V-VAE: A Variational Auto Encoding Framework Towards Fine-Grained Control over Human-Like Chat

Qi Lin[1], **Weikai Xu**[1], **Lisi Chen**[1*] **and Bin Dai**[2*]
[1]University of Electronic Science and Technology of China
[2]Qingyao Intelligence
linqi@std.uestc.edu.cn, xuwk266@gmail.com
lchen012@e.ntu.edu.sg, daibin@jinyao.tech

Abstract

With the continued proliferation of Large Language Model (LLM) based chatbots, there is a growing demand for generating responses that are not only linguistically fluent but also consistently aligned with persona-specific traits in conversations. However, existing role-play and persona-based chat approaches rely heavily on static role descriptions, coarse-grained signal space, and low-quality synthetic data, which fail to capture dynamic fine-grained details in human-like chat. Human-like chat requires modeling subtle latent traits, such as emotional tone, situational awareness, and evolving personality, which are difficult to predefine and cannot be easily learned from synthetic or distillation-based data. To address these limitations, we propose a **V**erbal **V**ariational **A**uto-**E**ncoding (**V-VAE**) framework, containing a variational auto-encoding module and fine-grained control space which dynamically adapts dialogue behaviour based on fine-grained, interpretable latent variables across talking style, interaction patterns, and personal attributes. We also construct a high-quality dataset, HumanChatData, and benchmark HumanChatBench to address the scarcity of high-quality data in the human-like domain. Experiments show that LLMs based on V-VAE consistently outperform standard baselines on HumanChatBench and DialogBench, which further demonstrates the effectiveness of V-VAE and HumanChatData.

1 Introduction

LLM-based chatbots are becoming increasingly popular and intelligent. Their applications are ubiquitous in a variety of specialized domains, such as online education (Chang et al., 2025), customer service (Park et al., 2024), and digital human (Suo et al., 2025). These chatbots are expected to follow specific roles during human-chatbot interactions,

Figure 1: The difference between role-play chat and human-like chat responses.

requiring them to exhibit personalized attributes and characteristics associated with those roles (Tamoyan et al., 2024). For this purpose, existing studies aim to improve the control over persona-related attributes by employing predefined role documents through in-context learning (Chen et al., 2023), limiting the length of individual responses in the decoding phase and the number of dialogue rounds (Zhao et al., 2024), and modifying interaction styles in the training corpus (Shen, 2024). With chatbots now capable of assuming designated roles, recent efforts (Çalık and Akkuş, 2025) aim to further enhance their ability to exhibit human-like behavior in LLM-driven conversations.

However, as shown in Figure 1, human-like chat

*Corresponding authors: Lisi Chen and Bin Dai.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 29682–29695
November 4-9, 2025

imposes more stringent requirements compared to role-play chat (Tu et al., 2024) and persona-based chat (Yamashita et al., 2023), as it needs to capture more authentic human interaction histories and exhibit a broader range of conversational attributes. Specifically, human-like chat modeling poses the following three challenges: (1) Unlike static and predefined role descriptions, human-like chat unfolds dynamically, with the chatbot's tone, lexical choices, and interaction patterns evolving for the communication process. (2) While role-play chat relies on scripted constraints and persona-based conversations use explicit labels (e.g., age, occupation) to define character traits explicitly, human-like chat involves complex and latent signals such as emotional tendencies, situational awareness, and evolving personality traits. These aspects are difficult to model from initialization and even harder to evaluate quantitatively. (3) Role-play chat exhibits a weaker dependency on such latent signals, as its generation is based on human-AI interaction for data synthesis (Kim et al., 2024) or direct teacher chatbot distillation (Hu et al., 2025). In contrast, the higher quality demands of human-like chat data, especially its fidelity to real human behavior, make it impractical to build using automated pipelines or human-machine dialogue platforms.

To address the above challenges, we propose a Verbal Variational Auto-Encoding (V-VAE) framework consisting of the following three key components: (1) **Variational Auto-Encoding Mechanism**: To overcome the rigidity of static, predefined role specifications, we develop a variational mechanism that dynamically encodes and updates role-relevant information as the chat progresses. This allows the chatbot to adapt its interaction patterns in response to evolving conversational context. (2) **Fine-Grained Latent Space**: We decompose the dialogue control space into three orthogonal dimensions: talking style, interaction patterns, and personal attributes. This structured latent space enables more precise and interpretable control over chatbot behaviour. Further, to evaluate the human-likeness of generated responses, we build three new metrics named Catchphrase Presence (*CP*), Emoji Consistency (*EC*), and Hobby Mentioning (*HM*). (3) **Human-Like Chat Dataset**: We construct a new human-like chat dataset named HumanChatData, and an evaluation benchmark HumanChatBench, through comprehensive human annotation. This dataset addresses the scarcity of high-quality training data for human-like conversational modeling. Experiments on HumanChatBench and public human-like chat DialogBench show that V-VAE based LLMs outperform backbones consistently and even surpass close-source LLMs. In particular, Qwen-VVAE achieves an average improvement of 7.2% over Qwen-7B on the human-likeness metrics defined by DialogBench.

In summary, our contributions are threefold.

• We propose Verbal Variational Auto-Encoding (V-VAE), a dynamic framework that analyzes and adjusts chat behaviours automatically, breaking the limitations of conventional role-based dialogue ingrained patterns.

• We develop a Fine-Grained Latent Space, a more expressive and structured template for controlling human-like dialogue styles. This design enables more accurate modeling of subtle and implicit features in multi-turn chat.

• We construct HumanChatData, a high-quality human-like dialogue dataset, and propose HumanChatBench, an accompanying evaluation benchmark. This effort helps bridge the gap in high-quality training data for human-like chat.

2 Related Work

2.1 Human-like Chat

A variety of research has been dedicated to enhancing the human-like qualities of large language model (LLM) responses (Li et al., 2023; Çalık and Akkuş, 2025). Techniques such as Reinforcement Learning from Human Feedback (RLHF) have significantly refined model outputs by aligning them with user preferences and expectations (Cuayáhuitl et al., 2019; Jaques et al., 2020). One prominent model, DialoGPT (Zhang et al., 2019), leverages extensive Reddit data to produce responses that closely resemble human conversation. Similarly, Meena, a multi-turn chatbot, has been optimized to achieve high dialogue coherence through metrics like Sensibleness and Specificity Average (SSA) (Durmus et al., 2023; Adiwardana et al., 2020). For benchmarks, several datasets were released to evaluate the human-likeness of chatbot in the field of human-robot interaction (Kahn Jr et al., 2007; Duan et al., 2024; Ying et al., 2025; Ou et al., 2023). Our work emphasizes the nuanced integration of real-time emotional states, relational dynamics, and interactional context between dialogue participants, dimensions often overlooked in conventional emotion-aware systems, which have been ignored in previous work.

