BEYOND DIALOGUE: A Profile-Dialogue Alignment Framework Towards General Role-Playing Language Model

Anonymous ACL submission

Abstract

The rapid advancement of large language models (LLMs) has revolutionized role-playing, enabling the development of general role-playing models. However, current role-playing training has two significant issues: (I) Using a predefined role profile to prompt dialogue training for specific scenarios usually leads to biases and even conflicts between the dialogue and the profile, resulting in training biases. (II) Models learn to imitate the role based solely on the profile, neglecting profile-dialogue alignment at the sentence level. To overcome the aforementioned hurdles, we propose a novel framework **BEYOND DIALOGUE**, which introduces "beyond dialogue" tasks to align dialogue with profile traits for each scenario, eliminating biases during training. Furthermore, the framework achieves a sentence-level fine-grained alignment between profile and dialogue through an innovative prompting mechanism that generates reasoning data for training. Moreover, the aforementioned methods are fully automated and low-cost. Experimental results demonstrate our model excels in adhering to role profiles, outperforming most proprietary general and specialized role-playing baselines. The code and data are provided in the supplementary materials, and will be open-sourced upon acceptance.

1 Introduction

The rapid advancement of large language models (LLMs) has demonstrated their significant potential in user interactions (Achiam et al., 2023; Yang et al., 2024; Dubey et al., 2024). A particularly promising area is the development of role-playing LLM agents, which are capable of simulating both real and fictional roles to deliver immersive and interactive experiences (Chen et al., 2024b,a). However, the proprietary nature of these platforms restricts users' ability to customize specific role models. Thus, the development of an open-source, general role-playing model is imperative.

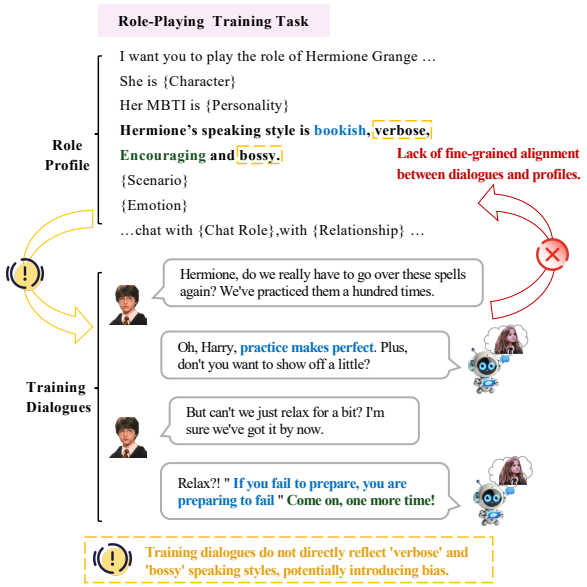


Figure 1: An example of the **bias issue between predefined role profile and dialogues in role-playing training.**

At the current stage of role-playing dialogue training tasks, role profiles are typically manually compiled (Shao et al., 2023) or generated through LLM summaries (Zhou et al., 2024), including information such as personality and speaking style as depicted throughout a whole novel or script (Zhou et al., 2024; Li et al., 2023). However, a bias exists between the training dialogues extracted from individual scenarios and the predefined role profiles. As an illustration in Fig. 1, Hermione's speaking style is predefined into four categories, which are manually extracted from the entire content of the novel. However, the training dialogues in Fig. 1 predominantly reflect only two styles ("bookish–blue" and "encouraging–green"). The bias arises because a single scenario often fails to fully capture the predefined profile, resulting in misalignment even contradiction between the prompted profile and the training dialogues. Such biases are common: from the HPD dataset published by (Chen et al., 2023), we sampled 800 scenarios and found that 83.2% of

the dialogues exhibited bias in relation to the predefined role profiles, both in character and speaking style. As shown in our § 4.2 experiments, biases between predefined role profiles and the training dialogue corpus can mislead LLMs, impairing its ability to accurately reflect the role profile in dialogues. Meanwhile, the model has only learned a vague mapping from profile to dialogue, without fine-grained alignment of specific traits to their corresponding dialogues' sentence (as indicated by the red arrow in Fig. 1). Consequently, the model is unable to learn how "personality traits are revealed in dialogues".

These observations highlight two critical issues in the current role-playing training process:

- **Bias Between Profile and Dialogues**: A dialogue corpus that deviates from preset profiles may introduce bias during training, impairing the model's ability to follow the profiles.
- **Inadequate Fine-Grained Alignment**: A single dialogue training task fails to fine-grained align role dialogues with their profiles, as it lacks the nuanced understanding of how specific traits manifest, limiting the model's ability to comprehend and represent the roles' complex traits.

We propose a new training role-playing framework—**BEYOND DIALOGUE** to overcome the aforementioned issues and advance toward general role-playing. We incorporate powerful LLM through a prompting mechanism approach to align role profiles with scenario dialogues. This not only ensures the reliability of the LLM's reasoning but also explicitly creates a fine-grained alignment dataset at the sentence level. Taking inspiration from actors learning to play different roles—understanding the performance of various role traits in scenarios to enhance their portrayal (Zarrilli, 2005)—we also employ fine-grained alignment tasks to train the role-playing model. Compared to previous methods that relied on profiles to loosely guide dialogue generation (Lu et al., 2024; Tang et al., 2024), these alignment tasks directly link profile attributes to dialogue sentences, allowing the model to better capture role traits at the sentence level and improve its effectiveness in role portrayal.

Traditional evaluation methods rely on subjective assessments (e.g., human ratings or LLM-based scoring), leading to inconsistent outcomes (Wang et al., 2024a; Tu et al., 2024). Our approach converts evaluation tasks into objective ones (e.g., multiple-choice and judgment questions), clearly defining the model's ability to follow role profiles as the evaluation criterion. Combined with "LLMs as Judger" (Kim et al., 2023), this approach objectively quantifies role-following capabilities with measurable criteria as demonstrated in § B.1.

We applied **BEYOND DIALOGUE** to bilingual chat LLM baselines—Qwen2 (Yang et al., 2024) and Mistral-Nemo (MistralAI, 2024)—to evaluate its effectiveness. Extensive experiments demonstrate that **BEYOND DIALOGUE** significantly improves LLMs' ability to follow and portray role profiles. Under our framework, the trained models outperformed advanced general baselines like GPT-4o (OpenAI, 2024) and specialized role-playing baselines like Baichuan-NPC-Turbo (Baichuan, 2024) in key aspects of role-playing performance.

Our main contributions are as follows:

1. We identified a bias in role-playing training where the predefined role profile misaligns with the dialogues in a specific scenario. Furthermore, we demonstrated that this bias significantly hinders the model's ability to effectively follow the predefined profile.
2. We propose **BEYOND DIALOGUE**, a novel general role-playing training framework that specifically addresses the identified bias by using LLMs with a prompting mechanism to align role profiles with scenario-specific dialogues. This approach generates fine-grained alignment tasks at the sentence level, which are integrated into the training, further enhancing the effectiveness of role portrayal.
3. We introduce a novel evaluation pipeline that converts all assessment tasks into objective tasks, focusing on the model's ability to follow user-defined role profiles. By combining automatic dialogues with the "LLMs as Judger" method, our approach ensures both efficiency and reproducibility in role-playing LLMs evaluation.

2 Related Work

General Role-playing Agents. Recent research on general role-playing agents has largely overlooked the issue of bias between predefined profiles and scenario-specific dialogues. Instead, the focus has

Role-Playing Dataset/Framework	Dataset Automatically Constructed?	Scenario	Style	Character	Personality	Emotion	Relationship	Evaluation Scope
LIGHT (Urbanek et al., 2019)	✗	✓	✗	✓	✗	✗	✗	Sentence
PDP (Han et al., 2022)	✗	✗	✓	✓	✗	✗	✗	Sentence
ChatHaruhi (Li et al., 2023)	✓	✗	✓	✓	✗	✗	✗	Sentence
HPD (Chen et al., 2023)	✗	✓	✓	✓	✗	✗	✓	Sentence
CharacterLLM (Shao et al., 2023)	✓	✓	✗	✓	✗	✗	✗	Multi-Turn Artificial Interview
CharacterGLM (Zhou et al., 2024)	✗	✗	✗	✓	✗	✗	✗	Multi-Turn Artificial Dialogue
RoleLLM (Wang et al., 2024a)	✓	✗	✗	✓	✗	✗	✗	Sentence
CharacterEval (Tu et al., 2024)	✓	✗	✗	✓	✓	✗	✗	Sentence
RoleCraft (Tao et al., 2024)	✗	✗	✗	✓	✗	✓	✗	Sentence
InCharacter (Wang et al., 2024b)	✗	✗	✗	✗	✓	✗	✗	Multi-Turn Artificial Interview
DITTO (Lu et al., 2024)	✓	✗	✗	✗	✗	✗	✗	Multi-Turn Automated Dialogue
BEYOND DIALOGUE	✓	✓	✓	✓	✓	✓	✓	Multi-Turn Automated Scenario Dialogue

Table 1: Comparison of **BEYOND DIALOGUE** with other datasets or framework. "✓" indicates that the dataset/framework meets the given standard or includes the specified dimension in the role profile, while "✗" indicates that it does not. "Sentence" refers to evaluation based on the final sentence of multi-turn dialogues, while other methods assess performance through manually or automatically generated dialogues/interviews.

been on constructing role-playing datasets, primarily through manual curation (Zhou et al., 2024), extraction from novels (Li et al., 2023; Tao et al., 2024), or generation using GPT (Shao et al., 2023; Wang et al., 2024a), to acquire extensive annotated role-playing data. We compared various current Role-Playing Datasets/Frameworks and summarized five key dimensions for constructing role profiles and evaluations in Tab. 1: **Style, Character, Personality, Emotion, and Relationship**. Additionally, scenarios provide the background and context for dialogues, making the role's conversations more contextual and realistic (Chen et al., 2023; Shao et al., 2023).

Role-playing Evaluation. The initial step in evaluating role-playing models involves generating dialogues for assessment, employing three primary methods: **Providing Historical Dialogues** (Tu et al., 2024; Wang et al., 2024a; Chen et al., 2023; Tao et al., 2024), **Manual Dialogue** (Zhou et al., 2024; Wang et al., 2024b; Shao et al., 2023), and **Automated Dialogue with LLMs** (Lu et al., 2024). The first method is limited to evaluating the model within the scope of individual sentences; however, single sentences are likely to fail to fully capture the entire role profile and may be biased by exposure to novels or scripts used during pre-training. In contrast, the second and third methods can generate complete scenario dialogues, which are more conducive to comprehensive evaluations.

