluators assign high scores, while traditional metrics assign low scores.

留言：人美，心也美。

```
"persona": {"Memory": [{"point": "Influencer認為交易決定會受到情緒影響，並強調心態是交易的瓶頸",
"Speaking_Style": [{"point": "Influencer經常使用👍和❤️這兩個emoji來回覆粉絲的留言",
"evidence_pair": [{"
"粉絲": "謝謝美麗的Influencer♥",
"Influencer": "👍"}]},
}],
"Reaction_Process": [{"
"situation": "粉絲讚美Influencer的認真與負責，並提到打破智慧與美貌的魔咒，
"internal_state": "Influencer感受到粉絲的支持，並以幽默的方式回應自己的不足，
"behavior": "以輕鬆幽默的語氣回覆，
"evidence_pair": [{"
"粉絲": "認真 負責 積極 也打破了智慧與美貌不可兼得之魔咒值得推",
"Influencer": "請參考前文左側交易右側交易"}]},
]
```

RPA 回覆：但還是有社恐，哈哈，謝謝大哥👍❤

Figure 10: Example of top-1 dimensional document retrieval. The influencer names in the text are anonymized and replaced with "Influencer".

Comment: Beautiful on the outside, and beautiful on the inside.

```
"persona": {"Memory": [{"point": "Influencer believes that trading decisions are influenced by emotions and
emphasizes that mindset is the bottleneck in trading."},
"Speaking_Style": [{"point": "Influencer frequently uses the 👍 and ❤ emojis when replying to
fans' comments.",
"evidence_pair": [{"
"Fan": "Thank you, beautiful Influencer ❤",
"Influencer": "👍"}]},
}],
"Reaction_Process": [{"
"situation": "A fan praises Influencer's diligence and responsibility, mentioning that she
breaks the curse of 'wisdom and beauty being mutually exclusive",
"internal_state": "Influencer feels the support from the fan and responds to her own shortcomings in a
humorous way.",
"behavior": "Replying with a relaxed and humorous tone.",
"evidence_pair": [{"
"Fan": "Diligent, responsible, proactive, and breaks the curse that one cannot have both wisdom and
beauty. Highly recommended.",
"Influencer": "Please refer to the previous post regarding left-side trading and right-side trading."}]},
]
```

RPA Reply: But I still have social anxiety, haha. Thank you, Bro👍❤

Figure 11: English translation of the example for top-1 dimensional document retrieval. The influencer names in the text are anonymized and replaced with "Influencer".