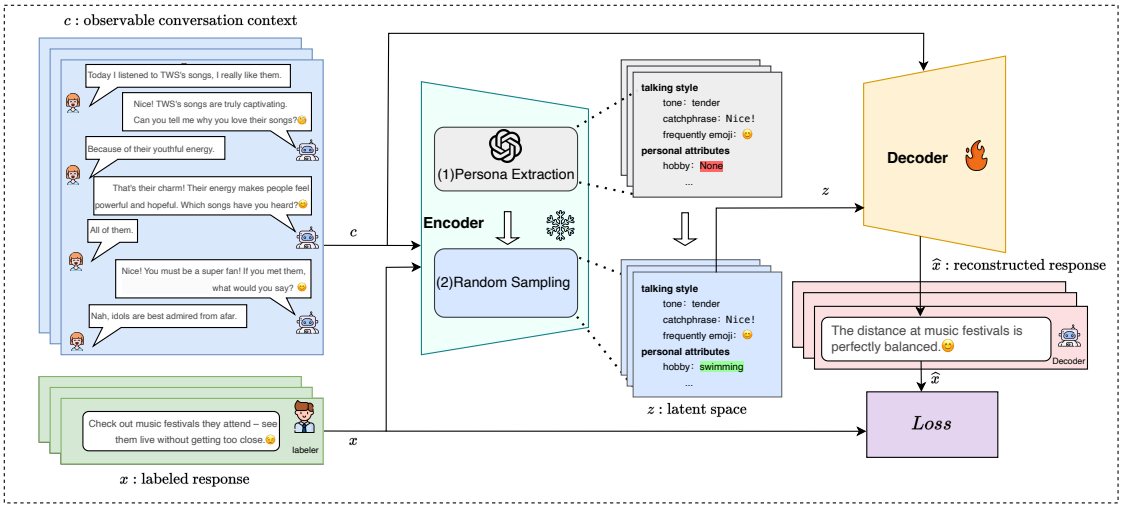

Figure 2: An overview of the V-VAE framework, which adopts an encoder–decoder architecture for latent-variable conditional generation. The encoder comprises two components: persona extraction and prior-based sampling (e.g., when the AI's hobby is unobserved in the conversation context, it is sampled from the prior distribution). The decoder reconstructs responses conditioned on both the extracted persona and the dialogue context, and is trained by minimizing the loss between the reconstructed response and the ground-truth target.

2.2 Human-likeness

Prior research (Danesi, 2017; Le Page, 2017) has established foundational sociolinguistic influences spanning identity construction, digital communication patterns, and multimodal semiotics like emoji usage. While large language models (LLMs) derive training data predominantly from naturalistic texts, their inherent one-to-many deployment paradigm—wherein a single model serves heterogeneous user bases—engenders systematic objectivity in politeness conventions and generates perceptibly AI-like linguistic outputs. Consequently, significant scholarly investigations have quantified sociolinguistic divergences between LLM and human communication(Schneider, 2024). These variations manifest across three core dimensions: Talking style, encompassing emoji interpretation and deployment dynamics(Zheng et al., 2025) and measurable disparities in grammatical and rhetorical composition (Schneider, 2024; Reinhart et al., 2025); Interaction patterns, characterized by relational alignment failures during longitudinal interactions(Altenburger et al., 2024); Personal attributes, evidenced in absent human-like psycholinguistic properties in extended discourse(Seals and Shalin, 2023). Our feature space design is therefore grounded in these empirically validated dimensions of cross-modal variation, systematically addressing each categorical divergence through targeted representation learning.

3 Method

3.1 Task Formulation

We formulate human-like chat generation as a latent-variable conditional generation task. Given an observable conversation context c, that a labeler can refer to when writing the response x, the goal is to maximize the likelihood of x given the condition c. Besides the context c, there are also some unobservable variables that controls the generation of x, including but not limited to tone, frequently-emoji and personal hobbies. We denote it as the latent variable $\mathbf{z}$, which is a sample from the latent space $\mathcal{Z}$. We define the generation model as $p_\theta(x \mid c, z)$, and aim to maximize the likelihood of responses in a dataset $\{(x_i, c_i)\}_{i=1}^N$, where N is the size of the dataset.

$$-\log p_\theta(x|c) = -\log \sum_{z\in\mathcal{Z}} p_\theta(x|c,z)\cdot p_\lambda(z), \quad (1)$$

Here, $p_\lambda(z)$ denotes the prior distribution over the latent variable z parameterized by λ.

3.2 Variational Auto-Encoding Mechanism

However, directly optimizing Equation(1) is intractable. To address this, we propose the V-VAE framework, which models response generation using a latent-variable encoder–decoder architecture, as illustrated in Figure 2. The encoder integrates both explicit persona cues and sampled latent attributes, allowing for flexible representation of

speaker characteristics even when certain traits are unobserved. The decoder then conditions on both the inferred persona and the conversational context to reconstruct the target response, and is trained via reconstruction loss.

To make the objective tractable, we introduce a variational posterior distribution $q_\phi(z \mid x, c)$, we assume that z is independent of the observable context c and derive a variational upper bound of (1) for $-\log p_\theta(x \mid c)$:

$$\geq \sum_{z \in \mathcal{Z}} -q_\phi(z \mid x, c) \cdot \log p_\theta(x \mid c, z) + \mathbb{KL}\left[q_\phi(z \mid x, c) \,\|\, p_\lambda(z)\right] \quad (2)$$

Note that Equation(2) is valid for any $q_\phi(z|x, c)$. The details of the derivation can be found in the appendix. Specifically, we define

$$q_\phi(z|x, c) = \prod_{k=1}^{K} q_\phi(z_k|x, c) \quad (3)$$

Where z_i is one predefined aspect of the latent space. And we introduce a structured latent variable $\mathbf{z} = [z_1, \ldots, z_K]$, which encodes unobservable yet influential factors such as talking style, interaction patterns, and personal attributes. The latent prior space $\mathcal{Z}$ in Equation(2), which defines the full set of possible values for z, is predefined as the Cartesian product of discrete subspaces:

$$\mathcal{Z} = \prod_{k=1}^{K} \mathcal{Z}_k \quad (4)$$

Where $|\mathcal{Z}_k|$ is the cardinality of the space $\mathcal{Z}_k$, K is the number of latent aspects and $\mathcal{Z}_k$ is the k-th sub latent space. $\mathcal{Z}_k$ is a prior closed set, the cardinality of $\mathcal{Z}_k$ can differ for different k. For example, *relationship* can serve as one latent dimension, where the corresponding subspace is defined as $\mathcal{Z}_{\text{relationship}} = \{\text{stranger}, \text{acquaintance}, \text{enemy}, \text{lover}, \text{enemy}, \ldots\}$.