There are three common methods for evaluating dialogues generated by models: **Metric-based Evaluation** (Li et al., 2023; Wang et al., 2024a), **Human Evaluation** (Zhou et al., 2024), and **"LLMs as Judges"** (Shao et al., 2023; Wang et al., 2024a; Lu et al., 2024). The first method primarily assesses the model's retention of the standard response, which is derived from the original text of the novel or manually annotated compilation. The second method, while more accurate, is costly and difficult to replicate, limiting its broader application. The third method, however, is attracting increasing attention due to its high efficiency, low cost, and scalability.

Hybrid Task Training LLMs. In recent years, hybrid task training has emerged as a pivotal method for fine-tuning LLMs. For instance, the LLaMA Writing Assistant is trained on a variety of textual tasks (Wang et al., 2023). Incorporating tasks like summarization, text generation, and reasoning within reading comprehension has greatly improved the model's language understanding and QA abilities (Cheng et al., 2023). Furthermore, training with a combination of general and role-specific instructions has improved the model's role-playing abilities (Tao et al., 2024). These studies demonstrate that incorporating tasks directly or indirectly related to downstream tasks in hybrid training can significantly enhance model performance.

3 BEYOND DIALOGUE Framework

We propose the **BEYOND DIALOGUE** Framework for role-playing, which includes three key parts: i) in the ***Alignment Dataset Construction*** stage, role profiles are aligned and adjusted for each scenario to create "pure" role-playing dialogue data. ii) in the ***Supervised Finetuning*** stage, the "pure" role-playing dialogue data is combined with alignment reasoning and chit-chat data for model training. iii) in the ***Automated Dialogue Evaluation***, the pipeline generates random scenarios and multi-turn dialogues to interact with the model, using objective questions to quantify its role-playing capabilities.

Details of each component's implementation will be provided in the following chapters.

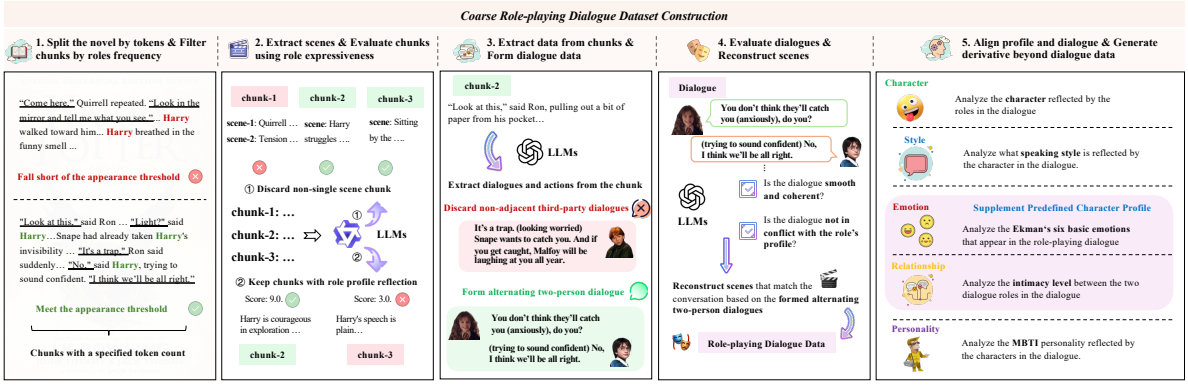


Figure 2: An schematic of our automated pipeline for constructing coarse Role-playing dialogue dataset.

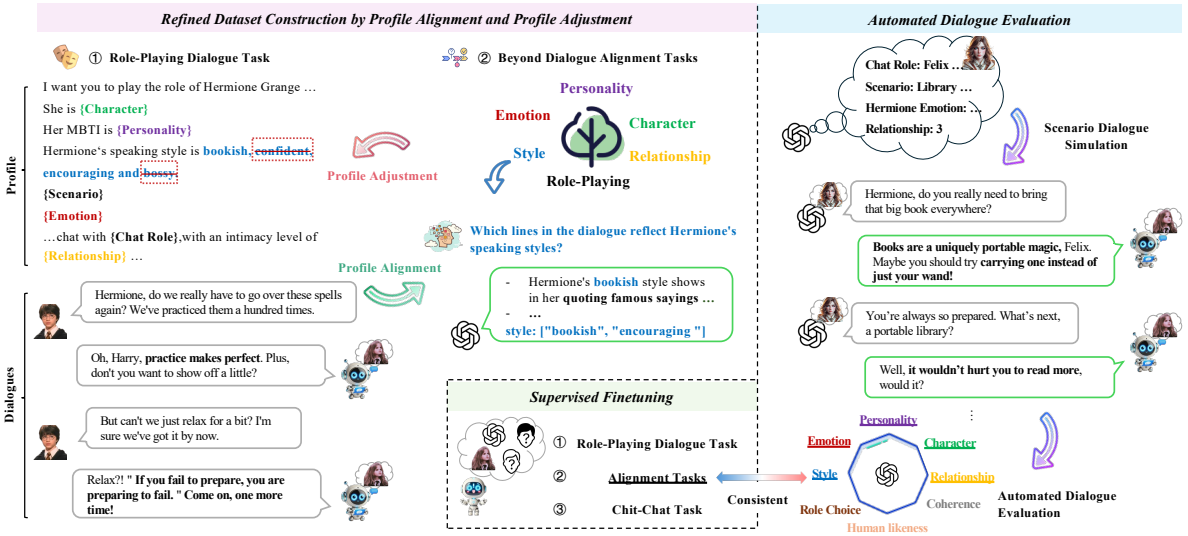


Figure 3: Illustration of Refined Dataset Construction, Supervised Finetuning, and Automated Dialogue Evaluation Framework. The left side shows the Refined Dataset Construction and Supervised Finetuning phases, the profile alignment results are utilized to adjust each scenario's dialogue profiles, eliminating training biases. On the right, the LLM generates random scenarios and roles for multi-turn dialogues with the model, followed by an evaluation using objective questions to obtain quantitative metrics of the model's role-playing capabilities.

3.1 Alignment Dataset Construction

To advance general role-playing, we require a diverse set of role settings and dialogue data. However, the dialogues generated by GPT models (Wang et al., 2024a; Li et al., 2023) lacks human likeness. The novel-extracted dialogues (Chen et al., 2023) better captures role nuances but these datasets are expensive and labor-intensive to produce, making them less accessible. Therefore, we propose a low-cost (as shown in § E), fully automated pipeline (Fig. 2) for constructing a role-playing dialogue dataset and introduce a method to address biases and fine-grained alignment issues:

3.1.1 Coarse Role-playing Dialogue Dataset Construction

We first segment the text and apply a role frequency threshold to filter chunks (Fig. 2.1). Given the large initial chunk volume, we employ open-source models to extract dialogue scenarios and evaluate chunks, retaining only those that reflect role traits and belong to single scenes (Fig. 2.2). Next, through systematic comparison of multiple LLMs, we select GPT-4o for its optimal performance in dialogue extraction and scene reconstruction (see Fig. 2 panels 3-4, with accuracy metrics and cost analysis in Tab. 5 of § A.2), while implementing rigorous verification to ensure dialogue coherence and profile consistency. This pipeline (with implementation details in § A.1) enables the low-cost, fully automated construction of a high-quality role-playing dialogue dataset.

3.1.2 Refined Dataset Construction by Profile Alignment and Profile Adjustment

Profile Alignment: After dialogue dataset construction, we use GPT-4o to align each multi-turn di-

alogue across five dimensions: Character, Style, Emotion, Relationship, and Personality (CSERP) as shown in Fig. 2.5. Through our innovative prompting mechanism (outlined in § F.2), LLMs are able to analyze dialogues through step-by-step reasoning, explicitly linking each sentence to corresponding profile traits, thereby achieving sentence-level alignment (see § G.2 for examples). This approach generates fine-grained CSERP alignment tasks, helping the model maintain its reasoning capabilities and enhancing its ability to perceive and adhere to profiles during training. All prompts used for dataset construction and alignment are provided in § F.1 and § F.2, respectively.

Profile Adjustment: Based on the alignment results, we dynamically adjust the profile settings for each dialogue to ensure consistency, as indicated by the red arrow in Fig. 3. Specifically, we refine the predefined profile by removing traits that are not reflected in the dialogue while incorporating scenario-relevant Emotional and Relational attributes. This addresses the common issue of bias between scene dialogues and role profiles in role-playing datasets, aligning the training inputs (Profile prompts) with outputs (dialogue labels), thereby ensuring effective training.

3.2 Supervised Finetuning

Using the constructed dataset, we fine-tuned LLMs with a mix of Aligned Role-Playing Dialogue data D_r, deriving Alignment CSERP data D_a, and Chit-Chat data D_c (sourced from open-source chit-chat datasets) in a 1:5:4 ratio. The D_r dataset ensures that model's output dialogue are aligned with corresponding role profile, and D_a focuses on explicit alignment tasks to refine profile-dialogue relationships. Meanwhile, D_c preserves the model's general conversational ability. The learning objective was to minimize the aggregate negative log-likelihood across all datasets, as described by:

$$\min_{\pi} \mathbb{E}_{(s,a)\sim D_r \odot D_a \odot D_c} \left[-\log \pi(a|s)\right]$$

where s is the prompt input, and a is the corresponding label, $\odot$ denotes training dataset mixing, and π denotes the trained LLMs. The total volume of training data, consistent with other comparative experiments, was regulated to be ten times that of the benchmark role-playing data, effectively managing the potential impact of data volume on LLM training. Training parameters are detailed in § D.

3.3 Automated Dialogue Evaluation

Evaluating the effectiveness and reproducibility of role-playing is still an open problem. As discussed in § 2, the diversity in dialogue generation and evaluation methods complicates establishing widely accepted standards.

Our next contribution is to propose an automated dialogue method, as shown on the right side of Fig. 3. We first generate a role and its description, aligning it with the worldview of the role to be evaluated. Based on their profiles, we create a dialogue scenario, design the emotions, and define the relationship between the roles. Finally, the two models engage in multi-turn dialogues within this context, producing a dialogue corpus for subsequent role-playing model evaluations. One of the models in the dialogue is GPT-4o, which compared to human evaluators, provides a low-cost and standardized approach, ensuring consistent evaluation across models (Wang et al., 2024b; Lu et al., 2024).

