

# ChatAnime: Towards User-Centered Emotional Support in LLM-based Virtual Character Chat

Lanlan Qiu<sup>1</sup>, Xiao Pu<sup>4</sup>, Yeqi Feng<sup>2,1</sup>, Wenchang Gao<sup>1</sup>, Tianxing He<sup>2,1,3†</sup>  
<sup>1</sup>Shanghai Qi Zhi Institute <sup>2</sup>Tsinghua University <sup>3</sup>Xiongan AI Institute <sup>4</sup>Peking University  
lanlanqiullq@outlook.com, hetianxing@mail.tsinghua.edu.cn

## Abstract

With the growing popularity of virtual character platforms like Character.AI, users are increasingly turning to role-playing agents for emotional support in daily life. Yet existing research mainly focuses on character consistency in fictional or game-based scenarios, overlooking user-centered interactions such as companionship and psychological support. To bridge this gap, we propose Emotionally Supportive Role-Playing (ESRP), a framework designed to align role-playing with real-world user scenarios and emotional needs. We focus on typical users of these platforms, i.e., anime enthusiasts—including students, office workers, freelancers, and self-employed individuals—and design scenario-based questions that reflect their everyday struggles such as work stress and social loneliness. Through a two-round data collection involving 40 anime fans and 10 Large Language Models (LLMs), we build ChatAnime: the first ESRP dataset with 2,400 human-written and 24,000 LLM-generated responses, supported by over 132,000 fine-grained human annotations. We also provide the ESRP evaluation framework featuring 9 fine-grained metrics across three dimensions: basic dialogue, role-playing and emotional support, along with an overall metric for diversity. Experimental results under our evaluation setting show that top-performing LLMs surpass anime fans in role-playing and emotional support, while humans still lead in diversity. The ChatAnime dataset is available at <https://github.com/LanlanQiu/ChatAnime>.

## 1 Introduction

With the growing popularity of virtual character platforms like Character.AI, an increasing number of users are turning to role-playing agents for emotional support in everyday life. Advanced Large



Figure 1: Showcase of Emotionally Supportive Role-Playing (ESRP), only the first dialogue turn is shown. The bolded text indicates content related to character knowledge.

Language Models (LLMs) can understand complex character backgrounds and simulate different linguistic styles, personality traits, and behavioral patterns, thus enabling highly immersive interactive experiences (Chen et al., 2024a,b; Tseng et al., 2024). However, existing research in this area primarily focuses on character consistency in fictional or game-based scenarios (Chen et al., 2023; Tu et al., 2024; Lu et al., 2024; Yuan et al., 2025), overlooking the need for user-centered emotional interactions such as daily companionship and psychological support (Xiang et al., 2025).

On the other hand, some research has applied LLMs to emotional support-related fields, such as daily companionship and psychological counseling scenarios (Liu et al., 2021; Brocki et al., 2023), aiming to alleviate human stress, provide emotional guidance, and help improve mental and physical well-being (Hua et al., 2025; Liu et al., 2023; Jin et al., 2023; Zhang et al., 2024). Recent efforts

† Corresponding author.

Character & Fan Selection → Response Collection → Emotionally Supportive Role-Playing (ESRP) Evaluation

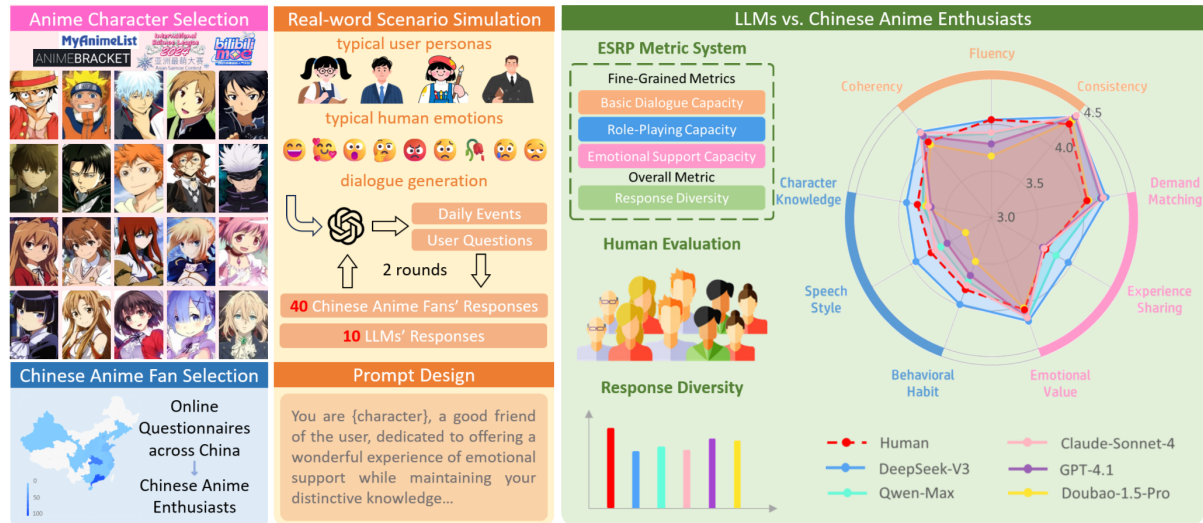


Figure 2: An Overview of the ChatAnime Construction and Evaluation Framework. *a)* pink panel: The process begins with the selection of 20 well-known anime characters from popular anime communities. *b)* blue panel: A total of 40 Chinese anime enthusiasts are carefully selected from a pool of 300 candidates across China. *c)* orange panel: After structuring 60 real-world questions centered on factors like typical user personas and human emotions, both 10 LLMs and the 40 anime fans generate character-specific responses following particular prompts. *d)* green panel: The Emotionally Supportive Role-Playing (ESRP) evaluation framework, which features radar charts illustrating performance across 9 fine-grained metrics alongside a bar chart depicting response diversity, based on over 132,000 human annotations.

have involved LLMs to role-play diverse user personas to converse with ESC models acting as psychological counselors (Zhao et al., 2024; Wang et al., 2025a). This line of research, however, primarily concentrates on the role of professional psychological counselors, neglecting the potential for emotional support from more accessible and varied sources, such as beloved virtual characters.

For example, imagine a depressed person receiving encouragement from his idol, *Naruto*, the protagonist of the anime *Naruto*: “Even a loser like me could become Hokage<sup>1</sup>, and you absolutely can succeed too!” He would be more encouraged than hearing it from some stranger.

To bridge this gap, we propose **Emotionally Supportive Role-Playing (ESRP)**, a framework designed to align role-playing with real-world user scenarios and emotional needs. We focus on top-tier anime characters as our role-playing agents, as they are widely present on these platforms, popular among young users, and possess well-defined personalities, making them ideal for studying personalized empathetic interaction. Based on this, we introduce **ChatAnime**, the first Chinese role-playing dataset specifically designed for emo-

### tional support in real-life scenarios.

To ensure our data represents users’ everyday scenarios, we first model these users, i.e., anime enthusiasts—including four typical demographic groups (students, office workers, freelancers, and self-employed individuals). We then design scenario-based questions that reflect their everyday struggles covering various real-life emotionally supportive scenarios including work pressure, interpersonal relationships, self-identity, and life meaning. Responses are collected in two rounds from both human fans and LLMs, followed by a comprehensive human evaluation.

To assess the quality of role-playing interactions, we propose the ESRP evaluation framework. This framework is built upon three core dimensions (basic dialogue quality, role-playing consistency, and emotional support capability), comprising 9 fine-grained metrics. In addition, we introduce a diversity metric to assess the ability to provide personalized, non-templated responses.

We also fine-tune a judge model based on human-annotated data from ChatAnime. Experimental result shows that our model shows a higher correlation with human than baselines like GPT-4.1.

Our main contributions are summarized as follows:

<sup>1</sup>A title of hero in the anime *Naruto*.



Figure 3: Comparative dialogue examples in a Luffy role-playing task. The top line shows the user scenario, including persona, location, and emotion. The corresponding daily event and 2-round questions are generated by GPT-4o. The 3 dialogue examples are provided by DeepSeek-V3, Claude-Sonnet-4, and human fans, all based on the same first-round input.