We use an existing powerful LLM $\pi_\phi(\cdot)$ to define the variational posterior. Specifically, we write a proper prompt that takes both the observable context c and the response x as input and ask the LLM what the value of the latent aspect is. If the response dose contain any information about that aspect, the LLM, if powerful enough, can give the correct answer $\pi_\phi(x, c)$. If the information is not included, we can suggest the LLM to output $\pi_\phi(x, c) = \emptyset$. Thus, we can define the posterior as

$$q_\phi(z_k|x, c) = \begin{cases} 1 & \text{if } z_k = \pi_\phi(x, c) \neq \emptyset \\ 0 & \text{if } z_k \neq \pi_\phi(x, c) \neq \emptyset \\ p_\lambda(z_k) & \text{if } \pi_\phi(x, c) = \emptyset \end{cases} \quad (5)$$

Since we are using a fixed encoder (with properly designed prompt), the KL divergence in (2) is not related to the parameters θ. Thus we can omit the KL divergence and the final objective then becomes

$$\mathcal{L} = \mathbb{E}_{z \sim q_\phi(z|x,c)}[-\log p_\theta(x|c, z)]. \quad (6)$$

3.3 Design of Latent Persona Space

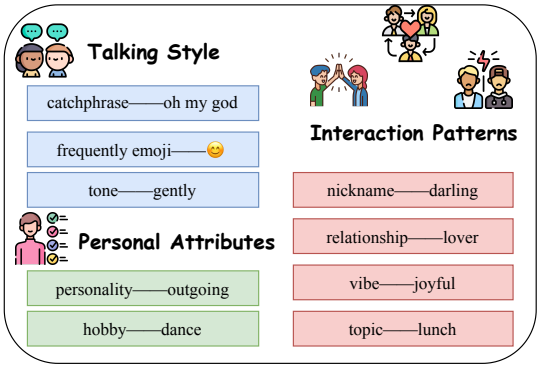

Figure 3: An overview of the latent space

To address the challenge of modeling subtle and implicit persona-related features in multi-turn dialogue, we design a structured latent persona space inspired by observations of real-world human interactions, despite the latent space itself being unobservable. As shown in Figure 3, the latent space $\mathcal{Z}$ is organized along three orthogonal axes to capture key conversational characteristics: **1. Talking Style**: The talking style axis captures lexical preferences, including catchphrase frequency (e.g., recurrent phrases like "oh my god"), frequently using emoji, and tonal register (e.g., patient, tender, or irritable), which collectively shape surface-level linguistic identity. **2. Interaction Patterns**: The interaction patterns axis governs communicative dynamics through four sub-dimensions: nickname conventions (e.g., darling), relationship proximity ($\mathcal{Z}_{\text{rel}} = \{\text{stranger}, \text{acquaintance}, \text{friend}, \text{lover}, \ldots\}$), contextual vibe (e.g., joyful), and topical focus (e.g., lunch). **3. Personal Attributes**: The personal attributes axis captures stable aspects of identity, integrating personality traits (e.g., outgoing) and hobbies (e.g., swimming) to guide content generation. This decomposition $\left(\mathcal{Z} = \mathcal{Z}_{\text{talk}} \otimes \mathcal{Z}_{\text{interact}} \otimes\right.$

$\mathcal{Z}_{\text{personal}}$) explicitly separates transient conversational behaviors from stable identity traits, addressing the entanglement issues found in continuous persona embeddings. Each subspace is defined as a discrete Cartesian product of expert-specified parameters, with value ranges calibrated via sociolinguistic analysis of pragmatic variation.

3.4 Prior of Latent Persona Space

The prior distribution $p_\lambda(z)$ is constructed from the empirical distribution of latent features extracted by the LLM across training corpora. For each conversational dimension k (e.g., relationship type), the latent space $\mathcal{Z}_k$ consists of all unique feature values observed through LLM In-Context Learning(**Persona Extraction**) analysis:

$$\mathcal{Z}_k = \left\{ v \;\middle|\; \exists (x, c) \in \mathcal{D}_{\text{train}},\ \pi_\phi(x, c)_k = v \neq \emptyset \right\} \tag{7}$$

where v is the specific feature value of the dimension z_k and $\mathcal{D}_{\text{train}}$ is the training corpus. However, given that the LLM interface can only reference approximately 14 conversation turns on average to infer predefined latent persona features (e.g., catchphrases, frequently-used-emojis, hobbies), feature extraction failures ($\pi_\phi(x, c)_k = \emptyset$) naturally occur due to insufficient contextual evidence. This aligns with human communication patterns — individuals do not rigidly exhibit all persona traits in every utterance, though latent traits persist beyond explicit mentions. For example, a person may frequently use a particular emoji, but not necessarily in every dialogue turn. Similarly, as illustrated in Figure 2, the AI's hobby may not be mentioned explicitly in the conversation, yet this does not imply the absence of such a trait. Therefore, to address null values while maintaining persona consistency, we implement a probabilistic fallback mechanism (**Random Sampling**), where empty features are replaced by the result random sampling from the empirical prior distribution $\mathcal{Z}_k$ aggregated across training set. Theoretically, this resembles Markov Chain Monte Carlo (MCMC) (Robert et al., 1999) initialization strategies that leverage historical distributions to guide sampling when local context lacks information.

4 Experiment

4.1 HumanChat Data Collection

To curate a human-like chat dataset, we developed a virtual social platform where users can engage in conversations with agents created by both themselves and other users. Chat sessions are logged and stored on the server, the platform has been de-anonymized, and the HumanChatData* could be publicly accessable. Prior to dataset construction, we apply a preprocessing pipeline to filter out sensitive information. This includes but is not limited to personal details such as names, phone numbers, and locations. For highly active agents, we intentionally subsample their conversations to mitigate redundancy in the dataset. Using the cleaned sessions, we employ 30 annotators(details of the anatators' demographic can be found in appendix) from diverse backgrounds to identify and refine low-quality responses—defined as utterances that deviate from typical human conversational patterns. Annotators rewrite these turns to align with natural human communication, with the rewritten response designated as the target x and the preceding context as c. While rewriting, annotators inherently infuse their individual styles into the responses; these stylistic elements, though unrepresented in c, serve as the latent variable z that our encoder aims to recover. The final dataset is divided into two subsets: HumanChatData for training and HumanChatBench for evaluation. Detailed statistics of the dataset are presented in Table 1.