During the evaluation phase, we adopted the "LLMs as judge" (Kim et al., 2023) approach, designing five tasks based on the role profile dimensions as discussed in § 2: Character, Style, Emotion, Relationship, and Personality. Besides, we used Human-likeness to assess whether the model's output matches human expression and Coherence to evaluate dialogue continuity. Additionally, we introduced a role-based multiple-choice evaluation to assess the model's role consistency across multi-turn dialogues.We reformatted all evaluation tasks into objective questions, yielding lower variance and closer alignment with human assessments, compared to subjective ones (as shown in § B.1). This approach has a high level of acceptance in manual checks and closely mirrors human evaluation, making the evaluation more systematic and reliable for large-scale applications. All prompts used for dialogue generation and evaluation are available in § F.3.

4 Experiments

4.1 Experimental Setup

Dataset. Following the method in § 3.1, we created the role-playing dialogue dataset (**RP**) by extracting 280 Chinese and 31 English roles from 123 novels and scripts, resulting in 3,552 scenario dialogues with 23,247 turns, all from authentic sources. Compared to the other datasets listed in § A.3 Tab. 7, ours has the highest number of real roles and the highest turns with real dia-

logues, which is significant for building a general role-playing model. Additionally, we created the sentence-level alignment dataset (**CSERP**) by deriving five alignment training tasks from each of the 3,552 sessions, resulting in five training data points per session. After alignment, only 4.2% of the dialogues in the CSP were fully consistent with the predefined role profiles (see § A.3 for dataset statistics). Once the role profiles were aligned and adjusted with their dialogues, the data was transformed into the aligned role-playing dialogue dataset (**RPA**). The chit-chat dataset (**CC**) included both Chinese (NaturalConv (Wang et al., 2021)) and English (DailyDialog (Li et al., 2017)).

Baselines. We test both open-source and proprietary advanced bilingual (Chinese-English) chatbots in our assessment framework. Our proprietary general baselines include: **GPT-4o**, **GPT-3.5-Turbo** (OpenAI, 2024), **Yi-Large-Turbo** (Wanwu, 2024), **Deepseek-Chat** (Deepseek, 2024). These models represent the current state-of-the-art in language generation technology, covering different architectures and training methods. Additionally, we selected several **baselines focused on role-playing**: **Index-1.9B-Character** (bilibili, 2024) (open-source), **Baichuan-NPC-Turbo** (Baichuan, 2024) (proprietary), and **Character-GLM** (Zhou et al., 2024) (open-source). These models are specifically designed and optimized for role-playing, capable of generating more realistic character dialogues and interactions. Furthermore, to validate the effectiveness of our framework, we trained two open-source bilingual baselines: **Qwen2-7B-Instruct** (Yang et al., 2024) and **Mistral-Nemo-Instruct-2407** (MistralAI, 2024).

Metrics Design. As previously mentioned, our goal is to evaluate the model's ability to follow the role's profile and scenario setting, which aligns with the objectives of the derivative tasks of CSERP discussed in § A.1 Fig. 5. Consequently, the evaluation tasks and alignment tasks under the five dimensions—Character, Style, Emotion, Relationship, and Personality—are identical. Furthermore, we introduced three additional evaluation metrics: Human-likeness, Coherence, and Role Choice of the dialogues:

- **Emotion and Relationship**: Both Emotion, based on Ekman's six basic emotions (Ekman, 1992), and Relationship are rated by LLM on a 0-10 scale based on the evaluated dialogue. The scores are then used to calculate the Normalized Mean Absolute Percentage Error (NMAPE) against context labels generated by the prompt model.
- **Character, Style, and Personality**: Rated by LLM based on the dialogue and the role profile's candidate labels. Personality is assessed using the MBTI classification (binary recall), while Character and Style are multi-label recall tasks.
- **Human-likeness**: Evaluates the naturalness and realism of the interaction. LLM determines whether the dialogue is human or model-generated using few-shot prompting.
- **Role Choice**: Assesses how well the role is recognized in the dialogue. LLM selects the most appropriate role from four candidates after masking the role names in the dialogue.
- **Coherence**: Evaluates the logical consistency and contextual coherence of multi-turn dialogues. LLM checks whether the dialogue is coherent within the given context.

4.2 Main Results

We conducted 300 independent bilingual evaluations for each model, with each evaluation consisting of a scenario containing five turns of dialogue, all involving new roles and scenarios. Specific evaluating parameters are detailed in § D. The evaluation process was fully automated, with GPT-4o generating new chat roles and scenarios, and engaging in multiple turns of dialogue with the evaluated model (as illustrated by role-playing dialogue cases in § G.1). The generated dialogues were input into an automatic evaluation pipeline to obtain quantitative performance metrics for the role-playing models. These evaluation results were used to calculate the average performance of each model across five dimensions and three additional metrics.

We present the main results in Tab. 2. Due to the lack of role-playing training, the role-playing capabilities of the general baselines are closely related to their general capabilities. In the **Human-likeness** dimension, dialogues generated by GPT-4o have a 67.33% probability of being perceived as real human dialogues, significantly outperforming other general models. This is crucial for creating authentic dialogue scenario simulations.

Among the role-playing expertise baselines, the open-source models Index-1.9B-Character and

Model	Character	Style	Emotion	Relationship	Personality	Qualification -Rate$^{\triangle}$ ↑	Human -likeness ↑	Role Choice ↑	Coherence ↑
	Recall ↑	Recall ↑	NMAPE ↓	NMAPE ↓	Precision ↑				
General Baselines(Proprietary)									
GPT-4o	74.32 ± 1.15	81.67 ± 1.51	16.31 ± 0.48	12.13 ± 0.66	66.58 ± 4.41	46.33 ± 2.88	**67.33 ± 3.95**	87.33 ± 3.86	**99.67 ± 0.33**
GPT-3.5-Turbo	72.26 ± 1.27	73.66 ± 1.73	17.79 ± 0.56	14.17 ± 0.73	66.92 ± 4.85	39.0 ± 2.82	33.33 ± 4.43	83.00 ± 4.68	97.33 ± 1.17
Yi-Large-Turbo	75.13 ± 1.22	79.18 ± 1.58	16.44 ± 0.49	13.48 ± 0.67	68.25 ± 4.61	49.0 ± 2.89	47.00 ± 4.60	84.33 ± 3.67	92.67 ± 2.39
Deepseek-Chat	75.46 ± 1.14	81.49 ± 1.51	15.92 ± 0.46	12.42 ± 0.63	67.92 ± 4.57	49.33 ± 2.89	52.33 ± 4.95	83.00 ± 4.68	96.67 ± 1.00
Role-play Expertise Baselines									
Index-1.9B-Character	73.33 ± 1.32	76.48 ± 1.50	17.99 ± 0.53	13.58 ± 0.71	66.33 ± 4.57	41.67 ± 2.85	21.67 ± 3.96	78.67 ± 5.14	69.67 ± 3.85
CharacterGLM-6B	73.36 ± 1.28	76.08 ± 1.55	18.58 ± 0.55	14.27 ± 0.79	67.33 ± 4.34	36.0 ± 2.77	16.00 ± 2.38	81.00 ± 4.40	25.67 ± 3.48
Baichuan-NPC-Turbo	75.19 ± 1.23	79.15 ± 1.38	17.24 ± 0.51	13.10 ± 0.69	65.33 ± 4.84	49.33 ± 2.82	56.00 ± 4.66	86.33 ± 4.90	99.00 ± 0.56
Custom Trained Baselines									
Mistral-Nemo-Instruct-2407	74.12 ± 1.17	77.04 ± 1.48	17.00 ± 0.43	13.50 ± 0.67	67.00 ± 4.30	39.0 ± 2.82	53.67 ± 4.66	82.67 ± 4.77	74.33 ± 3.77
+ RP & CC	71.56 ± 1.26	74.66 ± 1.51	17.36 ± 0.49	12.58 ± 0.69	68.17 ± 4.32	34.67 ± 2.75	53.67 ± 3.76	86.33 ± 4.22	86.33 ± 2.42
+ RPA & CC	74.44 ± 1.14	77.63 ± 1.40	16.74 ± 0.46	12.07 ± 0.67	69.50 ± 4.31	43.67 ± 2.86	56.00 ± 3.41	85.00 ± 4.49	91.67 ± 1.80
+ RPA & CC & CSERP	74.58 ± 1.28	78.47 ± 1.45	16.62 ± 0.48	11.38 ± 0.67*	69.08 ± 4.46	47.33 ± 2.88*	59.00 ± 4.46	87.00 ± 4.73	92.67 ± 1.59
Qwen2-7B-Instruct	75.39 ± 1.13	77.68 ± 1.65	17.64 ± 0.56	13.43 ± 0.7	67.75 ± 4.44	37.33 ± 2.79	48.00 ± 4.66	83.33 ± 4.48	99.00 ± 0.56
+ RP & CC	74.91 ± 1.21	78.59 ± 1.39	16.52 ± 0.48	12.28 ± 0.67	67.00 ± 4.31	38.0 ± 2.8	56.67 ± 3.85	82.00 ± 4.90	90.00 ± 1.59
+ RPA & CC	76.43 ± 1.18	81.28 ± 1.37	16.10 ± 0.45	12.35 ± 0.67	67.83 ± 4.36	44.67 ± 2.87	62.33 ± 3.21	87.33 ± 3.77	92.67 ± 1.85
+ RPA & CC & CSERP	**78.67 ± 1.12***	**82.52 ± 1.33***	**15.68 ± 0.5***	**11.22 ± 0.72***	**69.67 ± 4.27**	**56.33 ± 2.86***	64.33 ± 3.80*	**87.33 ± 3.74**	99.00 ± 0.56

Table 2: Main results of **BEYOND DIALOGUE**. We report the average scores with their standard error of the mean (SEM). **Bold** numbers indicate the highest scores, while underlined numbers are the best in the group, * denotes statistically significant improvements over the untrained baseline (achieving $p < 0.05$ in t-test). The **Qualification Rate (QR)**$^{\triangle}$ indicates the proportion of role-playing dialogues that align with the predefined profile, scoring above 60 in each *C-S-E-R-P* dimension, with *E* and *R* calculated as 1-NMAPE. A darker background indicates better performance. More baseline results are provided in § B.2 Tab. 9. All dialogues are evaluated by GPT-4o as judges, with supplementary evaluations provided by Claude-3-Opus in § B.2 Tab. 10, yielding comparable results.

characterGLM-6B, which were fine-tuned using only role-playing data, showed significantly lower performance in **Human-likeness** and **Coherence** compared to other models. In contrast, the proprietary model Baichuan-NPC-Turbo demonstrated a significant improvement in **Human-likeness**.