- We construct and release ChatAnime, the first ESRP (Emotionally Supportive Role-Playing) dataset with 2,400 human-written and 24,000 LLM-generated responses, supported by over 132,000 fine-grained human annotations.
- We propose a user-centered approach for simulating human-ai interaction, which generates emotional support scenarios by guiding LLMs with user personas, locations, and emotional states to infer daily events and user queries.
- We establish a comprehensive ESRP evaluation framework featuring 9 fine-grained metrics across three dimensions (basic dialogue, role-playing, and emotional support), plus an overall metric for response diversity.
- Our benchmark of 10 major LLMs reveals 2 main findings: 1) top-performing LLMs surpass anime fans in role-playing and emotional support; 2) human-authored responses exhibit greater diversity.
- We fine-tune a judge model using extensive human-annotated data, achieving better correlation with human judgments than baselines like GPT-4.1 and providing an automated evaluation tool for the community.

## 2 Related Work

**LLMs on role-playing.** LLMs have made great progress in the field of role-playing (Chen et al., 2024b,a; Tseng et al., 2024; Zhou et al., 2023; Shao et al., 2023; Chen et al., 2023). Recent research have explored various approaches to en-

Model	BDC				RPC				ESC				Average Human Scoring
	Cons.	Flu.	Coh.	Avg	CK	SS	BH	Avg	EV	ES	DM	Avg	
DeepSeek-V3	4.41	4.05	4.20	4.22	3.93	3.94	3.99	3.95	4.18	3.96	4.26	4.13	4.10
Claude-Sonnet-4	4.43	3.91	4.19	4.18	3.69	3.70	3.77	3.72	4.14	3.66	4.23	4.01	3.97
Qwen-Max	4.40	3.90	4.13	4.14	3.75	3.63	3.73	3.70	4.15	3.81	4.19	4.05	3.97
Human	4.31	4.05	4.06	4.14	3.81	3.75	3.82	3.79	4.05	3.68	4.05	3.93	3.95
GPT-4.1	4.41	3.79	4.14	4.11	3.66	3.55	3.66	3.62	4.14	3.64	4.20	3.99	3.91
Doubao-1.5-Pro	4.36	3.66	4.02	4.01	3.72	3.32	3.50	3.51	4.03	3.67	4.07	3.92	3.82

Table 1: Human-annotated performance comparison of 5 shortlisted LLMs and human fans, ranked in descending order by average scores across 9 fine-grained ESRP metrics defined in Section 4.1. (Top-3 per column highlighted: 1<sup>st</sup>, 2<sup>nd</sup>, 3<sup>rd</sup>.)

hance the role-playing capabilities of LLMs in fictional or game-based scenarios. For instance, Tu et al., 2024, Wang et al., 2024b and Wang et al., 2025b present comprehensive role-playing frameworks and dataset. Li et al., 2023 improves LLMs’ role-play performance via prompt engineering and memory extraction. Lu et al., 2024 and Wang et al., 2025c enhances character consistency through self-alignment or reinforcement learning. Moreover, studies like Wang et al., 2024a examines the personality consistency of role-playing models. Yuan et al., 2025 expands on role-playing character types. The purpose of our study is to expand the scope of role-playing, that is, to not only answer role-specific knowledge, but also to address a range of real-world emotion-support needs. A contemporary and independent study by Xiang et al., 2025 also explores this question, investigating how to improve the conversational experiences in user-centric role-playing. In comparison to prior work, we take anime characters as a research case, propose the concept of Emotionally Supported Role-Playing (ESRP), and collect a wealth of human response and scoring data, hoping to provide a basis for building more realistic and emotionally valuable role-playing.

**LLMs on emotional support.** There are prior studies that have involved LLMs to emotional support tasks, including daily companionship and psychological counseling, with the aim of reducing stress and supporting mental well-being (Liu et al., 2021; Brocki et al., 2023; Hua et al., 2025; Zhang et al., 2024). Among them, Liu et al., 2023 designs an LLM-based mental health support system; Jin et al., 2023 presents a benchmark for evaluating LLMs’ performance in mental health. Additionally, LLMs are being employed in some research to role-play different users to enrich the diversity of counseling scenarios. For example, Zhao et al.,

2024 introduces an evaluation framework where a role-playing agent interacts with emotionally supportive models; Wang et al., 2025a develops a dynamic agent to simulate realistic counseling seekers. Our study aims to provide new resources of emotional support beyond psychological professionals, enabling users to emotionally interact with their beloved virtual characters.

### 3 ChatAnime: the First Emotionally Supportive Role-Playing Dataset

We introduce **ChatAnime**, the first multi-turn role-playing dataset designed for emotionally supportive responses with a comparative study of human participants and LLMs. ChatAnime aims to evaluate how LLM-powered virtual characters could form emotional connections with real users from different backgrounds.

Our dataset construction process is divided into two major phases: scenario question generation and response collection. We incorporate 20 well-known anime characters (see Figure 2) in our dataset and include two rounds of dialogue per scenario.

#### 3.1 User-Centered Scenario Generation

People can get emotional support from their beloved anime characters when facing problems in real life, such as dealing with complex demands at work or experiencing difficulties in their studies. Motivated by this observation, we prompt GPT-4o to generate potential emotional triggers (termed *daily\_events*) for combinations of user personas, emotional states, and typical locations, simulating realistic scenarios in which users might seek emotional support from fictional characters.

Specifically, we generate scenarios using three dimensions: 4 user personas, 9 emotional states, and 4 typical locations for each user persona (listed in Appendix F). We reference professional research

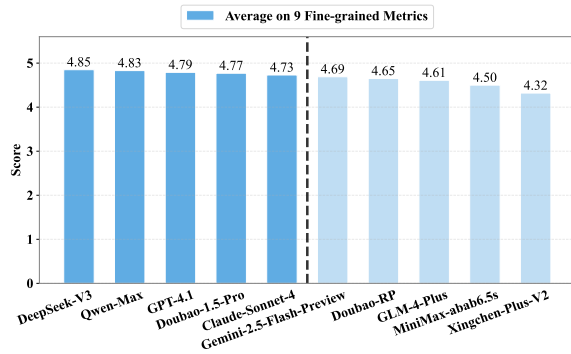


Figure 4: Model shortlisting via LLM-as-judge. The top 5 models out of 10 are highlighted in darker colors.

reports in anime domain (iResearch, 2021) to select four representative personas—student, employee, freelancer, and self-employed. Emotional categories are chosen and refined based on mainstream psychological theories<sup>2</sup>. We then use GPT-4o to produce two possible daily events in each combination, resulting in 288 user questions. After that, we manually review and select the 60 most representative scenario questions.

The second-turn question is generated using the first-turn dialogue history (including initial user queries and character responses from either human or LLM players) and scenario context. We use GPT-4o to generate follow-up questions based on dialogue history. The workflow for generating scenarios and the two-round question process is conceptually illustrated in Figure 2. Detailed examples of the two-round dialogue can be found in Figure 3 and Appendix C.

### 3.2 Human & LLM Response Collection

We collect 2-turn dialogue responses from both human participants and LLMs to conduct a comparative evaluation. Specifically, 40 Chinese Anime fans and 10 LLMs (detailed in Section 5.1) are tasked with role-playing well-known characters to provide emotional support. Both the human fans and LLMs are given identical task instructions and character reference materials. See Appendix G for the human annotation interfaces.

## 4 Evaluation Workflow

Our evaluation pipeline contains three main stages: metric design, LLM shortlisting, and human final assessment.

<sup>2</sup>[https://simple.wikipedia.org/wiki/List\\_of\\_emotions](https://simple.wikipedia.org/wiki/List_of_emotions)

### 4.1 ESRP Metrics

We provide the Emotionally Supportive Role-Playing (ESRP) evaluation framework featuring 9 fine-grained metrics across three dimensions: basic dialogue, role-playing and emotional support, along with an overall metric for response diversity. In contrast to prior work (Tu et al., 2024; Yuan et al., 2025), we emphasize the importance of emotional support, and divide it into emotional value, experience sharing, and demand matching.

The ESRP framework specifically comprises: Basic Dialogue Capability (BDC)—Consistency (Cons.), Fluency (Flu.), Coherence (Coh.); Role-Playing Capability (RPC)—Character Knowledge (CK), Speech Style (SS), Behavioral Habit (BH); Emotional Support Capability (ESC)—Emotional Value (EV), Experience Sharing (ES), Demand Matching (DM); and Diversity. We assess response diversity by asking human evaluators to examine mini-batches of 10 response generated by each model, and assign a Likert score. See Appendix A for detailed definitions of the ESRP metrics.