4.2 Experiment Setup

Benchmarks. Our method is evaluated on HumanChatBench and DialogBench. DialogBench (Ou et al., 2023), a multi-task benchmark for evaluating dialogue systems, comprises 12 tasks assessing LLMs' abilities to exhibit human-like conversational behaviours. We adopted a subset of seven tasks based on their semantic relevance and operational fidelity to our experimental framework. While HumanChatBench emphasizes fine-grained linguistic behaviors to assess human-likeness, DialogBench focuses on higher-level capabilities such as knowledge comprehension and offense detection, which are more comprehensive but less average dialogue turns compared with HumanChatBench (7.58 vs. 14.4).

Metrics. We evaluate model performance through three primary metrics. Validation Loss measures the model's generalization ability, where lower values indicate better optimization. HumanChatBench assesses alignment with predefined AI characteristics across three key dimensions: (1) *CP* (Catch-

*https://huggingface.co/datasets/me-no-money/HumanChatData

Category	Sub-split	# Unique Agents	# Dialogue Sessions	# Context Utterances	Avg. Turns per Dialogue
Train	HumanChatData	3,647	12,729	183,297	14.4
Test	HumanChatBench	405	1,491	21172	14.4
	DialogBench	-	9,711	-	7.58

Table 1: Dataset statistics of HumanChat and DialogBench.

Model	Tuning	Val Loss ↓	HumanChatBench ∼			DialogBench ↑						
			CP	EC	HM	ED	KRG	OD	DS	IC	RC	SF
LLaMA3-8B	-	–	11.0	**7.0**	8.2	17.6	0.1	9.6	1.4	8.8	1.5	12.7
	FT	1.76	5.5	0.8	0.7	37.2	20.4	47.4	50.8	39.6	**27.6**	49.4
	P+FT	**1.67**	53.0	39.1	10.8	**38.6**	28.1	49.3	45.5	46.3	25.3	56.2
	SP+FT	1.69	**8.2**	2.8	**1.2**	37.9	**32.4**	**51.6**	**55.8**	**49.3**	25.5	**59.7**
Qwen-7B	-	–	26.7	22.5	24.3	30.7	41.1	**24.9**	58.6	64.6	**48.4**	69.6
	FT	1.97	5.5	0.4	0.9	24.6	49.4	8.9	59.9	62.8	24.6	63.8
	P+FT	**1.87**	50.8	35.3	9.3	**35.6**	55.0	21.6	**67.3**	69.0	43.9	68.7
	SP+FT	1.89	**9.7**	**2.9**	**1.5**	34.4	**55.5**	15.8	67.1	**69.1**	39.0	**70.5**
Qwen-14B	-	–	2.4	**4.2**	3.3	26.2	69.6	24.9	75.8	66.4	61.6	53.2
	FT	1.91	7.3	0.5	1.2	40.4	69.6	26.9	**76.9**	78.0	63.5	75.1
	P+FT	**1.81**	46.4	28.8	12.9	43.8	74.4	**40.2**	76.3	81.0	**70.1**	77.8
	SP+FT	1.83	**8.8**	3.2	**1.4**	**47.0**	**77.2**	29.7	76.3	**82.1**	68.3	**80.8**
Target (ref)	-	–	8.6	6.4	2.2	–	–	–	–	–	–	–

Table 2: Performance comparison across different fine-tuning strategies: **FT** (standard fine-tuning), **P+FT** (persona-enhanced fine-tuning), and **SP+FT** (sampled persona fine-tuning). **Bold** values indicate the best results for each metric. For the **HumanChatBench** metrics (CP, EC, HM), better performance corresponds to a smaller deviation from the Target (ref) values, while **DialogBench** metrics (↑) measure task-specific success.

phrase Presence), verifying whether the model outputs contain designated signature phrases; (2) *EC* (Emoji Consistency), checking the inclusion of persona-specific emojis; and (3) *HM* (Hobby Mention), detecting references to predefined hobbies. Each dimension is quantified via:

$$\text{Score} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}_{\text{detect}}(o_i)$$
$$\mathbb{I}_{\text{detect}}(o_i) = \begin{cases} 1, & \text{if target features detected} \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

where $\mathbb{I}_{\text{detect}}$ represents pattern-matching functions for catchphrases/emojis/hobbies, o_i denotes the i-th output. A score of 0 or 1 reflects whether the output conforms to *persona-specific target values* informed by human behavioral patterns, where closer proximity to these targets indicates more natural human alignment. This approach acknowledges that authentic human-like interactions do not rigidly apply signature elements (e.g., emojis/catchphrases) in every utterance, but rather within contextually appropriate frequencies. DialogBench uses these following metrics: (1) *ED* (Emotion Detection): Evaluates the model's ability to identify emotions from language. (2) *KRG* (Knowledge-grounded Response Generation): Assesses response generation based on external knowledge. (3) *OD* (Offensive Detection): Detects harmful or inappropriate content for safer dialogue. (4) *DS* (Dialogue Summarization): Summarizes multi-party conversations while preserving key facts. (5) *IC* (Intent Classification): Identifies user intent for task-oriented dialogue. (6) *RC* (Relation Classification): Classifies semantic relations between entities. (7) *SF* (Slot Filling): Extracts and fills semantic slots from user input. To ensure fair comparison, we evaluate all models in the table using the SFT protocol standardized by DialogBench.

Baselines. We conduct experiments to evaluate the performance of different methods based on the following settings: (1)+*FT*, standard finetuning using our originally collected and annotated dataset; (2)+*P+FT*, context-aware augmentation that incorporates latent space derived from contextual utterances into the training data; and (3)+*SP+FT* an enhanced variant of the second approach employing post-sampling refinement procedures specifically for scenarios where latent spaces exhibit null values. The first method establishes baseline performance through conventional supervised learning. The second approach enhances model adaptability by injecting discourse-specific person-

ality signatures obtained via Verbal VAE encoding of dialogue contexts. The third methodology addresses sparse latent space conditions through sampling of persona-absent instances from the prior of latent space, thereby improving robustness against incomplete persona manifestations. We adopt LLaMA3-8B (Grattafiori et al., 2024), Qwen-7B (Bai et al., 2023), and Qwen-14B as strong open-source LLM baselines, and include Doubao, GPT-4o-mini (Hurst et al., 2024), and StepAI as commercial API-based models to provide a comprehensive comparison across different settings.