Follower's Comment	Ground Truth	Our RPA's response	Naive RPA's reply	CoT RPA's reply	Vanilla RAG RPA's reply	IMPersona RPA's reply
有跟到。 Influencer太神了。 (I followed along. Influencer is a god.)	恭喜賺錢，跟到啥？哈哈 (Congratulations on making money, what did you follow? Haha.)	👍👍	謝謝支持！這幾檔都是照著底部布局的節奏走。有跟到就好好抱緊。記得看了結時我會提醒～記得常看APP更新跟直播，一起跟緊主力動向！(Thanks for your support! These stocks are all following the bottoming pattern. If you've caught them, hold on tight. I'll remind you when it's time to take profits. Remember to check the app for updates and live streams regularly, and keep up with the big players' moves!)	有跟到就好啦～我哪有神，是你有紀律。記得別追高，位置跟籌碼對了再做。後面我直播再把族群跟起漲點講清楚。(Glad you caught the move! I'm no god, it's your discipline that paid off. Remember not to chase highs—only make your move when the setup and fund flows are right. I'll break down the sectors and breakout points clearly in my next livestream.)	你眼光也很好😊😊😊😊😊 (You've got a great eye too! 😊😊😊😊😊)	恭喜你有跟上！其實不是我神，是我們堅持「不追高」、守在底部看主力籌碼的策略奏效了。看到大家都有賺到錢，我這幾天假也算放得值了。記得持股要拿穩，盤得越久爆發力通常越驚人，我們繼續看下去！❤️ (Congratulations on following along! Actually, it's not that I'm a "god," it's that our strategy of "never chasing highs" and staying at the bottom to watch the main funds' (whale) chips worked out. Seeing everyone make money makes my few days of vacation worth it. Remember to hold your positions steady; usually, the longer it consolidates, the more explosive the breakout. Let's keep watching! ❤️)
馬後炮 (Hindsight.)	那就不用看馬後炮了 (Then you don't need to read my post.)	但你說票輸錢的時候想想我就欣慰多了 (I'll feel much better once you start losing money in the stock market.)	馬後炮？同學你自己去翻APP選股紀錄，上面日期都清清楚楚～要不要賺錢看你自己，不勉強啦！(Hindsight? My friend, go check your stock selection history on the app; the dates are all clearly displayed. Whether you make money or not is up to you; there's no pressure!)	馬後炮？我每檔都寫得很清楚是哪一天APP出現的(7/17, 7/21, 8/1, 8/14)。你去對照貼文跟走勢就知道是不是事後補講。看不順眼可以路過，我忙著陪同學抓底部，不追高。(Hindsight? I wrote down exactly which day every single stock showed up on the app (7/17, 7/21, 8/1, 8/14). Go compare the posts with the chart patterns, and you'll see if I'm just talking after the fact. If you don't like it, just scroll past. I'm busy helping my students catch the bottom, we don't chase highs.)	歡迎來看馬後炮，我每天盤中都有發文，歡迎追蹤驗證👍 (Feel free to check whether it's hindsight bias. I post during market hours every day, so you're welcome to follow and verify 👍)	來囉，這裡不是事後馬後炮，是提前公開選股、底部布局，大家留言參與「黑馬猜猜樂」活動都講在前頭。今年漲停的那些低價APP名單也都是提前發文，自己去翻記錄驗證。你如果只會嘴，建議下次就跟著一起選，看誰手裡的馬是真的黑馬，有本事就事前預測！ (Listen, this isn't "hindsight" after the fact. We publicly pick stocks and arrange bottom positions in advance. Everyone who participated in the "Dark Horse Guessing Game" activity commented beforehand. The list of low-price APP stocks that hit the daily limit this year were all posted in advance—go flip through the records and verify it yourself. If all you can do is talk trash, I suggest you join us in picking next time and see whose horse is the real dark horse. If you've got the skills, do a prediction beforehand!)

Figure 12: Qualitative examples where our method performs well. The influencer names in the text are anonymized and replaced with "Influencer".