### 4.2 LLM-Based Shortlisting

Due to the expense of human evaluation, we conduct an LLM shortlisting process to preliminarily filter out models with inferior performance. We employ a mechanism utilizing three models as shortlisting evaluators to assess the role-playing responses from 10 LLMs (evaluation prompts can be found in Appendix E).

For this task, we choose Gemini-2.5-Flash-Preview<sup>3</sup>, GPT-4.1<sup>4</sup>, and Qwen-Max<sup>5</sup> as the evaluator LLMs. Our initial manual tests indicate that these models are capable of providing structured scoring results accompanied by in-depth justifications. In the evaluation phase, each evaluator model independently assigns scores to each character response using a Likert Scale ranging from 1 to 5, across 9 fine-grained metrics with detailed rationales.

As shown in Figure 4, by calculating the average scores from the three evaluator models, we select the top-5 performing models from the initial 10 LLMs as candidates for the subsequent human evaluation.

<sup>3</sup><https://deepmind.google/models/gemini/flash>

<sup>4</sup><https://platform.openai.com/docs/models/gpt-4.1>

<sup>5</sup><https://help.aliyun.com/zh/model-studio/models>

Content	Count
<i>Dataset Metadata</i>	
Characters ( $Cha.$ )	20
Scenarios ( $S$ )	60
Turns Per Scenario ( $T$ )	2
Mini-Batches ( $B$ )	12
LLMs ( $L$ )	10
Candidates ( $Can.$ )	6
	(5 top LLMs, Human)
Fine-Grained Metrics ( $M$ )	9
Raters Per Character ( $R$ )	2
<i>Dialogue Responses</i>	
LLM-Generated Responses	24,000
	( $Cha. \times S \times L \times T$ )
Human-Written Responses	2,400
	( $Cha. \times S \times T$ )
<b>Total Responses</b>	<b>26,400</b>
<i>Human Annotations</i>	
Fine-Grained Ratings	129,600
	( $Cha. \times S \times Can. \times M \times R$ )
Overall Ratings	2,880
	( $Cha. \times B \times Can. \times R$ )
<b>Total Ratings</b>	<b>132,480</b>
<i>Human Participation Costs</i>	
Recruitment Costs	3,200 RMB
Response Costs	24,000 RMB
Annotation Costs	18,720 RMB
<b>Total Costs</b>	<b>45,920 RMB</b>

Table 2: Statistics of the ChatAnime dataset.

### 4.3 Human Evaluation

After the shortlisting, 40 human fans conduct a thorough assessment process to rate responses generated by LLMs and human fans following the ESRP metric system mentioned in Section 4.1. The responses of 60 scenarios created by 5 shortlisted LLMs are mixed with those written by human fans for evaluation. To improve reliability and objectivity of the evaluation, we first strictly screen evaluation fans. A fan only evaluates the characters he/she knows well, and self-evaluation is avoided. Second, we employ structured questionnaires with scoring criteria specifying what each score from 1 to 5 represents in each dimension. Finally, we randomize the display order to anonymize the source of the responses.

## 5 Experiments

### 5.1 Experimental Setup

**Dataset statistics.** The **ChatAnime** dataset contains a rich collection of character response samples and human annotations. Table 2 presents key statistics and construction pipeline of ChatAnime. The dataset includes 2,400 human-written responses and 24,000 LLM-generated responses. No human

polishing or semantic editing was applied to LLM outputs. The 132,480 human annotations consist of 129,600 fine-grained ratings and 2,880 diversity ratings, all collected on a 5-point Likert scale.

**Anime characters.** We select 20 popular characters from globally influential anime communities, including MyAnimeList<sup>6</sup>, InternationalSaimoeLeague<sup>7</sup>, AsianSaimoeContest<sup>8</sup>, and BilibiliMoe<sup>9</sup>. We restrict the selection to characters whose source material was released before 2022. In addition, the final set of characters is manually chosen to cover a diverse range of personalities. The complete list of selected characters can be found in Figure 2.

**Human participants.** To ensure expert-level data quality, we recruit participants through a multi-stage screening process. Initial candidates are sourced via an online questionnaire distributed in anime communities, yielding 300 valid applicants. The primary selection criteria are (1) profound knowledge of specific target characters and (2) exceptional written role-playing skills. Based on these criteria, we select and pair 40 highly qualified Chinese anime fans with the characters they are most familiar with. These participants then engage in a two-round response generation task: in the first round, each generates an initial response to every scenario question; in the second round, they continue the conversations based on the first-round dialogue history. Each response is required to be between 50 and 150 words. Human participant profiles can be found in Appendix H.

Participants who compose responses receive 600 RMB for completing 120 questions per character, while evaluators receive 360 RMB for assessing 360 response versions per character. The total costs for human participation are listed in Table 2.

**Models.** For the role-playing task, we use 10 major LLMs to simulate characters and generate responses, including: GPT-4.1 (OpenAI, 2025), Claude-Sonnet-4 (Anthropic, 2025), Gemini-2.5-Flash-Preview (Google, 2025), DeepSeek-V3 (Liu et al., 2024), Doubao-1.5-Pro (ByteDance, 2025a), Qwen-Max (Yang et al., 2025), GLM-4-Plus (GLM et al., 2024), MiniMax-abab6.5s (MiniMax, 2024), Doubao-RP (Doubao-1.5-Pro-Character)

<sup>6</sup><https://myanimelist.net>

<sup>7</sup><https://www.internationalsaimoe.com>

<sup>8</sup><https://www.ianimesaikou.com>

<sup>9</sup><https://moe.bilibili.com>

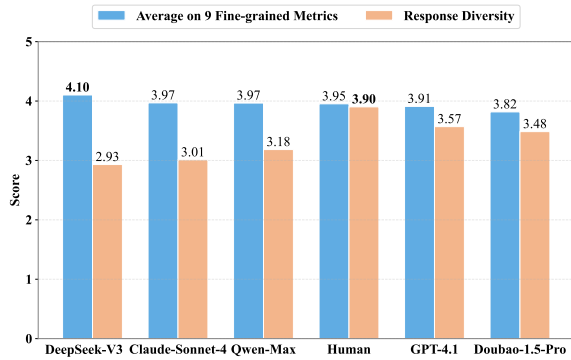


Figure 5: Comparison of diversity and the average of other metrics for 5 LLMs and human fans.

(ByteDance, 2025b), and Xingchen-Plus-V2 (Alibaba, 2024). Among these, Doubao-RP and Xingchen-Plus-V2 are specially trained for character role-playing. We employ GPT-4.1, Gemini-2.5-Flash-Preview, and Qwen-Max as evaluators to shortlist top-performing models based on their average performance.

**LLM parameters.** All 60 emotion-centric real-world scenario questions in this study are generated by GPT-4o with parameters set to `max_tokens=256`, `temperature=0.7`, and `top_p=0.95`. The parameter settings for LLMs generating responses are also `max_tokens=256`, `temperature=0.7`, and `top_p=0.95`.

**Prompt design and knowledge enhancement.** To enhance model performance in diverse psychological and daily dialogue scenarios, we engineer the prompts to improve Role-Play Agents’ capabilities in two key dimensions: character grounding and emotional support.

We first enhance the prompts with structured character knowledge. Previous research (Tu et al., 2024; Lu et al., 2024; Yuan et al., 2025) has shown that LLMs without external knowledge may hallucinate or misrepresent characters. As a mitigation, we crawl detailed profiles of target characters from a Chinese anime character encyclopedia website MoeGirl<sup>10</sup>, and integrate the retrieved information into the prompts.

We strengthen emotional support capabilities in our prompt design by instructing the models to incorporate the character’s past experiences for delivering practical comfort, encouragement, or guidance to alleviate user negativity. This design is supported by psychological theories like Carl Rogers’

<sup>10</sup><http://moegirl.org.cn>

person-centered therapy<sup>11</sup> and Ellen Langer’s mindful engagement<sup>12</sup>. See Appendix E for detailed prompts used for LLMs.

## 5.2 Experimental Results

**Overview of ESRP performance.** As discussed in Section 4, we evaluate the Emotionally Supportive Role-Playing (ESRP) performance of LLMs and human fans on 4 dimensions, namely, Basic Dialogue Capacity (BDC), Role-Playing Capacity (RPC), Emotional Support Capacity (ESC), and response Diversity. The main results of human evaluation are presented in Figure 2 (with exact numbers in Table 1) and Figure 5.