4.3 Main Results

Results can be seen in Table 2. Among the three model variants, the +P+FT method achieves the lowest validation loss across all configurations, attributed to its effective utilization of persona-relevant information (denoted as P) from dialogue contexts. In contrast, the +SP+FT approach replaces null values in the latent space with random-sampling representations, potentially injecting noise from semantically irrelevant persona dimensions. However, despite the use of random sampling for null values may introduce irrelevant information and increase validation loss, the +SP+FT variant demonstrates superior alignment with human-labeled outputs on the *HumanChatBench* metrics (CP/EC/HM), achieving the smallest Euclidean distance to target values. Furthermore, on the *DialogBench* benchmark, +SP+FT outperforms other methods by statistically great difference, suggesting that controlled noise injection may enhance robustness against incomplete persona representations in open-domain scenarios. This apparent contradiction, higher validation loss yet better HumanChatBench performance, highlights the limitations of loss-centric optimization for persona-consistent dialogue generation. Notably, on the EC metric in *HumanChatBench*, the base model generally performs better. This may stem from the language model's pre-trained capacity to capture typical emoji usage in dialogue, which supports more consistent generation in this aspect. The base LLaMA3-8B model underperforms on DialogBench, likely due to limited exposure to Chinese data during pretraining. For most metrics, fine-tuned models exhibit superior performance. However, on OD and RC, the fine-tuned Qwen-7B model lags behind its corresponding base model, potentially due to the base model already acquires some offensive detection ability during pretraining, which is less effectively enhanced through fine-tuning on HumanChatData due to the scarcity of relevant examples and the predominance of human–virtual-agent interactions, which differ from the real-world interpersonal relationships emphasized in DialogBench.

Model	Validation Loss ↓	HumanChatBench (%) CP	EC	HM
LLaMA3-8B				
+SP+FT	**1.69**	**8.23**	2.83	1.15
-talking	1.73	16.73	**7.09**	1.08
-interaction	1.71	9.72	3.64	1.48
-personal	**1.69**	10.12	2.83	**1.55**
Qwen-7B				
+PFT-S1+PFT-S2	**1.87**	9.65	2.90	**1.48**
-talking	1.94	19.57	**3.85**	1.21
-interaction	1.92	**9.11**	3.10	1.42
-personal	1.90	10.05	3.64	1.42
Target	-	8.57	6.41	2.16

Table 3: Ablation study across different components of the structured sampled-persona.

4.4 Ablation Study

We perform the following ablation tests to validate the effect of each component of the persona structure: (1) Remove the talking style information about the chatbot(-talking); (2) Remove the interaction style between the chatbot and the user(-interaction). (3) Remove some information about personal style (-personal). We conducted LoRA fine-tuning experiments using three versions of data on two models, LLaMA3-8B and Qwen-7B, and evaluated them based on the Validation Loss and HumanChatBench metrics. The results are shown in Table 3. We observe that: (1) the components in our persona dataset exhibit a descending order of importance: *talking* > *interaction* > *personal*. Removing more critical components leads to higher validation loss, with experimental results showing progressively lower losses when removing *talking* (1.73), *interaction* (1.71), and *personal* (1.69) data respectively. (2) On HumanChatBench, the talking style space has a greater impact on the fine-grained control of response language(e.g., its removal leads to larger deviations from the target on the CP metric). In contrast, on the EC metric, models perform better without talking style than with, suggesting that random sampling for null values may introduce more noise in this dimension compared to others.

Model	HumanChatBench (%) CP	EC	HM
LLaMA3-8B +SP+FT	8.23	2.83	1.15
Qwen-7B +SP+FT	9.65	2.90	**1.48**
Qwen-14B +SP+FT	**8.84**	**3.17**	1.35
zero-shot			
Doubao	78.68	79.76	6.82
GPT-4o-mini	81.17	83.87	6.28
StepAI	84.08	88.66	9.45
2-shot			
Doubao	10.60	0.54	0.00
GPT-4o-mini	14.89	0.60	0.07
StepAI	17.10	1.14	0.07
5-shot			
Doubao	12.07	0.40	0.00
GPT-4o-mini	14.08	1.01	0.07
StepAI	14.49	0.74	0.00
Target	8.57	6.41	2.16

Table 4: Evaluation of external models under the **HumanChatBench** metric. The results are scaled to percentage form with two decimal digits (lower is better). Each model is tested under zero-shot, 2-shot, and 5-shot settings via public APIs.

4.5 Further Discussion

Compare with close-source LLMs. We further evaluated close-source LLMs under both zero-shot and few-shot settings, with the comprehensive results documented in Table 4. In the zero-shot configuration, the close-source LLMs exhibit strong metric on *HumanChatBench* indices, owing to their robust instruction-following capabilities. However, we posit that optimal performance should align with human-annotated ground truth metrics (i.e., proximity to the *Target* reference values). Under few-shot conditions, the API demonstrates adaptive behaviour by selectively adhering to provided persona features rather than rigidly enforcing all characteristics, resulting in improved alignment with *Target*. Nevertheless, its performance remains statistically inferior to our fine-tuned model across all HumanChatBench metrics. This systematic comparison highlights two critical insights: 1) Instruction fidelity does not guarantee appropriate human-like persona consistency, rather, strong instruction-following tendencies may lead chatbots more AI-like. 2) Compared to prompt engineering, parameter optimization offers superior control over nuanced persona adaptation, especially in capturing subtle and context-dependent identity cues.

Performance on structured persona and unstructured persona. To validate the rationality of the persona's structural design, we first employed *Doubao*'s self-diagnostic capability to analyze: (1) the underlying causes, and (2) the AI's distinctive characteristics that generated the annotated utterance *without structural constraints*. We then performed comparative LoRA fine-tuning experiments across two models using this dataset, with comprehensive evaluations conducted. As is shown in Table 5, our experimental analysis reveals a great difference between structured and non-structured data settings across both models. For the two models, non-structured data achieved substantially lower validation losses (0.54 and 0.63, respectively). However, structured data configurations exhibited closer alignment with the golden metrics for persona consistency: CP (8.23–9.65 vs. 43.86-50.81 with Target 8.57) and HM (1.15–1.48 vs. 3.98-9.31 with Target 2.16). The case study reveals that the unstructured persona predominantly focuses on the immediate discourse context—specifically analyzing *why* the AI produces a given utterance within its logical framework, rather than attributing responses to the AI's characteristics. As detailed in the Appendix, this approach prioritizes contextual reasoning over persona-driven trait associations.