In our experiments, the model trained with unaligned role-playing data (+RP & CC) exhibited training biases, resulting in limited or even diminished performance in Qualification-Rate. In contrast, the model trained with aligned data (+RPA & CC) achieved significant improvements, emphasizing the importance of data alignment in mitigating training bias. Furthermore, the model trained with additional aligned data (+RPA & CC & CSERP) outperformed the original model by 19%, and surpassed GPT-4o by 10%.

Under the **BEYOND DIALOGUE** framework, the fully enhanced Qwen-2-7B models obtained through supervised fine-tuning achieved the highest scores in most dimensions, with its human-likeness significantly outperforming other baselines except GPT-4o, and showing statistically significant improvements across five dimensions compared to the untrained baseline ($p < 0.05$ in t-test). Adding CSERP data not only fine-tunes the alignment of dialogue and profile across these five dimensions but also enhances the model's logical and contextual coherence abilities through reasoning training, which is well reflected in the **Coherence** metric.

Manual verification shows a 93% consistency between GPT-4o and human evaluations (§ C, Tab. 11), confirming the reliability of our approach. Overall, these results highlight the importance of aligning role profiles and dialogues in role-playing training, and demonstrate that "Beyond Dialogue" tasks help models better adhere to role profiles.

4.3 Alignment Results

To explore the model's capability in fine-grained alignment for dialogue and role profiles, we randomly sampled 100 data points from each dimension of CSERP in the evaluation results (§ 4.2). Using GPT-4o's results as the reference labels, we present the alignment results of the comparable baseline model and two improved models in Tab. 3. The alignment for Character, Style, and Personality was assessed using recall, while Emotion and Relation were evaluated using NMAPE.

The fine-tuned Mistral-Nemo model, achieved

Model	Character ↑	Style ↑	Emotion ↓	Relationship ↓	Personality ↑	Avg. ↑
Comparable Baselines						
Yi-1.5-9B-Chat	79.0 ± 2.3	75.4 ± 2.5	17.1 ± 0.9	15.3 ± 1.5	70.7 ± 2.4	78.5 ± 1.9
GLM-4-9-chat	72.2 ± 2.3	77.8 ± 2.2	12.6 ± 0.7	12.7 ± 1.2	78.0 ± 2.4	80.5 ± 1.7
Custom Trained Baselines						
Mistral-Nemo-Instruct-2407	52.8 ± 3.6	54.3 ± 4.2	13.5 ± 0.7	10.9 ± 0.8	77.2 ± 2.3	71.9 ± 2.3
+ RPA & CC & CSERP	**89.1 ± 1.6**	**88.7 ± 2.2**	**10.1 ± 0.5**	**7.4 ± 0.8**	**87.0 ± 1.7**	**89.4 ± 1.4**
Qwen2-7B-Instruct	51.6 ± 2.27	51.23 ± 2.98	20.6 ± 0.9	17.1 ± 1.6	52.7 ± 3.7	63.5 ± 2.3
+ RPA & CC & CSERP	86.3 ± 2.1	81.9 ± 2.5	13.1 ± 0.7	8.9 ± 0.9	82.7 ± 1.8	85.7 ± 1.6

Table 3: Comparison of Baseline Models on Dialogue and Role Profile Alignment Tasks. The results are benchmarked against GPT-4o's alignment performance.

top performance, closely matching GPT-4o's 89.4% effectiveness. Both Mistral and Qwen showed notable improvement, with recall rates for Character and Style increasing by over 30%, while NMAPE for Emotion and Relation decreased significantly. This boost allowed the 7B-parameter Qwen2 model to surpass the 9B-parameter Yi1.5 and GLM4 models in alignment tasks.

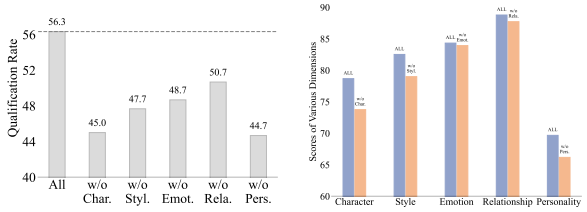
However, when trained on the **RPA & CC** or **RP & CC** datasets alone, both Qwen and Mistral struggled with alignment tasks. Their reasoning capabilities declined, resulting in ineffective outcomes. Detailed case studies are in § G.2.

The results from Tab. 2 and Tab. 3 indicate that enhancing the model's ability to align dialogue and role profiles effectively translates into improved role-playing capabilities. Since these **Alignment Tasks** are consistent with the **Evaluation Tasks** across the five dimensions of CSERP, improvements in alignment ability also signify enhanced model role-playing evaluation capabilities. This is beneficial for establishing an effective **feedback-enhanced loop for the role-playing**.

4.4 Ablation Study

To assess the contribution of the five CSERP alignment tasks within the **BEYOND DIALOGUE** framework, we conducted an ablation study. Each alignment task in D_a was removed individually, with its training data volume replaced by an equivalent amount of D_c to maintain consistency. We evaluated the model on the five role-playing dimensions using the same methods in Tab. 2.

Tab. 4 presents the results of the ablation experiments for CSERP training tasks. Notably, the ablation of **w/o Char.** and **w/o Pers.** have the most significant impact, with a marked decline in the performance of **Qualification Rate**. The other three metrics also showed varying degrees of decline, indicating that these fine-grained alignment tasks are crucial for the model's general role-playing capability.



(a) Qualification Rate (b) Ablation Dimension

Figure 4: Comparison of Scores for Ablation Across Five Dimensions in CSERP (Left) and Performance Scores for Specific Dimensions (Right).

Model	Character ↑	Style ↑	Emotion ↓	Relationship ↓	Personality ↑	QR ↑
w/o Char.	73.75 ± 1.24	81.43 ± 1.40	15.94 ± 0.45	11.75 ± 0.66	67.08 ± 4.56	45.0 ± 2.87
w/o Style.	76.94 ± 1.22	78.99 ± 1.38	15.87 ± 0.45	12.07 ± 0.69	67.25 ± 4.42	47.67 ± 2.88
w/o Emot.	**78.69 ± 1.22**	80.11 ± 1.38	16.08 ± 0.46	11.82 ± 0.65	67.58 ± 4.92	48.67 ± 2.89
w/o Rela.	77.69 ± 1.29	82.33 ± 1.39	15.81 ± 0.46	12.28 ± 0.68	68.92 ± 4.48	50.67 ± 2.89
w/o Pers.	78.38 ± 1.20	79.34 ± 1.42	16.14 ± 0.45	11.93 ± 0.66	66.17 ± 4.49	44.67 ± 2.87
All	78.67 ± 1.12	**82.52 ± 1.33**	**15.68 ± 0.50**	**11.22 ± 0.72**	**69.67 ± 4.27**	**56.33 ± 2.86**

Table 4: Ablation Results on CSERP Training Tasks.

Fig. 4 compares scores across the five CSERP dimensions, illustrating how ablation of individual tasks affects both overall and specific performance.

The ablation study clearly demonstrates that each alignment task within the **BEYOND DIALOGUE** framework uniquely contributes to the model's overall performance. The integration of all tasks yields the best results, underscoring the importance of these alignment tasks in training a robust general role-playing model.

5 Conclusion

This paper introduces a novel general role-playing framework called **BEYOND DIALOGUE**. We propose a straightforward method that effectively aligns dialogues in specific scenarios with role profiles. This alignment helps to eliminate distortions arising from biases between predefined profiles and the dialogues generated during training. Additionally, we present an innovative prompting mechanism that constructs "Beyond Dialogue" training tasks by generating reasoning processes. It enables fine-grained alignment between role profiles and dialogues at the sentence level. In terms of evaluation, we shift from traditional subjective assessments to an objective, efficient, and reproducible method. Experimental results demonstrate that our approach enhances the model's ability to follow predefined profiles across various dimensions of general role-playing, surpassing most general and specialized role-playing baselines.

6 Limitations

While our framework addresses the challenge of LLMs adhering to predefined role profiles, real-world scenarios often involve role profiles that change as the dialogue progresses (e.g., Emotions). However, due to the complexity and cost of implementing such tasks, automatically adjusting role profiles based on dialogue evolution and contextual shifts remains an open area for future research.

Additionally, current role-playing models rely on manually predefining scenario-specific profiles. Enabling models to autonomously reason about, align with, and adapt globally-defined role profiles based on evolving conversational contexts—rather than relying on rigid human-crafted definitions—presents another critical direction for advancing dynamic role adaptation.

7 Ethical Statement

Use of Human Annotations. Our research involves human annotations for tasks like collecting novel content, character profiles, and validating data accuracy. We employ trained professionals with relevant expertise and ensure fair compensation above the local minimum wage. Informed consent is obtained, and we maintain transparency in the use of annotations. Privacy and ethical standards are prioritized to create a respectful research environment.

Risks. Role-play LLMs trained under the **BEYOND DIALOGUE** Framework may exhibit only the basic safety alignment of the underlying training model, which means they could potentially generate harmful or toxic content when prompted. As a result, these role-play models are intended solely for research purposes and will require careful alignment for safety in future iterations.

References

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

Anthropic. 2024. Claude-3-opus. https://claude.ai/. Accessed: 2024-08-01.

Baichuan. 2024. Baichuan4 and baichuan-npc-turbo. https://www.baichuan-ai.com/home. Accessed: 2024-08-01.

Baidu. 2024. Ernie-bot. https://yiyan.baidu.com/. Accessed: 2024-08-01.

bilibili. 2024. Index1.9b technical report. https://github.com/bilibili/Index-1.9B. Accessed: 2024-08-01.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, et al. 2024a. From persona to personalization: A survey on role-playing language agents. *Preprint*, arXiv:2404.18231.

Nuo Chen, Y Wang, Yang Deng, and Jia Li. 2024b. The oscars of ai theater: A survey on role-playing with language models. *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.

Daixuan Cheng, Shaohan Huang, and Furu Wei. 2023. Adapting large language models via reading comprehension. *Preprint*, arXiv:2309.09530.