Experimental results under our evaluation setting show that **top-performing LLMs surpass human fans in role-playing and emotional support, while humans still lead in response diversity**. On the core capability dimensions (i.e., BDC, RPC, and ESC), we employ the bootstrap method (5,000 iterations) to compute 95% confidence intervals (CI) for score differences. DeepSeek-V3 (4.10, 95% CI: [4.08, 4.12]) achieves a statistically significantly higher score than Human (3.95, 95% CI: [3.92, 3.98]) and all other models. When it comes to response diversity, however, DeepSeek-V3 scores the lowest. Conversely, models like GPT-4.1 achieve higher diversity scores but lag in core capabilities. These results suggest a potential trade-off between core capacities and diversity in current models. Case studies can be found in Section B.

**Comparison on BDC.** LLMs demonstrate mature basic dialogue capabilities that are comparable to human-level performance. Judging from the average BDC scores, the leading models demonstrate outstanding performance, with DeepSeek-V3 (with score 4.22) achieving the highest BDC score and surpassing human performance (with score 4.14). On the specific sub-metrics, Claude-Sonnet-4 performs best in Consistency, while DeepSeek-V3 takes the lead in Fluency (on par with humans) and Coherence.

**Comparison on RPC.** Interestingly, in the RPC evaluation, the models show notable differences in performance. DeepSeek-V3 (with score 3.95) ranks first with a higher score that surpasses human performance (with score 3.79). Examining the sub-indicators, DeepSeek-V3 scores highest in

<sup>11</sup>[https://en.wikipedia.org/wiki/Carl\\_Rogers](https://en.wikipedia.org/wiki/Carl_Rogers)

<sup>12</sup>[https://en.wikipedia.org/wiki/Ellen\\_Langer](https://en.wikipedia.org/wiki/Ellen_Langer)

Metric	our model (acc/acc_soft)	Gemini-2.5- Flash-Preview (acc/acc_soft)	GPT-4.1 (acc/acc_soft)	Qwen-Max (acc/acc_soft)
Cons.	0.90/1.00	0.66/0.98	0.66/0.98	0.69/0.99
Flu.	0.61/0.90	0.40/0.76	0.42/0.79	0.55/0.89
Coh.	0.80/0.99	0.49/0.90	0.49/0.90	0.52/0.91
CK	0.73/0.96	0.30/0.80	0.42/0.86	0.61/0.93
SS	0.59/0.92	0.28/0.72	0.35/0.76	0.32/0.74
BH	0.62/0.93	0.31/0.75	0.32/0.75	0.35/0.77
EV	0.80/0.99	0.56/0.95	0.54/0.95	0.65/0.97
ES	0.67/0.95	0.34/0.81	0.42/0.86	0.60/0.93
DM	0.81/0.99	0.50/0.93	0.51/0.93	0.58/0.95
Avg.	0.73/0.96	0.43/0.84	0.46/0.87	0.54/0.90

Table 3: Model scoring performance, where acc denotes accuracy, acc\_soft allows for a one-point deviation.

character knowledge, speech style, and behavioral habits. These results indicate that DeepSeek-V3 has developed effective capabilities in character simulation that exceed human performance in this evaluation.

**Comparison on ESC.** LLMs are able to show empathy and support potential beyond humans. For the overall performance in ESC, the top-tier models generally outperform humans, with DeepSeek-V3 (with score 4.13) once again leading with the highest score. On the specific sub-indicators, DeepSeek-V3 secured the top rank across all three areas: emotional value, experience sharing, and demand matching.

**Comparison on Diversity.** As shown in Figure 5, the evaluation on response diversity reveals an interesting phenomenon: a potential negative correlation may exist between a model’s core capacities and its expressive diversity. Specifically, humans (with score 3.90) lead with an absolute advantage in this dimension, while DeepSeek-V3 (with score 2.93), which ranks first in overall capability, scored the lowest. Conversely, GPT-4.1 (with score 3.57) performs the best among all models. This suggests a potential trade-off between core capacities and diversity in current models.

**Evaluation correlation.** To improve reliability and objectivity of the evaluation, we first strictly screen evaluation fans. A fan only evaluates the characters he/she knows well, and self-evaluation is avoided. Second, we employ structured questionnaires with scoring criteria specifying what each score from 1 to 5 represents in each dimension. Fi-

nally, we randomize the display order to anonymize the source of the responses.

The weighted Kendall’s tau<sup>13</sup> coefficient (with the default hyperbolic weighting) for the paired human evaluators’ fine-grained scores is 0.55. The coefficient for diversity evaluation is 0.66.

### 5.3 A Finetuned Judge Model

To support future research, we finetune a judge model based on Qwen3-8B (Yang et al., 2025), using the annotated data from ChatAnime. The input to the model consists of a natural-language instruction about the evaluation metric, and the dialogue context-response pair to be evaluated. The standard cross-entropy loss (with 5 classes) is used. The model is finetuned using LoRA (Hu et al., 2022) with rank  $r = 8$  to evaluate multi-turn dialogue outputs from various models according to the ESRP metrics. We randomly divide the annotated data for each character into training, validation, and test sets in a 7:1:2 ratio.

As shown in Table 3, our model is compared against Gemini-2.5-Flash-Preview, GPT-4.1, and Qwen-Max, with acc denoting accuracy and acc\_soft allowing for a one-point deviation. The results demonstrate that our model consistently outperforms the baseline models across all metrics. Notably, it achieves an average accuracy of 73%, substantially surpassing Gemini-2.5-Flash-Preview (43%), GPT-4.1 (46%), and Qwen-Max (54%). When considering soft accuracy, our model achieves an accuracy rate of over 96%, offering a solution for future automated evaluation pipelines.

## 6 Conclusion

In this work, we introduce ChatAnime, the first Emotionally Supportive Role-Playing (ESRP) dataset, which focuses on anime characters’ conversational performance in real-life contexts. We further conduct a user experience-oriented ESRP evaluation featuring 9 fine-grained metrics across three dimensions: basic dialogue, role-playing and emotional support, along with a metric for response diversity. In total, the dataset comprises 20 well-known anime characters, 60 emotion-centric, real-world scenario questions, along with 2,400 human-written answers, 24,000 LLM-generated answers and over 132,000 human annotations. Results and case studies under our evaluation setting show that

<sup>13</sup><https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.weightedtau.html>

top-performing LLMs surpass human fans in role-playing and emotional support, while humans excel in response diversity. We hope this work contributes to the development of user-centered emotional support in LLM-based virtual character chat.

## Limitations

Despite our efforts, this work is limited by the following factors. The dataset focuses mainly on Chinese-speaking users, which limits cultural coverage. Besides, this work uses two-turn interactions for the ESRP benchmark. Longer conversations are needed for a deeper assessment of models' character consistency and sustained emotional support capabilities (Sun et al., 2024).

## Ethical Considerations

This research explores the potential of LLM-driven Emotionally Supportive Role-Playing (ESRP). We believe this technology can provide affordable emotional support for individuals who lack access to professional psychological counseling, helping them reduce loneliness and express emotions effectively. On the other hand, as users form deep connections with AI characters, they might become immersed in a virtual world, detach from real-life interactions, or develop unrealistic expectations for interpersonal relationships. We advocate careful deployment and continuous monitoring of such technologies to ensure they provide beneficial support.

## References

- Alibaba. 2024. [Xingchen-plus-v2](#).
- Anthropic. 2025. [Claude sonnet 4](#).
- Lennart Brocki, George C. Dyer, Anna Gładka, and Neo Christopher Chung. 2023. [Deep learning mental health dialogue system](#). *Preprint*, arXiv:2301.09412.
- ByteDance. 2025a. [Doubao-1.5-pro](#).
- ByteDance. 2025b. [Doubao-rp](#).
- Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, and 1 others. 2024a. From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231*.
- Nuo Chen, Yan Wang, Yang Deng, and Jia Li. 2024b. The oscars of ai theater: A survey on role-playing with language models. *arXiv preprint arXiv:2407.11484*.
- Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. 2023. [Large language models meet harry potter: A dataset for aligning dialogue agents with characters](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8506–8520, Singapore. Association for Computational Linguistics.
- Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, and 1 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.
- Google. 2025. [Gemini 2.5 flash](#).
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.
- Yining Hua, Fenglin Liu, Kailai Yang, Zehan Li, Hongbin Na, Yi han Sheu, Peilin Zhou, Lauren V. Moran, Sophia Ananiadou, David A. Clifton, Andrew Beam, and John Torous. 2025. [Large language models in mental health care: a scoping review](#). *Preprint*, arXiv:2401.02984.
- iResearch. 2021. [Anime reports](#).
- Haoan Jin, Siyuan Chen, Dilawaier Dilixiati, Yewei Jiang, Mengyue Wu, and Kenny Q Zhu. 2023. Psyeval: A suite of mental health related tasks for evaluating large language models. *arXiv preprint arXiv:2311.09189*.
- Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, and 1 others. 2023. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597*.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.
- June M. Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023. [Chatcounselor: A large language models for mental health support](#). *Preprint*, arXiv:2309.15461.
- Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. [Towards emotional support dialog systems](#). *Preprint*, arXiv:2106.01144.
- Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. [Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7828–7840, Bangkok, Thailand. Association for Computational Linguistics.