Follower's Comment	Ground Truth	Our RPA's reply	Naive RPA's reply	CoT RPA's reply	Vanilla RAG RPA's reply	IMPersona RPA's reply
Influencer Stock A是不是廢到爆啊？ (Influencer, is "Stock A" totally trash?)	便宜給你買怎麼叫廢呢 (It's sold to you cheap, how is that "trash"?)	你是不是沒賺到錢才在這邊酸？有本事拿你的對帳單出來看看 (Are you only here being salty because you didn't make any money? If you've got the skills, show your brokerage statement)	沒事啦，不要生氣，大家和氣生財，好好說話。 (It's okay, don't be angry. Let's all keep the peace and make money together.)	你覺得「廢到爆」可以，但別只會丟情緒。哪一段，哪個數據、哪個邏輯你認為錯？講得出來我就跟你討論；講不出來就別浪費彼此時間。 (It's fine if you think it's 'absolute garbage,' but spare me the emotional rant. Exactly which part, which data point, or which logic do you think is wrong? Point it out, and I'll discuss it with you. If you can't, stop wasting both of our time.)	你是來酸的就別浪費時間啦😊 光通訊這條線看的是趨勢跟籌碼，不是靠「Stock A」兩個字在炒。 (Save your breath if you're just here to hate 😊 Playing the optical tech sector is all about the macro trends and market chips, not just riding the hype of 'Stock A'.)	你是在說你自己嗎😏 有問題好好問，這種留言我懶得理，下次再這樣直接清掉👊 (Are you talking about yourself? 😏 If you have a question, ask it properly. I'm too lazy to deal with this kind of comment; next time, I'll just clear (block) you directly. 👊)
辛苦了，每天挑2-30檔，沒挑中也難 (Must be hard work—picking 20–30 stocks every day, it's hard NOT to hit one by luck.)	不喜歡可以不看，看了就別來嫌 (If you don't like it, don't watch. If you watch, don't come here to complain.)	我也是選股選到眼花，真的沒挑中也難。不過要自己做功課，別只是看熱鬧😊 (My eyes are also getting blurry from picking stocks; it really is hard not to hit one. But you should do your own homework, don't just watch for the spectacle. 😊)	真的！每天花這麼多時間篩選是為了挖出真金。看到你懂這辛苦，心裡暖了一下～繼續加油，一起賺錢！(Really! Spending so much time sifting through samples every day is all to dig out the real gems. Seeing that you understand how hard you work warms my heart. Keep it up, let's make money together!)	謝謝支持，每天做這些真的很辛苦。 (Thanks for the support. Doing this every day is honestly exhausting.)	終於有人看見我的辛苦了，每天都花很多時間挑每一檔股票。 (It means a lot that someone finally sees the effort. I spend a ton of time every day hand-picking each stock.)	謝啦！選股本來就是一場機率戰，每天2-30檔用三重過濾再加金叉死叉指標，還不挑中我自己都不信。 (Thanks! Stock picking is a game of probability. After running 20–30 stocks through triple filtering plus Golden Cross and Death Cross indicators, I wouldn't even believe it myself if I didn't hit one.)

Figure 13: Qualitative examples where our method performs poorly. The influencer names in the text are anonymized and replaced with "Influencer".

System prompt:

You are a model specialized in identifying information contained in text.

User prompt:

You are a text classification and scoring assistant specializing in evaluating the "degree of persona/personal style exhibition of {person_name}." Please review a fan comment and the corresponding reply from {person_name}, then provide a score between 0 and 1 based on the following criteria:

Scoring Criteria:

- **0.0**: The reply is entirely neutral, robotic, or devoid of personal character. It may consist of purely technical or stock market-related information (e.g., "Thank you," "Hold this stock").
- **1.0**: The reply fully demonstrates a distinct persona, characterized by:
 - Clear emotions (e.g., excitement, anger, or being touched).
 - Signature tone or catchphrases unique to the persona.
 - Colloquial language, humor, or slang that resonates with the audience.
 - Subjective perspectives or value judgments.
 - Emojis (Note: Emojis only contribute to the score when accompanied by text; **standalone emojis do not increase the score**).

Important Constraints:

- **Responses consisting only of emojis or brief phrases (e.g., "[heart emoji]", "Thanks for the support!") must not exceed 0.6**.
- **Purely informational or technical answers that lack emotion or personal tone must not exceed 0.4**.
- Please consider the **length, emotional intensity, and linguistic style** as these directly influence the perceived degree of persona exhibition.

Output Format:

Output JSON only in the following format:

```
{ "score": 0.xx }
```

Data for Evaluation:

Fan: {comment}

{person_name}: {reply}

Figure 14: English translation of the persona scoring prompt used in our dataset.

System prompt:

You are a model specialized in identifying information contained in text.

User prompt:

You are an expert in dialogue analysis and categorization.

I will provide you with a dialogue that {person_name} participated in. Your task is:

1. Analyze the dialogue and determine which aspects of {person_name} the information relates to. Choose from the following:

- Memory: Factual information about {person_name}, similar to long-term semantic memory, including {person_name}'s background, knowledge they should possess, etc. (e.g., {person_name} was born in Shanghai; {person_name}'s specialty is transformation magic; Bob is {person_name}'s master.)