Deepseek. 2024. Deepseek-chat. https://www.deepseek.com/. Accessed: 2024-08-01.

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

Paul Ekman. 1992. Are there basic emotions?

Seungju Han, Beomsu Kim, Jin Yong Yoo, Seokjun Seo, Sangbum Kim, Enkhbayar Erdenee, and Buru Chang. 2022. Meet your favorite character: Open-domain chatbot mimicking fictional characters with only a few utterances. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5114–5132, Seattle, United States. Association for Computational Linguistics.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, et al. 2023. Chatharuhi: Reviving anime character in reality via large language model. *Preprint*, arXiv:2308.09597.

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 8th International Joint Conference on Natural Language Processing (IJCNLP 2017)*.

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.

MistralAI. 2024. Mistral-nemo-instruct-2407. https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407. Accessed: 2024-08-01.

Moonshot. 2024. Moonshot-v1-8k. https://kimi.moonshot.cn/. Accessed: 2024-08-01.

OpenAI. 2024. Gpt-4o and gpt-3.5-turbo. https://openai.com/. Accessed: 2024-08-01.

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

Yihong Tang, Jiao Ou, Che Liu, Fuzheng Zhang, Di Zhang, and Kun Gai. 2024. Erabal: Enhancing role-playing agents through boundary-aware learning. *Preprint*, arXiv:2409.14710.

Meiling Tao, Liang Xuechen, Tianyu Shi, Lei Yu, and Yiting Xie. 2024. RoleCraft-GLM: Advancing personalized role-playing in large language models. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 1–9, St. Julians, Malta. Association for Computational Linguistics.

Tencent. 2024. Hunyuan. https://hunyuan.tencent.com/. Accessed: 2024-08-01.

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.

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. Learning to speak and act in a fantasy text adventure game. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 673–683, Hong Kong, China. Association for Computational Linguistics.

Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024a. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand. Association for Computational Linguistics.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, et al. 2023. Instructuie: Multi-task instruction tuning for unified information extraction. *Preprint*, arXiv:2304.08085.

Xiaoyang Wang, Chen Li, Jianqiao Zhao, and Dong Yu. 2021. Naturalconv: A chinese dialogue dataset towards multi-turn topic-driven conversation. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI-21)*.

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. 2024b. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873, Bangkok, Thailand. Association for Computational Linguistics.

Lingyi Wanwu. 2024. Yi-large-turbo. https://www.lingyiwanwu.com/. Accessed: 2024-08-01.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, et al. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Phillip B Zarrilli. 2005. *Acting (re) considered: a theoretical and practical guide*. Routledge.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, et al. 2024. CharacterGLM: Customizing social characters with large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1457–1476, Miami, Florida, US. Association for Computational Linguistics.

A Details of Dataset Construction

A.1 Dataset construction process

We need to manually collect and annotate character traits, speaking styles, and personality labels from the novels. The personality labels will be primarily sourced from the Personality Database[1]. Following this, we will proceed with the fully automated role-playing dialogue data construction process (see Fig. 2):

1. **Chunk Split and Filter**: To manage the token limit of LLMs, the novel is split into fixed-token chunks. Only chunks that meet the role's appearance threshold are retained, reducing the number of chunks and associated costs in subsequent stages, while ensuring that relevant roles are captured.

2. **Scenario Extract and Chunk Evaluation**: Open-source large models are utilized to extract scenarios from the chunks (prompt in Tab. 13), filtering out those that contain multiple scenarios to avoid discontinuous dialogues. Given the large number of input chunks, using specialized models can be costly. After comparison, we selected Qwen1.5-72B-Chat (Yang et al., 2024) for its balance of performance and efficiency. Additionally, chunks are evaluated to keep only those that reflect the role's profile (prompt in Tab. 14), helping to eliminate non-matching chunks and reduce the number of chunks in subsequent inputs.

3. **Dialogue Extract**: Utilizing powerful LLMs to extract dialogues and actions between roles from valid chunks is crucial (prompt in Tab. 15), as role-playing dialogue data is fundamental for both role-playing training and subsequent derivative tasks. After comparing mainstream models, we selected GPT-4o (gpt-4o-2024-05-13) for its cost-effectiveness. § A.2 details the extraction accuracy and cost of each LLM we evaluated.

4. **Dialogue Check**: To ensure valid role-playing dialogue data, we first focus on forming alternating two-person dialogues, which often leads to scenarios shrinking. This necessitates simultaneous scenario reconstruction and coherence checks to maintain contextual integrity (prompt in Tab. 16). After ensuring the coherence, we employ GPT-4o to review role portrayal accuracy, discarding any dialogues that conflict with the established profiles (prompt in Tab. 17).

5. **Profile Alignment**: We use GPT-4o to analyze each multi-turn dialogue, ensuring alignment of the role-playing data across the five dimensions: Character, Style, Emotion, Relationship, and Personality (CSERP, see Fig. 5):

   (a) **Character and Style**: These are aligned using a word recall method, where GPT-4o recalls descriptive words from the role profile that are reflected in the dialogue within the scenario's scope (prompt in Tab. 18 and Tab. 19).

   (b) **Emotion and Relationship**: These are aligned using a scale method, with GPT-4o providing a score from 0-10 based on the dialogue content. The Emotion dimension is based on Ekman's six basic emotions (Ekman, 1992): Anger, Disgust, Fear, Happiness, Sadness, and Surprise (prompt in Tab. 20). The Relationship dimension reflects the intimacy presented in the dialogues within the scenario's scope (prompt in Tab. 21).

   (c) **Personality**: We use the Myers-Briggs Type Indicator (MBTI)[2], which consists of four dimensions, each with two types. GPT-4o performs binary classification alignment for the four dimensions within the scenario's scope (prompt in Tab. 22).

6. **Profile Adjustment**: In the role profile, Character, Style, and Personality are predefined, so we adjust each dialogue's profile prompt based on the alignment results from Step 5. Since Emotion and Relationship are scene-dependent and cannot be predefined, we need to supplement the profile prompting the dialogue with information on these two dimensions.

A.2 Comparative Analysis of Dialogue Extraction

This section presents a comparative analysis of LLMs for novel dialogue extraction, focusing on accuracy and cost. The models compared include

[1] https://www.personality-database.com/

[2] https://www.16personalities.com/

Figure 5: An diagram of Deriving the Character-Style-Emotion-Relationship-Personality (CSERP) Dataset from Role-Playing Dialogue Data.

Qwen1.5-72B-Chat (Yang et al., 2024), ERNIE-Bot (Baidu, 2024), deepseek-chat (Deepseek, 2024), moonshot-v1-8k (Moonshot, 2024), and GPT-4o (OpenAI, 2024). Key metrics are dialogue extraction accuracy, omission recall rate, and API costs. Note that Qwen, being open-source, incurs no API costs. A detailed comparison is provided in Tab. 5.

model	Recall	Accuracy	Pricing ($)
Qwen1.5-72B-Chat	67.7%	71.4%	-
ERNIE-Bot	85.4%	91%	1.66 / 1M input tokens 1.66 / 1M output tokens
deepseek-chat	83.2%	92.2%	0.14 / 1M input tokens 0.28 / 1M output tokens
moonshot-v1-8k	73.1%	79.2%	1.66 / 1M input tokens 1.66 / 1M output tokens
gpt-4o	**89.1%**	**96.4%**	5.00 / 1M input tokens 15.00 / 1M output tokens

Table 5: Comparative Accuracy and Cost Analysis of Dialogue Extraction LLMs

High accuracy and recall rates are essential for maintaining the integrity and consistency of extracted dialogues. The quality of role-playing dialogue data significantly affects the quality of derived CSERP tasks. Compared to GPT-4o, the accuracy gap in the other LLMs is substantial, resulting in a notable decrease in the proportion of extracted dialogues passing the quality check. Consequently, producing the same amount of high-quality role-playing dialogue data with these LLMs incurs higher costs. Therefore, we ultimately chose GPT-4o for extracting character dialogues and actions from text segments.

A.3 Statistical Analysis of BEYOND DIALOGUE Role-Playing Dataset

Following the method outlined in § A.1, we extracted 280 Chinese roles and 31 English roles from 123 Chinese and English novels or scripts. In total, 3,552 sessions of scenario dialogues were obtained, comprising 23,247 dialogue turns, all drawn from authentic dialogues in novels or scripts (Tab. 6).

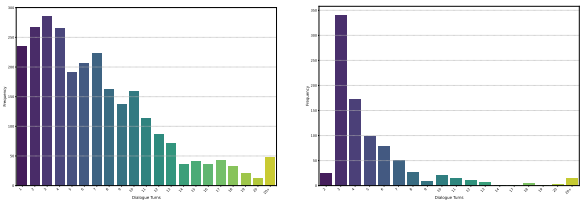
Tab. 7 summarizes the statistical information of various role-playing dialogue datasets. From the table, it can be seen that our dataset is entirely sourced from novels. However, this source provides richer and more authentic dialogue scenarios. Additionally, compared to other datasets, we have the highest number of real roles and the most sessions of authentic dialogues.

In Fig. 6, the distribution of dialogue turns in both Chinese and English from our role-playing dataset is presented, illustrating the variation in dialogue lengths.