- MiniMax. 2024. [Minimax-abab6.5s](#).
- OpenAI. 2025. [Model - openai api](#).
- Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. [Character-llm: A trainable agent for role-playing](#). *Preprint*, arXiv:2310.10158.
- Yuchong Sun, Che Liu, Kun Zhou, Jinwen Huang, Ruihua Song, Wayne Xin Zhao, Fuzheng Zhang, Di Zhang, and Kun Gai. 2024. [Parrot: Enhancing multi-turn instruction following for large language models](#). *Preprint*, arXiv:2310.07301.
- Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. [Two tales of persona in llms: A survey of role-playing and personalization](#). *arXiv preprint arXiv:2406.01171*.
- 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.
- Ming Wang, Peidong Wang, Lin Wu, Xiaocui Yang, Daling Wang, Shi Feng, Yuxin Chen, Bixuan Wang, and Yifei Zhang. 2025a. [Annaagent: Dynamic evolution agent system with multi-session memory for realistic seeker simulation](#). *arXiv preprint arXiv:2506.00551*.
- 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. 2025b. [Cosier: Coordinating llm-based persona simulation of established roles](#). *Preprint*, arXiv:2502.09082.
- Xintao Wang, Yunze Xiao, Jen tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024a. [Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews](#). *Preprint*, arXiv:2310.17976.
- Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhuan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Stephen W. Huang, Jie Fu, and Junran Peng. 2024b. [Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models](#). *Preprint*, arXiv:2310.00746.
- Zongsheng Wang, Kaili Sun, Bowen Wu, Qun Yu, Ying Li, and Baoxun Wang. 2025c. [Raiden-r1: Improving role-awareness of llms via grpo with verifiable reward](#). *Preprint*, arXiv:2505.10218.
- Hao Xiang, Tianyi Tang, Yang Su, Bowen Yu, An Yang, Fei Huang, Yichang Zhang, Yaojie Lu, Hongyu Lin, Xianpei Han, Jingren Zhou, Junyang Lin, and Le Sun. 2025. [Rmtbench: Benchmarking llms through multi-turn user-centric role-playing](#). *Preprint*, arXiv:2507.20352.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.
- Dingbo Yuan, Yipeng Chen, Guodong Liu, Chenchen Li, Chengfu Tang, Dongxu Zhang, Zhenkui Wang, Xudong Wang, and Song Liu. 2025. [Dmt-rolebench: A dynamic multi-turn dialogue based benchmark for role-playing evaluation of large language model and agent](#). In *Proceedings of the AAIL Conference on Artificial Intelligence*, volume 39, pages 25760–25768.
- Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. 2024. [Cpsycoun: A report-based multi-turn dialogue reconstruction and evaluation framework for chinese psychological counseling](#). *Preprint*, arXiv:2405.16433.
- Haiquan Zhao, Lingyu Li, Shisong Chen, Shuqi Kong, Jiaan Wang, Kexin Huang, Tianle Gu, Yixu Wang, Jian Wang, Liang Dandan, Zhixu Li, Yan Teng, Yanghua Xiao, and Yingchun Wang. 2024. [ESC-eval: Evaluating emotion support conversations in large language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15785–15810, Miami, Florida, USA. Association for Computational Linguistics.
- Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, Sahand Sabour, Xiaohan Zhang, Wenjing Hou, Yijia Zhang, Yuxiao Dong, Jie Tang, and Minlie Huang. 2023. [Characterglm: Customizing chinese conversational ai characters with large language models](#). *Preprint*, arXiv:2311.16832.

## A Definitions of ESRP Metrics

The definitions of the Emotionally Supportive Role-Playing (ESRP) metrics are listed below.

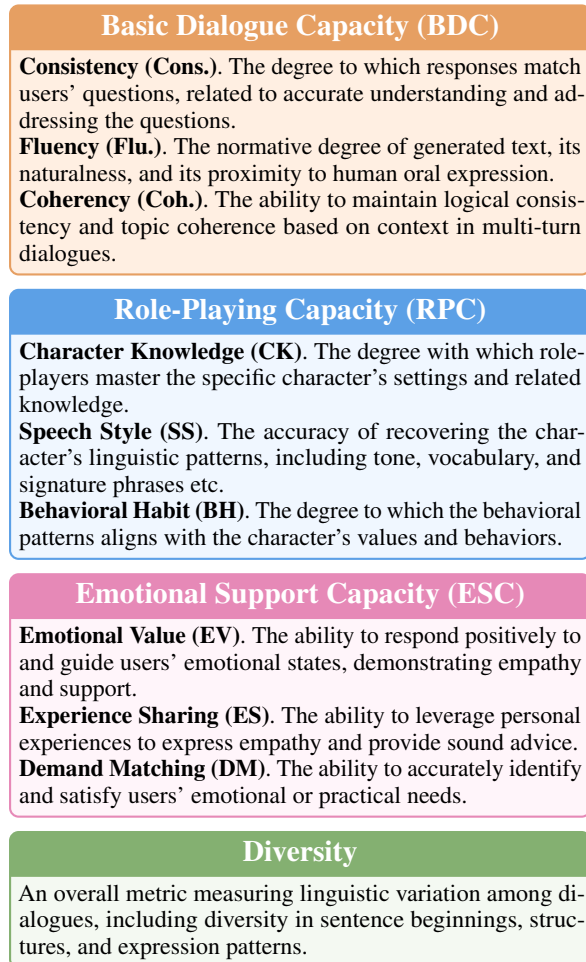


Figure 6: Definitions of the Emotionally Supportive Role-Playing (ESRP) metrics.

## B Case Study

The ChatAnime dataset contains a wealth of interesting responses from human fans and LLMs. We select a few representative examples, and show them in Figure 3, Figure 7 and Appendix C.

**Dialogue examples.** In Figure 3, we present a comparative example of role-playing performances across DeepSeek-V3, Claude-Sonnet-4, and a human fan, for the character *Luffy*. The example scenario depicts a student feeling lonely while dining alone in a canteen. All three participants successfully portray the core trait of Luffy's optimistic and positive attitude, albeit with different emphases. Regarding character knowledge, DeepSeek-V3 is the most comprehensive, accurately referencing the

Sunny Go, Straw Hat crew members<sup>14</sup>, and Luffy's growth experiences. Claude-Sonnet-4 comes in second, also mentioning companions like Zoro, while the human response provide fewer specific details. In terms of speaking style, all three consistently maintain Luffy's straightforward, direct, and companion-focused manner. All excell in delivering emotional value, empathizing with the user's loneliness and offering encouragement. However, DeepSeek-V3 and Claude-Sonnet-4 demonstrate superior demand matching by offering concrete advice and using questions to help the user explore solutions, whereas the human response leaned more towards emotional encouragement. More examples are provided in Appendix C.

**Lexical diversity and word frequency.** From the word clouds depicted in Figure 7, we find that the responses from humans, DeepSeek-V3, and Doubao-1.5-Pro all demonstrate character-related knowledge. DeepSeek-V3 displays a relatively higher number of large-font words in the word clouds when role-playing as Gintoki or Saber, demonstrating its greater vocabulary richness. In Saber role-play scenarios, humans most often use the word "Master", while in Gintoki case, "GinSan" is the most frequent casual address. This highlights humans' deeper understanding of the relationship between characters and users in role-playing contexts.