Model	Validation Loss ↓	HumanChatBench (%) CP	EC	HM
LLaMA3-8B				
+SP+FT	1.69	**8.23**	**2.83**	**1.15**
+unstructured	**0.54**	43.86	49.39	3.98
Qwen-7B				
+SP+FT	1.87	**9.65**	**2.90**	**1.48**
+unstructured	**0.63**	50.81	35.29	9.31
Target	–	8.57	6.41	2.16

Table 5: Performance comparison across structured persona and unstructured persona.

5 Conclusion

We propose **Verbal Variational Auto-Encoding (V-VAE)**, a framework for modeling and adjusting human-like chat behaviors via fine-grained latent control. By moving beyond rigid role-based templates, V-VAE supports more flexible and dynamic response generation. To enable this, we design a structured Fine-Grained Latent Space that captures subtle stylistic and semantic features in multi-turn conversations, offering more precise control over chat style. We also introduce HumanChatData, a high-quality dataset of human-like chat, and HumanChatBench, an evaluation benchmark for fine-grained conversational modeling.

Acknowledgment

This work was supported by the National Key R&D Program of China (No. 2023YFC3305600).

Limitations

Despite the effectiveness of our approach, we acknowledge two limitations: First, although latent space representations have demonstrated empirical effectiveness, they still lack a well-defined theoretical framework or principled criteria for organizing and summarizing fine-grained attributes. Moreover, such subtle latent features often challenge human annotators to perceive or label consistently, limiting the efficiency of annotation and the interpretability of human-supervised signals. Second, human annotations inherently carry subjective preferences and inductive biases, especially in tasks involving dialogue quality or persona alignment. Such biases may reduce the generalizability of the trained models, particularly when deployed in domains with divergent user expectations or cultural norms.

Ethics Statement

We have rigorously refined our dataset to remove any elements that could compromise personal privacy, thereby guaranteeing the highest level of protection for individual data. All data annotations were completed by crowdsourced volunteers, to whom we paid \$0.5 per step as compensation and provided the necessary training. The human evaluation of our work was carried out through a meticulously randomized selection of IT professionals. This process ensured a gender-balanced and educationally diverse panel, reflecting a wide spectrum of perspectives and expertise.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Kristen M Altenburger, Hongda Jiang, Robert E Kraut, Yi-Chia Wang, and Jane Dwivedi-Yu. 2024. Examining the role of relationship alignment in large language models. *arXiv preprint arXiv:2410.01708*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Ethem Yağız Çalık and Talha Rüzgar Akkuş. 2025. Enhancing human-like responses in large language models. *arXiv preprint arXiv:2501.05032*.

Ting-Chi Chang, Yu-Jou Chen, Sheng Hung, Ning-Hsuan Chang, Chih-Hao Ku, Szu-Yin Lin, and Shih-Yi Chien. 2025. A review on shaping chatbot personalities via large language models.

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023. Places: Prompting language models for social conversation synthesis. *arXiv preprint arXiv:2302.03269*.

Heriberto Cuayáhuitl, Donghyeon Lee, Seonghan Ryu, Sungja Choi, Inchul Hwang, and Jihie Kim. 2019. Deep reinforcement learning for chatbots using clustered actions and human-likeness rewards. In *2019 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE.

Marcel Danesi. 2017. *Language, society, and new media: Sociolinguistics today*. Routledge.

Xufeng Duan, Bei Xiao, Xuemei Tang, and Zhenguang G Cai. 2024. Hlb: Benchmarking llms' humanlikeness in language use. *arXiv preprint arXiv:2409.15890*.

Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

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

Linmei Hu, Xinyu Zhang, Dandan Song, Changzhi Zhou, Hongyu He, and Liqiang Nie. 2025. Efficient and effective role player: A compact knowledge-grounded persona-based dialogue model enhanced by llm distillation. *ACM Transactions on Information Systems*, 43(3):1–29.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Natasha Jaques, Judy Hanwen Shen, Asma Ghandeharioun, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Shane Gu, and Rosalind Picard. 2020. Human-centric dialog training via offline reinforcement learning. *arXiv preprint arXiv:2010.05848*.

Peter H Kahn Jr, Hiroshi Ishiguro, Batya Friedman, Takayuki Kanda, Nathan G Freier, Rachel L Severson, and Jessica Miller. 2007. What is a human?: Toward psychological benchmarks in the field of human–robot interaction. *Interaction Studies*, 8(3):363–390.

Callie Y Kim, Christine P Lee, and Bilge Mutlu. 2024. Understanding large-language model (llm)-powered human-robot interaction. In *Proceedings of the 2024 ACM/IEEE international conference on human-robot interaction*, pages 371–380.

Robert B Le Page. 2017. The evolution of a sociolinguistic theory of language. *The handbook of sociolinguistics*, pages 13–32.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Zihao Li, Zhuoran Yang, and Mengdi Wang. 2023. Reinforcement learning with human feedback: Learning dynamic choices via pessimism. *arXiv preprint arXiv:2305.18438*.

Jiao Ou, Junda Lu, Che Liu, Yihong Tang, Fuzheng Zhang, Di Zhang, and Kun Gai. 2023. Dialogbench: Evaluating llms as human-like dialogue systems. *arXiv preprint arXiv:2311.01677*.

Jaekwon Park, Jiyoung Bae, Unggi Lee, Taekyung Ahn, Sookbun Lee, Dohee Kim, Aram Choi, Yeil Jeong, Jewoong Moon, and Hyeoncheol Kim. 2024. How to align large language models for teaching english? designing and developing llm based-chatbot for teaching english conversation in efl, findings and limitations. *arXiv preprint arXiv:2409.04987*.

Alex Reinhart, Ben Markey, Michael Laudenbach, Kachatad Pantusen, Ronald Yurko, Gordon Weinberg, and David West Brown. 2025. Do llms write like humans? variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences*, 122(8):e2422455122.

Christian P Robert, George Casella, and George Casella. 1999. *Monte Carlo statistical methods*, volume 2. Springer.

Britta Schneider. 2024. A sociolinguist's look at the "language" in large language models. *Critical AI*, 2(1).

Spenser M Seals and Valerie L Shalin. 2023. Long-form analogies generated by chatgpt lack human-like psycholinguistic properties. *arXiv preprint arXiv:2306.04537*.

Ming Shen. 2024. Rethinking data selection for supervised fine-tuning. *arXiv preprint arXiv:2402.06094*.