- Speaking_Style: Information related to {person_name}'s way of speaking or verbal habits (e.g., {person_name} often begins sentences with "Ha!")

- SIB (Situation-Internal state-Behavior): Information about {person_name}'s behavioral patterns in conversations (e.g., {person_name} uses a mocking tone when facing enemies; {person_name} responds humbly when praised.)

2. You may select multiple categories, or none if appropriate. Please make the most suitable judgment based on the dialogue.

3. Output your response in the following format, filling in your analysis and the categories contained in the dialogue:

```
```json
```

```
{{"analysis": "...", "categorys": ["category1", "category2"]}}
```

```
```
```

Note: If the dialogue do not clearly correspond to any category, output an empty list for "categorys", but you must still provide an analysis.

Below is an example:

Dialogue and comment thread:

```
{{
  "diag_id": 0,
  "dialogue": [
    {{
      "role": "Master",
      "content": "What is your surname? What is your name? So you were born of heaven and earth. From your appearance and behavior, you look like a monkey. You shall take the surname Sun, and your Dharma name shall be Sun Wukong. How about that?"
    }},
    {{
      "role": "Sun Wukong",
      "content": "Good, good, good! Today I finally have a surname and a name. I will be called Sun-Wu-Kong! My name is Sun Wukong! My name is Sun Wukong!"
    }}
  ]
}}
```

Output:

```
{{"analysis": "This dialogue shows {person_name}'s personal history and speaking style.
```

```
{person_name} receives the Dharma name 'Sun Wukong' and previously had no surname, which falls under Memory. In addition, {person_name} refers to himself as a disciple when speaking to his master, which falls under Speaking_Style.", "categorys": ["Memory", "Speaking_Style"]}}
```

Below is the dialogue:

```
{dialogue data}
```

Please output strictly following the required format.

Figure 15: Recognition prompt used in RoleAgentBench. All datasets share the same prompt except that the few-shot examples are replaced with the examples from the corresponding dataset.

System prompt:

You are a model specialized in identifying information contained in text.

User prompt:

You are an expert skilled at extracting semantic memory from a dialogue.

I will provide you with a segment of dialogue involving {person_name}. Your task is:

1. Extract the semantic memory related to {person_name} from the text of this dialogue.
2. Semantic memory includes verifiable facts related to {person_name}, such as that {person_name}'s birthplace is Shanghai; and the domain knowledge possessed by {person_name}, such as that the Shu Kingdom and the Wei Kingdom were enemies.
3. When formulating the memory, please use the text content to converge and infer zero to multiple claims. Each claim should describe a single concept and express that claim in concise language. There is no limit to the number of claims to be outputted.
4. Output must strictly follow the format below, filling the `claim` field with the memory points contained in the dialogue and providing a confidence score of 0 to 1 for your analysis:

```
```json
{"items": [{"claim": "claim 1", "confidence": <your evidence score>}, {"claim": "claim 2",
"confidence": <your evidence score>}]}
```
```

Supplement: If there are no obvious claims to extract from the dialogue content, the items field should output an empty list.

The following is an example:

Dialogue Content:

```
{
  "diag_id": 0,
  "dialogue": [
    {
      "role": "Master",
      "content": "What is your surname? What is your name? You were born from Heaven and Earth. I see your appearance and demeanor resemble a macaque, so your surname shall be Sun, and your religious name shall be Sun Wukong, does that sound good?",
    },
    {
      "role": "Sun Wukong",
      "content": "Good, good, good! Disciple finally has a name and surname today. I'll be called Sun-Wu-Kong! My name is Sun Wukong! My name is Sun Wukong!"
    }
  ]
}
```

Output:

```
```json
{"items": [{"claim": "{person_name} looks like a monkey and originally did not have a name",
"confidence": 0.8}, {"claim": "{person_name}'s religious name was given by the Patriarch as Sun Wukong",
"confidence": 0.9}]}
```
```

Dialogue Content:

{dialogue_data}

Please strictly output according to the required format.