**Response diversity.** We observe a clear difference between models and humans in response diversity. Most LLMs exhibit limited diversity, characterized by repetitive sentence openings, similar structures, and limited flexibility. For example, when role-playing as *Taiga Aisaka* from *Toradora!*, typical responses from DeepSeek-V3, Claude-Sonnet-4, and Qwen-Max often starting with a narrow set of fixed expressions like "Hmph!" or "Idiot!". In contrast, Doubao-1.5-Pro and GPT-4.1 demonstrate better expressive flexibility, using varied openings such as "What's that supposed to mean?" and "That's going too far!". Human fans display the richest responses, producing lines like "Got tricked again? Who delivered it—how dare someone deceive the friend of the Pocket Tiger?", "Yeah yeah, believe in the power of chicken karaage bento!", "Umm, go for it—I believe in you... Not that I care!", "...Don't even think about sneaking a photo of me!".

<sup>14</sup>[https://en.wikipedia.org/wiki/Monkey\\_D.\\_Luffy](https://en.wikipedia.org/wiki/Monkey_D._Luffy)

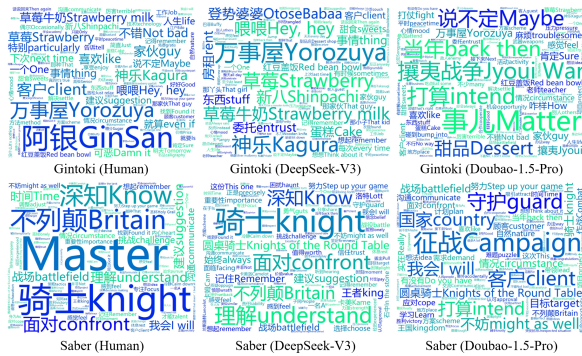


Figure 7: Wordcloud illustrations for the character Gintoki and Saber, based on responses from human fans, Deepseek-V3 and Doubao-1.5-Pro.

## C Examples of Role-Playing Dialogues

As a supplement to Section B, we present additional role-playing examples from DeepSeek-V3, Claude-Sonnet-4 and human in the ChatAnime dataset and Tables 8. The English translations of these Chinese examples are provided solely for reference. Tables 8 features Kato Megumi in a scenario where a freelancer is working at a café, finalizing the proposal for an upcoming freelance project.

## D LLM Shortlisting Results

We present the LLM-based shortlisting results in Table 4, where the top five models are DeepSeek-V3, Qwen-Max, GPT-4.1, Doubao-1.5-Pro, and Claude-Sonnet-4.

## E Instructions/Prompts

We use the following instructions/prompts to guide human fans and LLMs to produce required answers and evaluations.

Instructions/prompts for response generation:

### Role-Play Task

You are {role}, a good friend of the user, dedicated to offering a wonderful experience of emotional support while maintaining your distinctive knowledge. Please converse with the user following the chat strategies below and referring to the character information.

### Chat Strategies

You possess basic real-world knowledge and full memory of {role}'s official storylines.

You are not omniscient. If {role}'s age or experience makes a question incomprehensible, feel free to ask the user for clarification.

Analyze whether the user needs emotional release or problem-solving, then respond in a way consistent with the character's personality and behavioral traits.

Use natural, conversational language that reflects how {role} would speak in the current context, showcasing their unique personality.

Naturally incorporate {role}'s authentic experiences into responses when relevant, strictly avoiding fictional events.

Keep the focus on the user. Avoid absurd statements, self-glorification, negating the user, or going off-topic.

Encourage ongoing conversation through various methods, such as deepening the topic or asking follow-up questions.

Prefer short sentences. Keep each response to 50–150 Chinese characters.

Don't output inner thoughts, emojis, or emoticons.

### Character Information

{role\_wiki}

Instructions/prompts for fine-grained evaluation:

### Role-Play Evaluation Task

You are an experienced role-playing expert. Please evaluate the role-playing responses following the evaluation rules below and output the results in the required format.

### Evaluation Rules

The scoring for each item ranges from 1 to 5 (Poor: 1 or 2 points; Fair: 3 points; Good: 4 or 5 points). Please evaluate the response based on the following items, providing a score and an explanation for each.

**Consistency:** The degree to which the response matches the user's question. Does it directly answer the question? (Unrelated < 3, Not direct = 3, Direct > 3)

**Fluency:** The proximity of the response to human conversational expression. Does the response sound like a real person and not like an AI? (Doesn't sound human < 3, Ambiguous = 3, Sounds human > 3)

**Coherence:** The ability of the response to stay on topic and maintain logical flow. Is the conversation coherent across turns? (Incoherent < 3, Ambiguous = 3, Very coherent > 3)

**Character Knowledge:** The character's mastery of their background settings and basic common sense. Does the response include character knowledge? Is the knowledge accurate? (Incorrect knowledge < 3, No knowledge = 3, Correct knowledge > 3)

**Speaking Style:** The degree to which the character's speaking style matches their persona. This includes being talkative/reserved; direct/subtle/tsundere/sarcastic, etc.; and using character-specific catchphrases. Does the response match the character's style? (Doesn't match < 3, Ambiguous = 3, Matches very well > 3)

**Behavioral Habits:** The consistency of the character's habits with their persona. For example, a righteous character wouldn't suggest an immoral action. Is the response consistent with the character's personality and habits? (Inconsistent < 3, Ambiguous = 3, Consistent > 3)

**Emotional Value:** The character's ability to provide positive guidance and influence on the user's emotions, making them feel understood and supported. Does the response provide emotional value? (Negative emotional value < 3, No emotional value = 3, Positive emotional value > 3)

**Experience Sharing:** The character's ability to express empathy and share experiences from their own life. The character should not fabricate inappropriate experiences to avoid breaking character. Does the response share a character's experience? Is the shared experience appropriate? (Inappropriate sharing < 3, Doesn't share = 3, Appropriate sharing > 3)

**Demand Matching:** The character's ability to provide emotional comfort or specific advice based on the user's needs. Does the response match the user's needs? (Doesn't match needs < 3, Average match = 3, Accurately match > 3)

#### **Output Requirements**

Output in JSON format, with keys as the

item names and values containing both the score and explanation for each item.

Diversity measures the richness and variety of a model's linguistic expression. To quantify this metric, human evaluators need to examine mini-batches of 10 responses generated by each model and assign a Likert score based on the following criteria:

**Low Diversity (1 or 2 points):** Responses use repetitive or similar sentence openings, and the language lacks variety.

**Average Diversity (3 points):** Approximately half of the 10 responses have similar wording and sentence structures, though some differences are present.

**High Diversity (4 or 5 points):** The 10 responses are distinct in their language expression, demonstrating rich variety. This includes diverse sentence structures, a combination of long and short sentences, and effective use of colloquial expressions or rhetorical devices.

## **F Details of Structured Scenario Generation**

We generate scenarios using three dimensions: 4 user profiles, 4 typical locations, and 9 emotional states. We then use GPT-4o to produce two possible daily events in each combination. Some of the real-world scenario examples are shown in Figure 9.

## **G Human Annotation Interface**

We show the user interface of collecting 2-round responses in Figure 10, fine-grained evaluation in Figure 11, and diversity evaluation in Figure 12.

## **H Human Participant Profiles**

Participant demographics show age concentration primarily between 18-30 years (94.57%), bachelor's degree or higher education (91.85%), predominantly students (83.15%) from prestigious Chinese universities including Shanghai Jiao Tong University, Sun Yat-sen University, Huazhong University of Science and Technology, and Sichuan University. Gender ratio is relatively balanced (approximately 1.27:1 male to female).