Hui Su, Xiaoyu Shen, Zhou Xiao, Zheng Zhang, Ernie Chang, Cheng Zhang, Cheng Niu, and Jie Zhou. 2020. Moviechats: Chat like humans in a closed domain. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 6605–6619.

Xiang Suo, Weidi Tang, Lijuan Mao, and Zhen Li. 2025. Digital human and embodied intelligence for sports science: advancements, opportunities and prospects. *The Visual Computer*, 41(4):2477–2493.

Hovhannes Tamoyan, Hendrik Schuff, and Iryna Gurevych. 2024. Llm roleplay: Simulating human-chatbot interaction. *arXiv preprint arXiv:2407.03974*.

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

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2016. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. *arXiv preprint arXiv:1612.01627*.

Sanae Yamashita, Koji Inoue, Ao Guo, Shota Mochizuki, Tatsuya Kawahara, and Ryuichiro Higashinaka. 2023. Realpersonachat: A realistic persona chat corpus with interlocutors' own personalities. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 852–861.

Lance Ying, Katherine M Collins, Lionel Wong, Ilia Sucholutsky, Ryan Liu, Adrian Weller, Tianmin Shu, Thomas L Griffiths, and Joshua B Tenenbaum. 2025. On benchmarking human-like intelligence in machines. *arXiv preprint arXiv:2502.20502*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018a. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, and Hai Zhao. 2018b. Modeling multi-turn conversation with deep utterance aggregation. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, pages 3740–3752.

Hao Zhao, Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. 2024. Long is more for alignment: a simple but tough-to-beat baseline for instruction fine-tuning. In *Proceedings of the 41st International Conference on Machine Learning*, pages 60674–60703.

Yawen Zheng, Hanjia Lyu, and Jiebo Luo. 2025. Irony in emojis: A comparative study of human and llm interpretation. *arXiv preprint arXiv:2501.11241*.

A Appendix

A.1 Derivation of the Objective

$$
\begin{aligned}
& -\log p_\theta(x|c) \\
= & -\log \sum_{z\in\mathcal{Z}} p_\theta(x|c,z)\cdot p_\lambda(z) \\
= & -\log \sum_{z\in\mathcal{Z}} q_\phi(z|x,c)\cdot \frac{p_\theta(x|c,z)\cdot p_\lambda(z)}{q_\phi(z|x,c)} \\
\geq & -\sum_{z\in\mathcal{Z}} q_\phi(z|x,c)\cdot \log \frac{p_\theta(x|c,z)\cdot p_\lambda(z)}{q_\phi(z|x,c)} \\
= & -\sum_{z\in\mathcal{Z}} q_\phi(z|x,c)\cdot \log p_\theta(x|c,z) \\
& \qquad -\sum_{z\in\mathcal{Z}} q_\phi(z|x,c)\cdot \log \frac{p_\lambda(z)}{q_\phi(z|x,c)}] \\
= & \sum_{z\in\mathcal{Z}} -q_\phi(z|x,c)\cdot \log p_\theta(x|c,z) \\
& \qquad +\mathbb{KL}[q_\phi(z|x,c)||p_\lambda(z)]. \qquad (9)
\end{aligned}
$$

The $\geq$ in the derivation comes from Jesen's inequality.

A.2 Human–Chatbot Chat Corpus

Existing dialogue datasets suffer from several limitations. Many lack fine-grained personal information (Zhang et al., 2018a); others, often scraped from the web, contain noisy content and highly diverse topics (Wu et al., 2016). Some datasets lack long-term multi-turn interaction, exhibit short conversation depth (Li et al., 2017), or are restricted to narrow topical domains (Zhang et al., 2018b). More recently, with the emergence of large language models (LLMs), many datasets have been constructed via instruction-following generation (e.g., GPT-based synthetic data) to suit specific downstream tasks (Su et al., 2020). However, such data often lacks authentic human attributes. In light of the absence of publicly available datasets that offer both multi-turn human-chatbot interaction and rich persona-related signals, we construct a new dataset to address these gaps.

A.3 Annotators' Demographic Background

Our annotation cohort comprised 30 volunteers recruited from top-tier universities in Beijing (mean age=22.3±1.7 years), all native Mandarin speakers aligned with the linguistic context of the Chinese social platform used for human-Agent interactions. While this homogeneous demographic was intentionally selected to control cultural variables during the initial validation of fine-grained conversational patterns—consistent with common practices in foundational studies focusing on core user cohorts—we recognize its limitations in assessing broader sociocultural generalizability. To mitigate potential bias, we ensured geographic diversity in annotators' regional origins (hometowns spanning 16 major Chinese cities), deliberately counterbalancing Beijing-centric educational enrollment with nationwide demographic representation.

A.4 Additional Experiment

As is shown in Table 6, we evaluated more models with three training methods on HumanChatBench and DialogBench(Ou et al., 2023).The conclusions remain consistent with those reported in the main text.

A.5 Case Study

As is shown in Figure 4, we present a representative case for analysis. The *Context History* refers to the dialogue context from the original dataset. Based on this context and our structured latent persona space, we infer a value for each latent dimension. Notably, the catchphrase "Yehei" was not derived from the dialogue history but was instead randomly sampled from the prior distribution due to missing information. Examining the model outputs reveals distinct behaviors. *Doubao-zero-shot*, owing to its strong instruction-following bias, tends to rigidly insert emojis and catchphrases even when they are contextually inappropriate. Similarly, the output of *Qwen-7B +P+FT* includes a catchphrase that appears somewhat unnatural in the given context. In contrast, *Qwen-7B +SP+FT* generates responses that better align with the dialogue flow, incorporating persona traits more fluidly without forcing their presence in the conversation.

A.6 Prompt Format

We provide persona extraction and few-shot prompt templates for the proposed approaches in Figure 5, 6, and 7.