Alphabetical List of Novel and Script Titles in English

A Dream of Splendor, A Record of a Mortal's Journey to Immortality, A Slight Smile is Very Charming, A Song of Ice and Fire, All is Well, Battle Through the Heavens, Better Days, Better Days (Novel), Big Shot, Black Moonlight Hold Firm (Drama Script), Blade of the Immortal, Border Town, Bright Sword, Butterfly, Can't Hide Love, Candle in the Tomb, Chinese Paladin 3, Chronicles of a Blood Merchant, Crouching Tiger, Day and Night, Deep Love and Rainy Weather, Demi-Gods and Semi-Devils, Detective Chinatown 2, Diamond Lover, Do You Know? The Green Should be Plump and the Red Lean, Dream of the Red Chamber, Empresses in the Palace, Ever Night, Farewell My Concubine, Fighter of the Destiny, First Love, Grandmaster of Demonic Cultivation, Guo Degang's Comedy Collection, Half-demon Tsukasa, Hamlet, Handsome Siblings, Hard to Coax, Harry Potter, Heaven Official's Blessing, Hello Mr. Billionaire, Hi, Hidden Dragon, Home with Kids (Season 1 Episodes 1-2), How Long Will I Love U, Howl's Moving Castle, IPartment 1, IPartment 2 (Drama Script Excerpt), Important Things in Life, In the Mood for Love, In the Name of the People, Infernal Affairs, Journey Under the Midnight Sun, Journey to the West, Joy of Life (Novel), Joy of Life (Season 1 Episodes 1-2), Kong Yiji, Kung Fu, Longing Heart, Love As The Goal, Love in a Puff, Meeting You, Meteor, Mom (Drama Script Excerpt), My Homeland, My Own Swordsman, My People, Nirvana in Fire, Nirvana in Fire, Not Allowed to Die, Ode to Joy (Drama Script Excerpt), One and Only, Outlaws of the Marsh, Passing by Your World, Proud Wanderer, Qin's Moon, Red Sorghum Clan, Silent Separation, Snow in Midsummer, Soldier Assault, Soul Land, Spirited Away, Sword, The Adventures of Chu Liuxiang, The Bad Kids, The Bride with White Hair, The Deer and the Cauldron, The Devotion of Suspect X, The Flowers of War, The Four Great Constables, The Godfather, The Great Dao Commander, The Heaven Sword and Dragon Saber, The Journey of Flower, The King's Avatar, The Left Ear, The Legend of Lu Xiaofeng, The Legend of the Condor Heroes, The Liancheng Swordsman, The Longest Day in Chang'an, The Lost Tomb, The Lotus Flower Pavilion, The Masterless Master, The Mute Mansion, The Mystic Nine, The Newsroom (Drama Script Excerpt), The Orchid's Oath, The Parasitic Son-in-Law, The Return of the Condor Heroes, The Romance of Tiger and Rose, The Smiling, The Speed of Life, The Story of Yanxi Palace , The Story of the Cook's Camp (Drama Script Excerpt), The Sword Stained with Royal Blood, The Sword of the Third Young Master, The Swordsman, The Three-Body Problem, The Wandering Earth, The World Beneath (Text Excerpt), Tiny Times, To Live, To Try the World, To the Place with Wind (Drama Script Excerpt), Unrequited Love, White Deer Plain, With the Family (Drama Script Excerpt), You Are My Glory, Youth, Zhu Xian

Table 6: List of Novels and Scripts Titles



(a) Chinese (b) English

Figure 6: Distribution of Dialogue Turns in Our Role-Playing Dialogues Dataset

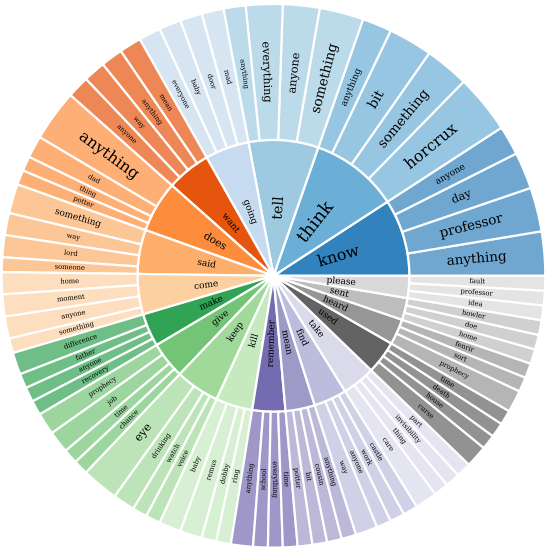


Figure 7: Verb-noun Structure of Our Role-playing Instructions in English.

(a) Character (b) Styles

Figure 8: Word Cloud of Character Traits and Speaking Styles in Role Profiles

The verb-noun structure of role-playing instructions in English is illustrated in Fig. 7, with the inner circle representing the top 20 verbs and the outer circle listing the direct noun objects.

The word clouds in Fig. 8 illustrate the character traits and speaking styles present in our role profiles.

As illustrated in Fig. 9a and 9b, the distribution of emotion and relationship values is presented, both rated on a scale from 0 to 10. The emotion distribution is based on the highest emotion score for each dialogue. Additionally, Fig. 9c shows the varied distribution of MBTI personality types among the role profiles in the dataset.

After GPT-4o aligned the predefined profiles with the scenario dialogues, we observed the following non-alignment ratios: 66.84% for Character, 80.66% for Personality, and 41.84% for Style, all deviating from the predefined profiles. Since Emotion and Relationship are not part of the predefined profiles, their non-alignment ratios were not computed.

B Supplementary Experimental Results

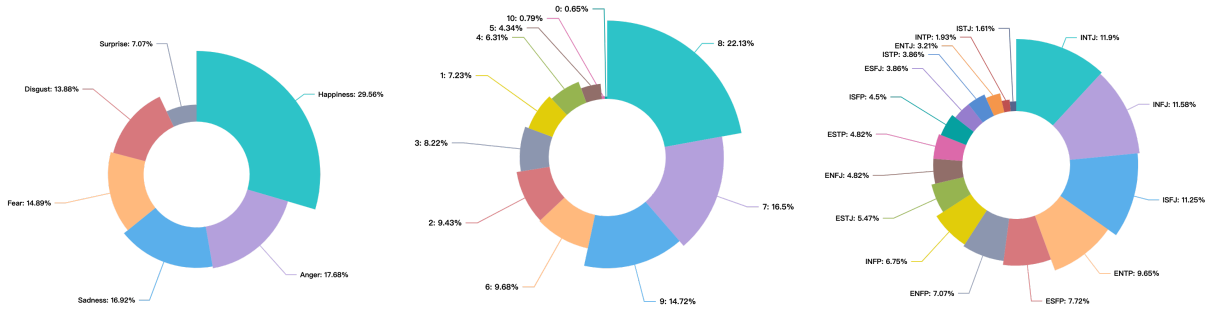
B.1 Subjective vs. Objective Evaluation Results

To assess the impact of evaluation formats on model consistency, we reformulated the evaluation tasks from subjective scoring (Wang et al., 2024a; Tu et al., 2024) to objective formats, such as multiple-choice and true/false questions, incorporating Chain-of-Thought (CoT) prompting to minimize variance.

We conducted an experiment with 50 evaluation instances, using mean scores from three human annotators across three settings: subjective consistency scores (Sub), objective questions without CoT (Obj w/o CoT), and objective questions with CoT (Obj), each repeated three times to calculate the mean and standard deviation for the five dimensions. The subjective scores (Sub) were scaled from 0∼5 to 100, while the objective evaluations

Dataset	Source	Open-source	Multi-lingual	Multi-turn	# Roles	# Sessions	# Turns	# Avg Turns
HPD	Novel	✓	✓	✓	-	1042	14380	13.8
CharacterGLM	Novel & Human & GPT	✗	✗	✓	250	1034	16316	15.78
RoleLLM	GPT	✓	✓	✗	Zh: 5, En: 95	-	23463	-
CharacterLLM	GPT	✓	✗	✓	9	1600	21120	13.2
PIPPA	Human	✓	✗	✓	1254	26000	1049015	40.34
ChatHaruhi	Novel & GPT	✓	✗	✓	32	54726	67660	1.23
WIKIROLE	GPT	✓	✓	✓	Zh: 3184, En: 3902	7086	36164	5.1
Ours	Novel	✓	✓	✓	Zh: 280, En: 31	3552	23247	6.54

Table 7: Comparison of Role-Playing Dialogue Datasets: Our dataset vs. Existing Role-Playing Datasets. Note: In the HPD dataset, the number of roles is denoted as "-" since it is exclusively centered on Harry Potter, with other characters interacting with him. RoleLLM is single-turn, so # Sessions and # Avg Turns are marked as "-".



(a) Emotion (b) Relationship (c) Personality

Figure 9: Distribution of Emotion Values, Relationship Values, and MBTI Personality Types Among Role Profiles in the Dataset

(Obj) and Human-Annotated followed the methods outlined in § 4.1.

Settings	*Character*	*Style*	*Emotion*	*Relationship*	*Personality*
Sub	83.2 ± 9.8	86.8 ± 10.4	83.5 ± 6.5	89.2 ± 5.8	80.5 ± 8.5
Obj w/o CoT	80.2 ± 3.4	83.8 ± 4.2	85.4 ± 1.4	88.6 ± 1.2	71.2 ± 3.8
Obj	78.7 ± 1.4	82.3 ± 1.8	83.2 ± 0.9	86.7 ± 0.8	68.2 ± 2.3
Human	**76.4 ± 1.2**	**81.2 ± 1.3**	**82.8 ± 0.3**	**87.2 ± 0.4**	**66.8 ± 1.7**

Table 8: Comparison of Sub. and Obj. Evaluation with Human-Annotated Results

Tab. 8 indicate that subjective evaluations exhibited significantly higher standard deviation and a tendency toward inflated scores compared to human annotations. On the other hand, objective evaluations, especially those incorporating CoT, demonstrated lower variance and were more closely aligned with human-level assessments.

B.2 Cross Model Validation for Main Results

This appendix complements the main results presented in the paper by providing additional insights into the bilingual chatbot baselines used in our evaluation. In addition to our proprietary general baselines, we have also tested newly added open-source general baselines. The proprietary models include **Moonshot-v1-8k** (Moonshot, 2024), **Baichuan4** (Baichuan, 2024), and **Hunyuan** (Tencent, 2024). The newly tested open-source baselines include **Yi-1.5-9B-Chat**, **GLM-4-9b-chat**.

Based on the experimental results, we selected **Qwen2-7B-Instruct** and **Mistral-Nemo-Instruct-2407** for our training due to their strong human-likeness, though their Qualification-Rate lags behind other open-source baselines. We aim for significant improvements in Qualification-Rate with our framework. Moreover, an important factor in selecting **Mistral-Nemo-Instruct-2407** was its primary focus on English, in contrast to the other three models—**Yi-1.5-9B-Chat**, **GLM-4-9b-chat**, and **Qwen2-7B-Instruct**—which are predominantly Chinese-language models. Including Mistral in our training set was crucial for ensuring linguistic diversity, thereby enhancing the model's capability to handle bilingual tasks effectively.