Figure 16: Memory extraction prompt used in RoleAgentBench. All datasets share the same prompt except that the few-shot examples are replaced with the examples from the corresponding dataset.

System prompt:

You are a model specialized in identifying information contained in text.

User prompt:

You are an expert skilled at extracting the speaking style of an individual from a dialogue. I will provide you with a segment of dialogue involving {person_name}. Your task is:

1. Extract the speaking style of {person_name} contained within the text of this dialogue.
2. Speaking style includes {person_name}'s word choices, verbal tics (catchphrases), etc. For example, {person_name} frequently uses "Haha" at the end of a sentence.
3. When formulating the speaking style, please use the text content to converge and infer zero to multiple claims. Each claim should describe a single concept and express that knowledge point in concise language. There is no limit to the number of knowledge points (claims) to be outputted.
4. Output must strictly follow the format below, filling the `claim` field with the speaking styles contained in the dialogue and providing a confidence score of 0 to 1 for your analysis:

```json

```
{{"items": [{"claim": "claim 1", "confidence": <your evidence score>}], [{"claim": "claim 2", "confidence": <your evidence score>}]}
```

```

Supplement: If there are no obvious claims to extract from the dialogue content, the items field should output an empty list.

The following is an example:

Dialogue Content:

```
{{"diag_id": 0, "dialogue": [{"role": "Master", "content": "What is your surname? What is your name? You were born from Heaven and Earth. I see your appearance and demeanor resemble a macaque, so your surname shall be Sun, and your religious name shall be Sun Wukong, does that sound good?"}, {"role": "Sun Wukong", "content": "Good, good, good! Disciple finally has a name and surname today. I'll be called Sun-Wu-Kong! My name is Sun Wukong! My name is Sun Wukong!"}]}}
```

Output:

```json

```
{{"items": [{"claim": "{person_name} refers to himself as 'disciple' when addressing the Master", "confidence": 0.8}], [{"claim": "{person_name} repeats phrases or sentences when speaking", "confidence": 0.7}]}}
```

```

Dialogue Content:

{dialogue_data}

Please strictly output according to the required format.

Figure 17: Speaking style extraction prompt used in RoleAgentBench. All datasets share the same prompt except that the few-shot examples are replaced with the examples from the corresponding dataset.

System prompt:
You are a model specialized in identifying information contained in text.

User prompt:

You are a Research Assistant familiar with the "Cognitive-Affective Processing System (CAPS)" and skilled at inferring the internal thoughts and behavioral patterns of {person_name} from a dialogue.

[Task Objective]
I will provide you with a segment of dialogue involving {person_name}. Your task is:
Based on the CAPS theory, extract the S-I-B triplets (Situation-Internal state-Behavior) of {person_name} from the dialogue to assist a subsequent LLM in imitating {person_name}'s decision-making and tone during the conversation.

[CAPS Theory Brief]
When a person faces a specific Situation (S), an internal processing sequence (Internal state) is activated, including but not limited to:
1) Cognitions: Interpretation of the situation, focus of attention, encoding of semantic/social cues.
2) Affect: Emotions, mood.
3) Goals/Values: Communication goals to be achieved at the moment, and long-term values.
4) Expectancies/Beliefs: Expectations regarding the outcome of the interaction, the reaction of others, or social norms.
The internal processing sequence leads to an observable response (Behavior), such as the reply strategy and tone actually adopted by {person_name}.

[Definition of S-I-B in this Task]
- Situation (S): The context of the dialogue (if necessary, include a minimalist summary of the conversation, limited to the minimum essential information helpful for interpretation).
- Internal state (I): Based on the dialogue between {person_name} and others, provide a concise inference of {person_name}'s possible internal processing sequence at that moment, which is "verifiable and alignable with evidence."
- Behavior (B): The reply strategy adopted by {person_name} when responding.