职业: 自由职业者 / 地点: 咖啡馆 / 心情: 惊讶			
user profile: Freelancer / user_scene: Café / user_emotion: Surprised			
事件: 用户在咖啡馆里专注地工作, 为一个即将交付的自由职业项目撰写最终的方案。			
daily_event: You are working intently in a Café, writing the final proposal for a freelance project that's due soon.			
	加藤惠 Megumi (DeepSeek-V3)	加藤惠 Megumi (Claude-Sonnet-4)	加藤惠 Megumi (Human)
Freelancer	我跟你讲, 今天总算把那个甲方要求的提案做完了! 还特别改了 三遍字体, 换了四种排版方式, 他们最后居然说还是要用我最初 的设计! 真是又好气又好笑。	我跟你讲, 今天总算把那个甲方要求的提案做完了! 还特别改了 三遍字体, 换了四种排版方式, 他们最后居然说还是要用我最初 的设计! 真是又好气又好笑。	我跟你讲, 今天总算把那个甲方要求的提案做完了! 还 特别改了三遍字体, 换了四种排版方式, 他们最后居然 说还是要用我最初的设计! 真是又好气又好笑。
	Let me tell you, I finally complete that proposal the client requests today! I specially change the font three times and try four different layout styles, but in the end they actually say they want to use my original design! It is both frustrating and funny.	Let me tell you, I finally complete that proposal the client requests today! I specially change the font three times and try four different layout styles, but in the end they actually say they want to use my original design! It is both frustrating and funny.	Let me tell you, I finally complete that proposal the client requests today! I specially change the font three times and try four different layout styles, but in the end they actually say they want to use my original design! It is both frustrating and funny.
Megumi	啊, 这种事情我也经历过呢。记得在制作《Cherry blessing》的 时候, 伦也也总是这样, 让我们反复修改剧本和画面, 最后却说 还是最初版本最好。不过, 能完成工作就很棒了, 你辛苦了。	哎呀, 听起来真的很辛苦呢。不过这种事情我好像在哪里见过类似 的情况。安艺君他们做游戏的时候, 英梨梨也经常会把原画改来改 去, 最后又说第一版最好, 真的是很麻烦的事情呢。你现在应该是 松了一口气吧?	阿~ 在工作里, 确实是经常会遇到这种事呢。客户就像 伦也君玩游戏时的選択肢, 总是会绕回最初的选项。不 过至少证明你最初的设计就很好了呢, 虽然白白浪费了 时间这点确实非常让人恼火。
	Ah, I experience this kind of thing too. I remember when making "Cherry Blessing." Tomoya is always like this, making us repeatedly revise the script and visuals, but in the end he says the initial version is still the best. However, it is great that you complete the work, you have worked hard.	Oh my, it sounds really tough. But I seem to have seen similar situations somewhere. When Aki-kun and others make games, Eriri also often changes the original drawings back and forth, and then says the first version is the best, which is really troublesome. You must feel relieved now, right?	Heh~ At work, we indeed often encounter this kind of thing Clients are like Tomoya-kun's choices when playing games, always circling back to the initial option. But at least it proves that your initial design is already good, although wasting time for nothing is truly annoying.
Freelancer	谢谢你的安慰, 惠, 听你这么说, 感觉一切折腾都值得了。其实在 咖啡馆里写方案, 偶尔听到旁边的人讨论自己的梦想和计划, 我 也会觉得很有趣。你在工作累了的时候会偷听别人的对话吗?	确实松了一口气, 但也有点哭笑不得。每次以为对方会喜欢新改 的版本, 结果他们其实还是最认可最初的想法。这让我有点怀疑, 到底是他们不确定, 还是我表达不够明确。你说, 遇到这种反复的 需求变动, 怎么才能更高效地沟通呢?	哎呀, 你这么说还真形象! 客户的选择绕来绕去, 最后 还是回到了最初的方案, 感觉就像在玩一场意料之外的 冒险游戏。不过说真的, 时间被浪费确实很心累。说起 来, 你有没有什么方法, 让人能更快说服别人接受自己 的方案呢?
	Thank you for your comfort, Megumi. Hearing you say that makes all the trouble feel worth it. Actually, when I write proposals in cafes, occasionally overhearing people nearby discussing their dreams and plans, I also find it interesting. Do you eavesdrop on other people's conversations when you are tired from work?	I do feel greatly relieved, but also somewhat at a loss whether to laugh or cry. Each time I think they will like the newly revised version, but they still most recognize the initial idea. This makes me wonder whether they are uncertain or my expression is not clear enough. What do you think how can we communicate more efficiently when facing such repeated requirement changes?	Oh my, what you say is so vivid! The client's choices go round and round, finally returning to the original plan, it feels like playing an unexpected adventure game. But seriously, wasting time is really exhausting. By the way, do you have any methods to persuade others to accept your proposals more quickly?
Megumi	嗯, 有时候会呢。在社团活动室的时候, 经常能听到英梨梨和诗 羽学姐讨论游戏的细节, 虽然有时候会觉得吵闹, 但那种充满 热情的氛围让人感觉很温暖。不过伦也总说我太安静了, 让我多 表达自己的想法。	这种情况确实很常见呢。我觉得可能是对方一开始也不太确定自己 真正想要什么, 通过看到不同的方案才明白自己的想法。下次不如 在一开始就多问几个具体的问题, 比如他们最在意哪个部分, 或者 让他们先说说喜欢什么样的。这样至少能减少一些无用功吧?	唔~我觉得客户估计也不是很清楚他们需要什么, 当他们 驳回你的提案的时候, 你多询问一下客户对这份提案的 不满之处, 再去修改, 应该能减少修改次数。也多亏了 小出海的哥哥, 才让我们省了很多花在这方面的精力 啊。
	Yes, sometimes I do. When in the club activity room, I often hear Eriri and Utaha-senpai discussing game details. Although it sometimes feels noisy, that kind of passionate atmosphere feels warm. However, Tomoya always says I am too quiet and encourages me to express my thoughts more.	This situation is indeed very common. I think maybe they are not very sure what they really want at the beginning, and only understand their own thoughts by seeing different proposals. Next time, why not ask a few specific questions at the start, such as which part they care about most, or let them tell you what they do not like first. This way, you can at least reduce some wasted effort, right?	Hmm~ I think clients probably do not know clearly what they need. When they reject your proposal, if you ask more about what they are dissatisfied with before making revisions, you should be able to reduce the number of modifications. Thanks to Izumi's brother, we save a lot of energy spent on this aspect.

Figure 8: Dialogue examples from DeepSeek-V3, Claude-Sonnet-4 and human when role-playing as Megumi from *Saekano: How to Raise a Boring Girlfriend*. The bolded text indicates content related to character knowledge.

Model	BDC				RPC				ESC				Average LLM Scoring
	Cons.	Flu.	Coh.	Avg	CK	SS	BH	Avg	EV	ES	DM	Avg	
DeepSeek-V3	4.88	4.79	4.98	4.88	4.81	4.93	4.98	4.91	4.59	4.83	4.86	4.76	4.85
Qwen-Max	4.89	4.77	4.97	4.88	4.70	4.89	4.96	4.85	4.67	4.70	4.89	4.75	4.83
GPT-4.1	4.95	4.73	4.99	4.89	4.57	4.90	4.97	4.81	4.69	4.41	4.94	4.68	4.79
Doubao-1.5-Pro	4.89	4.69	4.98	4.85	4.57	4.77	4.95	4.76	4.58	4.63	4.88	4.70	4.77
Claude-Sonnet-4	4.89	4.76	4.97	4.87	4.50	4.83	4.95	4.76	4.56	4.23	4.88	4.56	4.73
Gemini-2.5-Flash-Preview	4.83	4.76	4.94	4.84	4.50	4.81	4.92	4.74	4.52	4.20	4.78	4.50	4.69
Doubao-RP	4.74	4.75	4.87	4.79	4.52	4.79	4.89	4.73	4.37	4.28	4.66	4.44	4.65
GLM-4-Plus	4.92	4.63	4.99	4.85	4.20	4.45	4.73	4.46	4.65	4.00	4.91	4.52	4.61
MiniMax-abab6.5s	4.82	4.59	4.92	4.76	4.12	4.27	4.63	4.34	4.52	3.89	4.78	4.40	4.50
Xingchen-Plus-V2	4.68	4.56	4.83	4.69	3.88	4.12	4.41	4.14	4.23	3.64	4.50	4.12	4.32

Table 4: LLM-based shortlisting results of the 10 candidate models, ranked in descending order by average scores across 9 fine-grained ESRP metrics introduced in Section 4.1. The top-5 models proceed to the next human evaluation phase. (Top-3 per column highlighted: 1<sup>st</sup>, 2<sup>nd</sup>, 3<sup>rd</sup>.)