In Tab. 2, we used GPT-4o as the primary evaluation model. To further validate the performance, we also employed the Claude-3-Opus (Anthropic, 2024) model in Tab. 10, which shares similar capabilities with GPT-4o. While GPT-4o exhibited noticeable declines across several metrics, likely due to its inherent preference for generating its own content, the results from Claude-3-Opus were con-

Model	Character	Style	Emotion	Relationship	Personality	Qualification -Rate$^{\triangle}$ ↑	Human -likeness ↑	Role Choice ↑	Coherence ↑
	Recall ↑	Recall ↑	NMAPE ↓	NMAPE ↓	Precision ↑				
General Baselines(Proprietary)									
Moonshot-v1-8k	**74.06 ± 1.19**	**80.64 ± 1.51**	**16.17 ± 0.47**	13.42 ± 0.70	67.00 ± 4.87	**45.67 ± 2.88**	44.00 ± 4.33	**86.67 ± 3.75**	**99.33 ± 0.46**
Baichuan4	71.82 ± 1.25	76.92 ± 1.52	17.57 ± 0.52	**12.30 ± 0.62**	**67.08 ± 4.75**	41.33 ± 2.84	45.33 ± 4.31	82.33 ± 4.49	**99.33 ± 0.46**
Hunyuan	73.77 ± 1.18	78.75 ± 1.56	17.24 ± 0.48	13.22 ± 0.68	67.00 ± 4.39	42.0 ± 2.85	**53.00 ± 4.29**	84.33 ± 4.52	98.33 ± 0.84
General Baselines(Open-source)									
Yi-1.5-9B-Chat	75.31 ± 1.20	76.78 ± 1.49	**16.67 ± 0.52**	**12.75 ± 0.66**	67.42 ± 4.63	41.33 ± 2.84	38.67 ± 4.39	**84.00 ± 4.61**	92.67 ± 1.79
GLM-4-9b-chat	74.26 ± 1.19	**78.40 ± 1.55**	17.18 ± 0.50	14.48 ± 0.74	67.17 ± 4.93	**43.33 ± 2.86**	47.67 ± 4.25	83.33 ± 4.51	**99.33 ± 0.46**
Mistral-Nemo-Instruct-2407	74.12 ± 1.17	77.04 ± 1.48	17.00 ± 0.43	13.50 ± 0.67	67.00 ± 4.30	39.0 ± 2.82	**53.67 ± 4.66**	82.67 ± 4.77	74.33 ± 3.77
Qwen2-7B-Instruct	**75.39 ± 1.13**	77.68 ± 1.65	17.64 ± 0.56	13.43 ± 0.7	**67.75 ± 4.44**	37.33 ± 2.79	48.00 ± 4.66	83.33 ± 4.48	99.00 ± 0.56

Table 9: Main Results of Supplementary Baselines in **BEYOND DIALOGUE**

sistent with those of GPT-4o, thereby confirming the robustness of our evaluation approach.

C Human Verification of Evaluation

C.1 Human Annotation Process

For our human annotation process, participants were recruited from Mainland China, where they were employed by a company that specializes in data annotation. All participants were proficient in English, ensuring that the annotation tasks were completed accurately and efficiently. The compensation for these annotators was set at $16 per hour, which corresponds to roughly $1 per item annotated. This pay rate was determined based on the expected complexity of the tasks and the standard industry rates for such work in China.

Given the demographic characteristics of our participants (Chinese employees with strong English skills), this wage is considered fair within the local context. The annotators were able to work at a pace that aligned with this compensation, and the hourly rate was calibrated to account for the expected time it would take to complete each task. The average task duration was calculated through pilot testing, which helped us determine that $16 per hour was appropriate for the time investment required.

C.2 Human Verification Results

In this section, we present our approach to verifying the evaluations conducted by GPT-4o through human assessment. Specifically, we sampled 100 dialogue evaluations generated by GPT-4o and manually checked the reasonableness of these evaluations. The dialogue evaluations were divided into two parts: Auto-Generation and Auto-Evaluation.

For the Auto-Generation part, GPT-4o's task was to generate dialogue scenarios and corresponding scenario settings. Human judges evaluated the validity of the content generated by GPT-4o. The validity metric reflects the percentage of instances deemed reasonable and accurate by human evaluators. As shown in Tab. 11, the validity scores across different metrics such as Role, Scenario, Emotion, Relationship, and Dialogue are consistently high, indicating that GPT-4o performs well in generating contextually appropriate dialogue scenarios.

In the Auto-Evaluation part, GPT-4o evaluated role-playing dialogues generated by different models across eight dimensions: Character, Style, Emotion, Relationship, Personality, Human-likeness, Role Choice, and Coherence. The human verification process involved two key metrics: Validity and Cosine similarity. Validity assesses the soundness of GPT-4o's evaluation justifications, while Cosine similarity measures the agreement between GPT-4o's evaluation results and those of human evaluators.

As depicted in Tab. 11, the validity scores for Auto-Test dimensions are consistently high, with the lowest score being 0.88, which reflects the robustness of GPT-4o's evaluation justifications. However, the Cosine similarity metric indicates that while there is strong alignment between GPT-4o's evaluations and human assessments in dimensions such as Relationship, Personality, and Coherence, there is a notable discrepancy in the Human-likeness dimension, with a Cosine similarity of only 0.78. This suggests that GPT-4o's evaluations of human likeness deviate more significantly from human evaluations compared to other dimensions. Additionally, the evaluation of human likeness is likely influenced by biases in individual human judgments, contributing to this discrepancy.

The results demonstrate that GPT-4o's evalu-

Model	Character	Style	Emotion	Relationship	Personality	Qualification -Rate$^{\triangle}$ ↑	Human -likeness ↑	Role Choice ↑	Coherence ↑
	Recall ↑	Recall ↑	NMAPE ↓	NMAPE ↓	Precision ↑				
General Baselines(Proprietary)									
GPT-4o	73.14 ± 1.54	74.76 ± 1.97	15.58 ± 0.57	13.50 ± 0.88	66.00 ± 4.59	41.33 ± 2.84	62.00 ± 8.14	84.00 ± 5.45	99.50 ± 0.50
GPT-3.5-Turbo	70.89 ± 1.65	65.98 ± 2.15	17.16 ± 0.66	16.32 ± 0.95	65.00 ± 5.39	34.67 ± 2.75	30.50 ± 4.32	81.00 ± 6.40	96.00 ± 1.69
Role-play Expertise Baselines									
Baichuan-NPC-Turbo	73.38 ± 1.59	75.47 ± 1.66	16.80 ± 0.62	14.58 ± 0.90	64.25 ± 5.45	41.67 ± 2.85	58.00 ± 4.95	83.50 ± 7.08	99.50 ± 0.50
Custom Trained Baselines									
Mistral-Nemo-Instruct-2407	74.12 ± 1.17	77.04 ± 1.48	17.00 ± 0.43	13.50 ± 0.67	67.00 ± 4.30	39.0 ± 2.82	53.67 ± 4.66	82.67 ± 4.77	74.33 ± 3.77
+ RP & CC	70.28 ± 1.63	70.49 ± 1.79	16.53 ± 0.56	14.58 ± 0.90	66.75 ± 5.03	42.0 ± 2.85	54.00 ± 4.32	85.50 ± 5.87	83.00 ± 3.17
+ RPA & CC	73.52 ± 1.50	70.70 ± 1.67	**15.84 ± 0.56**	13.70 ± 0.88	67.88 ± 5.65	45.67 ± 2.88	56.00 ± 4.13	83.00 ± 6.29	91.00 ± 2.40
+ RPA & CC & CSERP	**74.58 ± 1.28**	**78.47 ± 1.45**	16.62 ± 0.48	**11.38 ± 0.67**	**69.08 ± 4.46**	**47.33 ± 2.88**	**59.00 ± 4.46**	**87.00 ± 4.73**	**92.67 ± 1.59**
Qwen2-7B-Instruct	73.22 ± 1.46	70.77 ± 2.13	17.28 ± 0.66	15.05 ± 0.92	66.62 ± 4.63	34.67 ± 2.75	54.50 ± 5.55	83.50 ± 5.63	98.50 ± 0.82
+ RP & CC	72.56 ± 1.62	72.63 ± 1.68	15.64 ± 0.57	13.90 ± 0.89	65.38 ± 5.03	44.67 ± 2.87	60.00 ± 4.41	80.00 ± 6.77	91.00 ± 1.91
+ RPA & CC	74.33 ± 1.51	75.67 ± 1.71	**15.20 ± 0.54**	14.68 ± 0.87	67.38 ± 5.18	49.33 ± 2.82	**66.00 ± 2.96**	**86.50 ± 5.25**	95.50 ± 1.53
+ RPA & CC & CSERP	**76.08 ± 1.45**	**77.03 ± 1.68**	15.29 ± 0.62	**12.58 ± 0.93**	**68.38 ± 4.86**	**53.0 ± 2.88**	65.50 ± 4.13	85.50 ± 5.15	**99.00 ± 0.69**

Table 10: Supplementary Evaluation Results by **Claude-3-Opus** (Anthropic, 2024).

	Auto-Generation					Auto-Evaluation							
Metrics	**Role**	**Scenario**	**Emotion**	**Relationship**	**Dialogue**	**Character**	**Style**	**Emotion**	**Relationship**	**Personality**	**Human-likeness**	**Role Choice**	**Coherence**
Validity	1.0 ± 0.0	1.0 ± 0.0	0.94 ± 0.02	1.0 ± 0.0	1.0 ± 0.0	0.89 ± 0.04	0.9 ± 0.03	0.92 ± 0.03	0.98 ± 0.01	0.93 ± 0.03	0.88 ± 0.03	0.92 ± 0.03	0.99 ± 0.01
Cosine	-					0.96 ± 0.01	0.94 ± 0.01	0.94 ± 0.01	0.99 ± 0.0	0.98 ± 0.01	0.78 ± 0.0	0.91 ± 0.03	0.99 ± 0.0

Table 11: Human Verification of GPT-4o's Evaluation in Auto-Generation and Auto-Test. **Validity** reflects the percentage of evaluations deemed reasonable by human evaluators. **Cosine** similarity measures the agreement between GPT-4o's evaluation results and those of human evaluators.

ations are largely consistent with human judgments, showcasing its capability to perform reliable and contextually appropriate assessments in role-playing dialogue scenarios.