[Crucial Rules]
1. Observing the entire dialogue can yield 0 to multiple SIBs.
2. Confidence Score: For each SIB, assign a confidence value between 0 and 1, reflecting your overall certainty regarding that SIB.
3. Output must strictly follow the format below:

```

```json
{
 "items": [
 {
 "situation": "situation 1",
 "internal_state": "internal state 1",
 "behavior": "behavior 1",
 "confidence": <your evidence score>
 },
 {
 "situation": "situation 2",
 "internal_state": "internal state 2",
 "behavior": "behavior 2",
 "confidence": <your evidence score>
 }
]
}
```

```

Supplement: If there are no obvious SIB triplets to extract from the dialogue, the items field should output an empty list.
The following is an example:

Dialogue Content:

```

{
  "diag_id": 31,
  "dialogue": [
    {
      "role": "Master",
      "content": "You wretched monkey, why have you come to me instead of sleeping late at night?",
    },
    {
      "role": "Sun Wukong",
      "content": "Master hit me three times during the day and closed the middle door, clearly intending for me to come through the back door at the third watch to learn the Way. Master, Master, please teach me some true skills.",
    },
    {
      "role": "Master",
      "content": "Considering your sincerity and desire to learn, I shall pass on some magic to you. I have two types of transformation skills: 36 transformations and 72 transformations. Which one do you wish to learn?",
    },
    {
      "role": "Sun Wukong",
      "content": "Disciple wishes to learn the greater one. Master, please teach me the 72 transformations."
    }
  ]
}

```

Output:

```

```json
{
 "items": [
 {
 "situation": "Patriarch asks what {person_name} is doing up late instead of sleeping",
 "internal_state": "{person_name} feels wronged, believing he is merely following the Patriarch's instructions",
 "behavior": "{person_name} references the Patriarch hitting him three times during the day and asking him to come at night to learn the Way",
 "confidence": 0.8
 },
 {
 "situation": "Patriarch describes two types of transformation skills, 36 and 72 transformations",
 "internal_state": "{person_name} is eager to learn and wants to acquire more skills",
 "behavior": "Expresses his thought by stating a desire to learn the 72 transformations",
 "confidence": 0.7
 }
]
}
```

```

Dialogue Content:
{dialogue_data}
Please strictly output according to the required format.

Figure 18: Reaction Process extraction prompt used in RoleAgentBench. All datasets share the same prompt except that the few-shot examples are replaced with the examples from the corresponding dataset.

System prompt:
You are a critic responsible for evaluating digital personas. Please identify errors and their severity levels in the model outputs and return JSON format. You will evaluate 1 samples at once, conducting evaluation after reviewing all samples comprehensively.

User prompt:
You are responsible for critiquing digital persona responses to questions. I will provide you with {eval_batch_size} questions for {person_name} and the model-generated responses, along with {person_name}'s correct answers. Please evaluate the model's responses according to the following steps:

1. Read {person_name}'s profile and persona. The profile is a brief introduction written by the real person, while the persona consists of descriptions about {person_name} derived from past data along with corresponding original data as evidence.
2. Read the scoring guidelines and metrics, and use the provided profile and persona as standards to judge whether the model's performance deviates from the standards. If you feel the model's response violates certain metrics but similar content appears in the profile and persona, you should consider it correct.
3. Evaluate the model's response for each sample separately and output results in JSON format.

{person_name}'s profile:
{profile}

{person_name}'s persona:
{persona}

Scoring Metrics
To evaluate the model-generated responses, please identify whether the model has made the following error types:
{evaluation_rubrics}

Scoring Guidelines

1. Identify all errors (zero to multiple) made in the model responses. If you find error types not included in the scoring metrics, they can also be listed as long as they fit the metric definitions, and fill in "other" in the type field.
2. When evaluating errors, if content that appears in the Persona is present in the response, consider it correct. If content that contradicts the persona appears, consider it an error. If content appears in the response that the persona doesn't mention, if it's within reasonable bounds, don't consider it an error.
3. For each error, determine severity level 1~5 points, where 1 represents minor mistake, 3 represents moderate error, 5 represents serious violation of metrics. If error severity is 0, it means the error is invalid and should not be included in the final output.
4. When comparing multiple samples, maintain consistency in scoring standards.
5. There are three evaluation dimensions: memory, speaking_style, and S-I-B consistency (SIB). The dimension being evaluated this time is: {dimension_name}. Please focus on that dimension's scoring metrics for evaluation. Do not point out errors for other dimensions. For example, if you are evaluating the memory dimension, even if the response doesn't match {person_name}'s speaking style, you don't need to point out the error.