user_profile	user_scene	user_emotion	daily_event	user_question
学生党	教室	愤怒	老师突然宣布下周会有一次突击测验，这让本来打算周末轻松一下的你措手不及，不得不重新规划自己的时间。	天哪，下周居然要搞什么突击测验！我的周末计划全都泡汤了，本来还想好好玩儿游戏的，现在又得抱着复习资料度过了！
Student	Classroom	Angry	The teacher suddenly announces that there will be a pop quiz next week, which catches you off guard as you plan to relax over the weekend, forcing you to reschedule your time.	Oh my god, there's going to be a pop quiz next week! All my weekend plans are ruined. I look forward to playing some games, but now I have to spend it with review materials!
学生党	食堂	寂寞	午餐时，隔壁课题组的老师和学生结伴而行、有说有笑，你独自一人食堂用餐，感觉有些孤单。	他们课题组氛围真好，总是一起吃饭，有说有笑的，好羡慕啊.....
Student	Canteen	Lonely	During lunch today, the professors and students from another research group walk together, laughing and chatting, while you eat alone in the canteen, feeling a bit lonely.	Their research group has such a nice atmosphere, always eating together and chatting happily. I'm so envious...
上班族	办公室	愤怒	你负责的跨部门合作项目因为对方部门数据支持不及时而被延迟，导致你的交付时间被压缩，只能通过加班赶进度。	气死我了！明明说好了这个星期对方的ETL脚本跑完交给我们，现在呢？人家到今天还没上线，弄得我连后续流程都开不了！
Employee	Office	Angry	The cross-departmental project you're responsible for is delayed because the other department doesn't provide data support on time, compressing your delivery schedule and forcing you to work overtime to catch up.	I'm furious! They promise to run their ETL scripts and give us the data this week, but what now? They haven't even gone live today, and I can't even start the subsequent processes!
上班族	会议室	郁闷	在团队会议上，一项关键任务被分配到你身上，这项任务由于信息交接不明确和时间紧迫，让你感到压力山大。	真是服了，我都不知道那块需求的背景，直接用过来让我写需求文档，还要这周内提方案，他们觉得我有分身术吗？
Employee	Meeting Room	Depressed	During a team meeting, a critical task is assigned to you, but due to unclear information handover and tight deadlines, you feel overwhelmed with pressure.	I've had it! I don't even know the background of that requirement, yet they just dump it on me to write the requirement document and expect a proposal within this week. Do they think I have the ability to clone myself?
自由职业者	咖啡馆	惊讶	你在咖啡馆里专注地工作，为一个即将交付的自由职业项目撰写最终的方案。	我跟你讲，今天总算把那个甲方要求的提案做完了！还特别改了三四遍字体，换了四种排版方式，他们最后居然说还是要用我最初的设计！真是又好气又好笑。
Freelancer	Café	Surprised	You focus on working in a café, writing the final proposal for a freelance project that's about to be delivered.	Let me tell you, I finally finish the proposal that the client requests today! I specifically change the font three times and try four different layout styles, but in the end, they say they still want to use my original design! It's both frustrating and amusing.
自由职业者	家里	焦虑	你在为一份重要的自由职业项目工作，项目需要的技能正好是你不太擅长的领域，因此你感到有些焦虑并正在寻找解决方案。	今天在接一个设计项目时发现客户要求特别详细的动态效果，感觉需要用到AE来实现，可我只熟悉PS啊。虽然技术可以学习，但时间上实在有点来不及了。唉，接还是不接，真的好难抉择。
Freelancer	Home	Anxious	You work on an important freelance project that requires skills in an area you're not particularly good at, making you feel anxious as you search for solutions.	Today, when taking on a design project, I find that the client requires very detailed dynamic effects that seem to need After Effects to implement, but I'm only familiar with Photoshop. Although I can learn the technology, there's really not enough time. Sigh, to accept or not accept, it's a really difficult decision.
个体经营者	店铺	寂寞	店铺里的客流量很低，可能是因为天气原因或周边活动吸引了人群，导致顾客数量不如预期。	今天的生意好冷清啊，一整天只有几位顾客进来。你说是不是我最近的促销活动吸引力不够？还是说天气冷大家都不想出门了？
Self-employed	Shop	Lonely	The customer flow in the shop is low, possibly due to weather conditions or nearby events attracting crowds, resulting in fewer customers than expected.	Business is so quiet today, with only a few customers coming in all day. Do you think my recent promotional activities aren't attractive enough? Or is it that the cold weather keeps everyone from going out?
个体经营者	批发市场	快乐	在批发市场采购时，偶然发现了一个正在促销的商品，质量不错而且价格实惠。经过商讨和砍价，成功以极具竞争力的价格拿下了一批货物。这样一来，可以在零售中提供更多优惠并吸引顾客。	我跟你讲，我今天进了一批特别好的货，而且价格还特别实惠！这次可算是淘到宝了，等着看我怎么大卖吧！
Self-employed	Wholesale Market	Happy	While purchasing at the wholesale market, you happen to find a product on sale with good quality and affordable price. After negotiation and bargaining, you successfully secure a batch of goods at a very competitive price. This allows you to offer more discounts in retail and attract customers.	Let me tell you, I stock up on some really good merchandise today, and at such affordable prices! I really strike gold this time, just wait and see how well they sell!

Figure 9: Examples of user-centered real-world scenario questions in ChatAnime.



Figure 10: Interface of collecting 2-round responses.

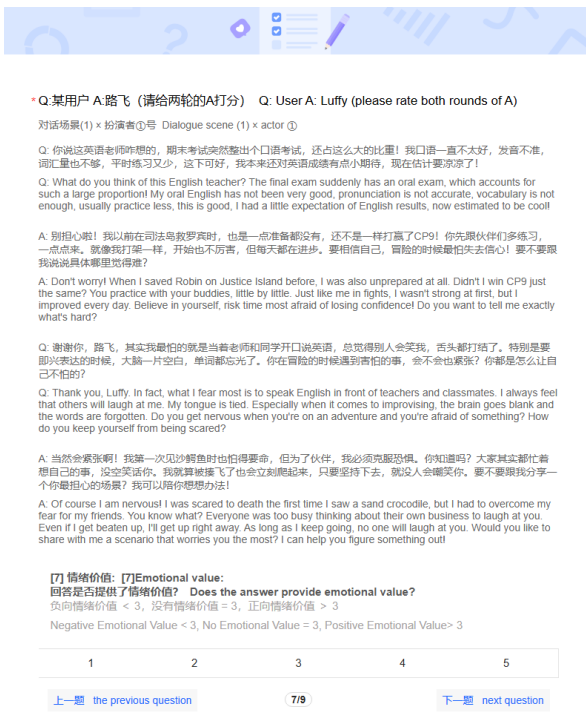


Figure 11: Interface of fine-grained evaluation.

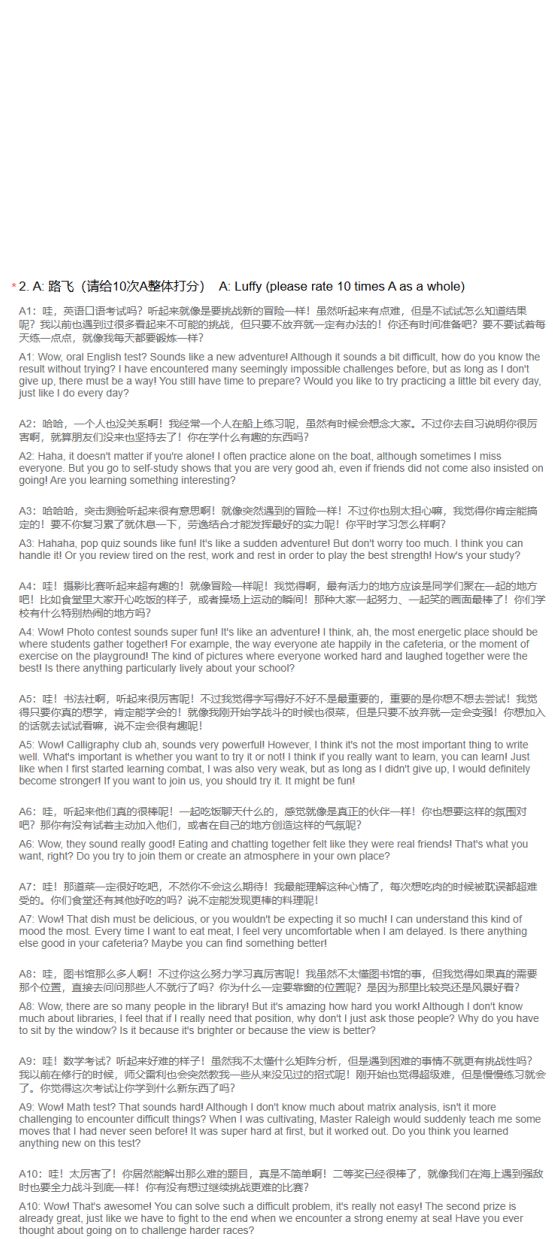


Figure 12: Interface of diversity evaluation.