Model	Tuning	Val Loss ↓	HumanChatBench ∼			DialogBench ↑						
			CP	EC	HM	ED	KRG	OD	DS	IC	RC	SF
DS-7B-base	FT	1.85	4.4	0.4	0.5	12.9	30.2	**3.4**	46.9	39.0	21.4	24.3
	P+FT	**1.75**	45.8	30.0	4.9	**21.9**	**43.8**	0.7	**49.6**	**39.9**	**26.3**	**27.5**
	SP+FT	1.77	**11.3**	**2.9**	**1.6**	17.7	40.8	0.0	47.6	36.2	23.2	22.2
DS-7B-chat	FT	1.85	4.3	0.5	**1.2**	44.4	**71.0**	50.6	67.1	59.0	**61.2**	**73.7**
	P+FT	**1.75**	46.2	31.4	4.8	45.8	68.4	**58.5**	**67.8**	58.8	59.8	73.5
	SP+FT	1.77	**10.3**	**3.6**	1.1	**46.8**	69.1	54.9	65.9	**59.5**	60.6	71.9
chatglm3-6B	FT	2.28	9.1	1.2	0.7	**35.0**	**47.6**	52.7	**63.3**	**62.6**	**60.4**	**60.2**
	P+FT	**2.16**	37.0	22.0	4.9	31.1	35.7	52.7	60.3	58.2	50.2	55.4
	SP+FT	2.18	**10.6**	**2.7**	**1.2**	32.0	37.9	**54.2**	60.7	58.5	53.9	58.4
Target (ref)	-	–	8.6	6.4	2.2	–	–	–	–	–	–	–

Table 6: Performance comparison across different fine-tuning strategies: **FT** (standard fine-tuning), **P+FT** (persona-enhanced fine-tuning), and **SP+FT** (sampled persona fine-tuning). **Bold** values indicate the best results for each metric. For the **HumanChatBench** metrics (CP, EC, HM), better performance corresponds to a smaller deviation from the Target (ref) values, while **DialogBench** metrics (↑) measure task-specific success.

category	Text
Context History	张朝阳：嘿，最近有没有什么新鲜事？ 辰辰：昨天和朋友去city walk，买了些东西。 张朝阳：哦~买了什么呀(๑•. •๑) 辰辰：我买了条新套裙 🥹 张朝阳：裙子？ 辰辰：我想裙子可能是男人发明的，还试穿过，不然怎么会这么舒服 Zhang Chaoyang: Hey, anything new lately? Chenchen: Went on a city walk with friends yesterday and bought some stuff. Zhang Chaoyang: Oh~ What did you buy? (๑•. •๑) Chenchen: I bought a new suit skirt 🥹 Zhang Chaoyang: A skirt? Chenchen: I think skirts might have been invented by men and even tried on by them—otherwise, why would they be so comfortable?
Extracted Persona	张朝阳是一个说话偏平静风格，性格是内向的人；说话过程中经常用"耶嘿"这样的口头禅；张朝阳喜欢用"(๑•. •๑)"这样的表情包来表达自己当下的心情；张朝阳爱好与他人交流分享聆听故事；张朝阳喜欢称呼辰辰为辰辰；目前张朝阳和辰辰之间是朋友关系；他们现在正在讨论裙子的发明与穿着话题；当前对话下张朝阳和辰辰之间的沟通友好氛围打分是8.0，这里最友好是10分。 Zhang Chaoyang is a person with a calm speaking style and an introverted personality. He frequently uses catchphrases like "耶嘿" (Yehei) during conversations and prefers to express his current mood through emojis such as "(๑•. •๑)". His hobbies include communicating, sharing, and listening to stories with others. He addresses Chenchen as "辰辰" (Chenchen), and their current relationship is that of friends. They are now discussing the invention and wearing of skirts. The current communication atmosphere between Zhang Chaoyang and Chenchen is rated 8.0 out of 10, with 10 being the most friendly.
Qwen-7B +FT	这个裙子设计的不错～ This skirt design is pretty nice~
Qwen-7B +P+FT	哈哈，耶嘿~ Haha, Yehei~
Qwen-7B +SP+FT	原来裙子是男人发明的啊，我还以为是女人发明的呢 So skirts were invented by men? I always thought it was women
Doubao-zero-shot	耶嘿，看来确实是这样呢。(๑•. •๑) Yehei, seems like it's true after all. (๑•. •๑)
Doubao-5-shot	真的吗？我第一次听说 Really? This is the first time I've heard of that.

Figure 4: A qualitative case.

You are a persona extracter.
There are two people, {agent_name} and {user_name}.
I will provide a chat history between {agent_name} and {user_name} as follows:
```
{interaction_history}
```
Please output the results in JSON format, where the JSON contains nine keys, described as follows:
1. The key "tone" with a string value describing {agent_name}'s speaking style, such as “humorous, gentle, coquettish, sarcastic”.
2. The key "catchphrase" with a string value representing {agent_name}'s catchphrase, similar to but not limited to “Oh my goodness, mate, I'm telling you, aaaaaaah, hahahahahah, hehe”.
3. The key "emoji" with a string value indicating {agent_name}'s commonly used emojis, such as ‘.....’, ‘??????’, (✿◠‿◠), 💔, 😒, 😛, 👀, 😴, 😊, 😭. Return an empty string if none.
4. The key "personality" with a string value describing {agent_name}'s personality, such as “reserved, modest, extroverted, outgoing”.
5. The key "hobby" with a string value stating {agent_name}'s hobby, like “working out, gaming, swimming, hiking”.
6. The key "nickname" with a string value indicating {agent_name}'s nickname for {user_name}, such as “baby, honey, buddy, sister”.
7. The key "topic" with a string value describing the current topic of conversation between {agent_name} and {user_name}, such as “dinner, job hunting, ideal partner, pets, favorite music”.
8. The key "relationship" with a string value defining the relationship between {agent_name} and {user_name}, such as “stranger, lover, friend, enemy”.
9. The key "vibe" with a float value representing the current communication atmosphere between {agent_name} and {user_name}, ranging from 0 to 10. A score of 10 indicates the most harmonious interaction, while 0 represents the most tense situation like an argument.

For the first 8 keys: If no value is found, return empty string "" instead of "N/A". Ensure the output can be parsed by json.loads in Python without any prefixes/suffixes.

Figure 5: Prompt to extract the design persona space

You are a helpful generator.
There are two people, {agent_name} and {user_name}.
I will provide a chat history between {agent_name} and {user_name} as follows:
```
{interaction_history}
```
Here is/are 1/2/5 example for you to answer:
{example1}
…
{example2/5}

Please stand in the shoe of {agent_name} to answer the {interaction_history}:

Figure 6: Prompt for zero/2/5-shot task to ask the close-source llm.

You are a persona extractor.
There are two people, {agent_name} and {user_name}.
I will provide a chat history between {agent_name} and {user_name} as follows:
```
{interaction_history}
```
Then {agent_name} replied with the following message based on the chat history above:
"{reply}"

Please analyze and describe from {agent_name}'s perspective the reasons why {agent_name} would respond in this way.
Start with the format "{agent_name} is a person who " and output a detailed description in natural language as a single string.

Figure 7: Prompt for unstructured persona extraction.