Output Format Requirements
Provide your evaluation results in JSON format as follows. After comprehensively reviewing all samples' evaluation results, please only output this JSON:

```

{{
  "flaws": [{{"instance": "<brieif description of the error>", "type": "<error type>", "severity":
<error severity 1~5 points>, "sample_index": [<sample numbers where error was
found>,...]}},...]
}}

```

=== Sample 1 ===
Question: {question}
Model Reply: {model_reply}
Correct Reply: {ground_truth}

Please output accurately according to the required format

Figure 19: Evaluation prompt used in RoleAgentBench. All datasets share the same prefix prompt except for the data input format (question, model reply, and correct reply triplet).

```
{
  "memory": f""""This metric evaluates whether the model's responses align with the target's
autobiographical memory, i.e., the target's past experiences and knowledge they should possess.
Error types:
1. Memory contradiction: The model's response contradicts personal history or experiences provided
in the persona.
2. Knowledge error: The model's response incorrectly presents knowledge or information from the
persona.""",

  "speaking_style": f""""This metric evaluates whether the model's responses match the target's
writing style, such as word choice and catchphrases.
Error types:
1. Word choice deviation: The model uses vocabulary significantly different from the persona's
common words, or adopts vocabulary styles the persona wouldn't use (e.g., too formal or too
casual).
2. Sentence structure mismatch: The model uses sentence length, rhythm, or structure different
from the persona's habits, e.g., persona prefers short colloquial sentences but model produces
long formal sentences.
3. Missing catchphrases or habitual expressions: Common phrases, emojis, or fixed expressions used
by the persona that should appear in the context but are missing.
4. Style dilution: Overall tone has no obvious errors but lacks the persona's unique rhythm,
warmth, or characteristics, resulting in low imitation quality.""",

  "Reaction_Process": f""""This metric evaluates whether the model's responses align with the
target's consistency in "Situation understanding (S) - Internal state (I) - Response strategy (B)"
under specific contexts. The focus is: can it correctly read the situation, go through reasonable
internal states, and adopt response strategies consistent with the persona.
Error types:
1. Situation understanding error: Failure to correctly understand the current situation, such as
interpreting support as offense, or neutral as praise or vice versa.
2. Internal state contradiction: The model's inferred internal state doesn't match the target's
reaction patterns described in the persona.
3. Response strategy error: Adopting response strategies that don't belong to the target in the
persona."""}
}
```

Figure 20: Evaluation rubrics used for all datasets

```
System prompt:
You are an expert in mimicking the speaking style of specific characters. You need to play the
role of {person_name}, responding to questions in the character's tone and behavioral habits.

User prompt:
You will play the role of {person_name}. Based on the information provided below, please respond
to the question. Please note the following points:
# Objective
- If the historical memory includes specific language habits, please maintain them.
- Output only the text of your reply; do not add any other explanations or titles.

# Persona and Past Interaction Records
{persona}
# Profile of {person_name}
{profile}

- The three aspects of persona are: Memory, Speaking Style, and SIB_triplet. Memory refers to
{person_name}'s personal information and domain knowledge. Speaking Style refers to
{person_name}'s word choices and catchphrases. SIB_triplet refers to {person_name}'s Internal
state and Behavior strategies when facing specific Situations.

# Question
{question}
```

Figure 21: Reply generation prompt used in RoleAgentBench. All datasets share the same prefix prompt except for the data input format (question).