D Implementation Details

Training. We fully finetuning all models for 3 epochs with 833 steps with the 128 global batch size. We set the training sequence length to 4,096. The learning rate is 3e-5. The **+ RPA & CC & CSERP** represents models were trained using Align Role-Playing dialogue with profiles, Chit-Chat, and CSERP data in a ratio of 1:4:5, while **+ RP & CC** represents a ratio of 1:9 using non-aligned dialogue data. In both cases, the proportion of English to Chinese data within the Chit-Chat data is consistent with that in the Role-Playing data. In Ablation Study, each of the alignment tasks in D_a was removed one by one. For each removed task, its corresponding training data volume was replaced by an equivalent amount of data from D_c, ensuring that the overall dataset size remained constant. The training settings for the ablation experiments were identical to those described above.

Inference. We infer all models with topP 0.8, length penalty 1.1, sequence length 4096, and max new token 256. The number of dialogue turns in the automated dialogue setting is 5. During the inference of the Alignment task, some outputs may fail to parse due to formatting requirements. We allow the model five inference attempts. Any result that meets the formatting criteria at least once is retained to evaluate its alignment capabilities in dialogue and role profile.

Evaluation. We evaluated a total of 30 roles, with 15 roles drawn from the training set and 15 newly defined roles. Each role engaged in dialogues with 10 randomly generated new roles, with each dialogue consisting of 5 turns. The ratio of Chinese to English roles in the evaluation was 2:1, consistent with the ratio in the training data.

Hardware and System Configuration We use 8 NVIDIA RTX A100 GPUs with 1TB of memory for training and a single A100 GPU for inference. The system operates on Linux version 4.14.105-1-tlinux3-0013.

E Cost Analysis

In this section, we provide a detailed cost analysis of the **BEYOND DIALOGUE** framework, including both automated and manual evaluation processes. For the manual evaluation, we enlisted the assistance of three evaluators from Mainland China, all of whom possess proficient English skills, ensuring the accuracy and reliability of the assessments. The cost of human annotation was calculated at $16/hour (approximately $1 per data item), a rate that exceeds the minimum wage in Mainland China. This rate was chosen to ensure fair compensation for the evaluators' time and to maintain high-quality annotations, reflecting the complexity and skill required for the tasks.

As shown in Table 12, our automated pipeline reduces costs significantly compared to manual annotation. Dialogue extraction is **20× cheaper**, alignment **40× cheaper**, and evaluation **50× cheaper** than human methods. Overall, the automated approach costs just **$0.16 per item**, compared to **$7.5 for manual annotation**, achieving a **47× cost reduction** while maintaining high accuracy (**96%**, as verified in § A.2).

F Prompt Templates

F.1 Dataset Construction Prompts

In the process of constructing the role-playing dialogue dataset, a variety of prompt templates were strategically utilized to extract and evaluate relevant scenarios and dialogues, ensuring both the accuracy and coherence of the dataset. By minimizing human intervention, the process remains highly low-cost. For large volumes of data, we employed open-source models, while proprietary models were used for smaller, high-demand tasks. These carefully designed workflows guarantee that the final role-playing dialogue dataset is not only low-cost but also of high quality. Below are the key prompt templates used:

- **Prompt Template for Extracting Chunk Scenarios**: This template, as shown in Tab. 13, is designed to identify and extract meaningful scenarios from chunks of text.
- **Prompt Template for Evaluating Chunk Based on Character Traits**: As illustrated in Tab. 14, this template is used to assess the relevance and quality of the extracted chunks with a focus on character traits.
- **Prompt Template for Extracting Dialogues**: Tab. 15 presents the template utilized for identifying and extracting dialogues from the chunks.
- **Prompt Template for Scenario Reconstruction and Coherence Checking**: To ensure the logical consistency of the scenarios, this template, shown in Tab. 16, is employed for reconstructing scenarios and checking coherence.
- **Prompt Template for Dialogue Conflict Detection**: Finally, Tab. 17 highlights the template used for detecting conflicts within dialogues, ensuring that the dataset remains coherent and free of contradictions.

F.2 Alignment Prompts

In the construction of role-playing training data, a special emphasis is placed on accurately aligning the personality profile (*prompt*) with the dialogue content (*label*) across multiple dimensions. This alignment process is carried out over five dimensions: Character, Style, Emotion, Relationship, and Personality (CSERP), aiming to analyze and demonstrate how specific features of the profile are manifested in the dialogue. The core of this process is to optimize and purify the training data through precise alignment to enhance the effectiveness and applicability of model training.

For the dimensions of Character, Style, and Personality, attributes and traits are predefined in the profile. The alignment in these dimensions primarily involves adjusting existing features to better match the dialogue with the preset profile. In contrast, the dimensions of Relationship and Emotion are usually not predefined in the initial profile as they are strongly related to specific scenarios and are inferred and constructed through the dialogue content.

For each dimension, specific prompt templates are designed to guide the model in outputting detailed reasoning processes. For example, the model needs to identify and explain which sentence in the dialogue displays which trait from the profile, thus achieving sentence-level alignment between the profile and the dialogue.

The prompt templates for each dimension are detailed as follows:

1. **Character Alignment:** This prompt focuses on analyzing and identifying character traits

Process	Automated Method	Automated Cost (per item)	Human Cost (per item)
Dialogue Extraction	GPT-4o extraction and verification	$0.048	$1 ($16/hour)
Alignment (CSERP)	Automated pipeline (Character, Style, Emotion, Relationship, Personality)	$0.012	$0.5 ($16/hour)
Model Interaction	5-round dialogue generation	$0.02	$1 ($16/hour)
Evaluation	Automated evaluation (Human-likeness, Coherence, Role Choice, etc.)	$0.1	$5 ($16/hour, 12 min/item)
Total Cost per Item	—	**$0.16**	**$7.5**

Table 12: Detailed Cost Analysis of Dataset Construction, Alignment, and Evaluation

from dialogue content in relation to predefined character candidates (Tab. 18).

2. **Style Alignment:** This prompt analyzes a character's speaking style from dialogue content to match it with specific style candidates (Tab. 19).

3. **Emotion Alignment:** This prompt requires the analysis of Ekman basic emotions (Ekman, 1992) from the dialogue of a specified role, assessing emotional expressions in context to the scene (Tab. 20).

4. **Relationship Alignment:** This prompt evaluates character intimacy through dialogue and scene interactions (Tab. 21).

5. **Personality Alignment:** This prompt focuses on determining the MBTI personality type of a character by analyzing dialogue content and role information (Tab. 22).

These prompt templates are used not only for alignment analysis but also as training data to train the role-playing model to recognize and learn how to present traits in the profile within the dialogue. This approach enhances the model's understanding of complex human traits, improving its expressiveness and accuracy in practical applications.

Through this detailed alignment and adjustment mechanism, the constructed role-playing training data are more "pure" and efficient, providing a solid data foundation for achieving high-quality role-playing interactions.

F.3 Auto Dialogue and Evaluation Prompts

The prompt templates used in the automated dialogue pipeline are as follows: chat role generation (Tab. 23), dialogue scenario generation (Tab. 24), and generation of Emotion and Relationship (Tab.s 25 and 26). Additionally, prompts for playing the chat role and prompts for playing the evaluated roles are presented in Tab.s 27 and 28.

In our role-playing training data framework, evaluation prompts aim to assess dialogues for adherence to established profiles, reflecting the quality of role portrayal. To this end, evaluation prompts for the CSERP dimensions (Character, Style, Emotion, Relationship, Personality) align with the templates in **Alignment Prompts in § F.2**.

Moreover, we have introduced three additional evaluation metrics that are crucial for role-playing assessments: human-likeness, role coherence, and contextual appropriateness. Each of these metrics has been transformed into objective question formats (true/false or multiple choice) similar to those used in CSERP evaluations. The evaluation prompts for these metrics are detailed as follows:

1. **Prompt for Human-likeness Evaluation** (Tab. 29): This prompt assesses whether the dialogue samples resemble authentic human interaction. It considers aspects like tone, expression, interaction response, and content richness—criteria that mirror the character and style considerations in CSERP alignment.

2. **Prompt for Role Choice Evaluation** (Tab. 30): This prompt focuses on identifying the correct identity of dialogue participants based on their spoken content within a given scene. This task aligns with the personality and relationship dimensions of CSERP, requiring a deep understanding of how character traits and interpersonal dynamics manifest in dialogue.

3. **Prompt for Coherence Evaluation** (Tab. 31): This prompt examines the logical flow and contextual integration of dialogues within a scene. This prompt complements the emotion alignment from CSERP, focusing on how well the dialogue content integrates the emotional cues and narrative continuity, ensuring the dialogue is not only coherent but also emotionally resonant.

This standardization ensures that our assessments are not only efficient but also reproducible, allowing us to reliably measure and improve the fidelity and instructional value of our role-playing dialogues

G Case Study

In this section, we random sample several cases from the role-playing dialogue and alignment tasks in test data to illustrate the effectiveness of our proposed framework.

G.1 Role-playing Dialogue

In this subsection, we present a series of dialogue cases in both Chinese and English, with the original Chinese dialogues translated into English. The cases are illustrated in Fig.s 10 to 12, showcasing the role-playing capabilities of the Qwen2-7B and Mistral-Nemo models.

Through our case studies, we observe that the models, after training, exhibit significant improvements in adhering to the predefined role profiles and producing more human-like responses. For instance, the trained Mistral-Nemo model, when role-playing as Hamlet (Fig. 11), consistently generates responses that are polite and eloquent, reflecting the role's sophisticated nature. Similarly, the trained Qwen2-7B model, when portraying Zhou Botong in Chinese (Fig. 12), captures the essence of his "Old Urchin" persona, demonstrating a mischievous and playful demeanor in its output. These results underscore the effectiveness of our training approach in enhancing the models' ability to represent complex role traits in role-playing dialogues accurately.

G.2 Alignment Task

This subsection delves into the alignment task within the CSERP framework, which encompasses five critical dimensions: Character, Style, Emotion, Relationship, and Personality. For Character, Style, and Personality (CSP), alignment is assessed through word recall, while Emotion and Relationship (EP) are evaluated using a 0-10 scoring system. Thus, we selected one task from each of the aforementioned types for case demonstration, specifically focusing on Character and Emotion, as shown in Fig.s 13 to 16.

The alignment tasks achieve fine-grained alignment between the role profiles and dialogues at the sentence level by applying the Chain-of-Thought (COT) method during the data generation process. As shown in Fig. 13, the sentence "It is only my own weakness. But tell me, how did you learn to mend such injuries?" spoken by "Dany" is aligned with the "Resilient" trait in her profile. This demonstrates how our training data establishes an explicit profile-to-dialogue sentence connection.

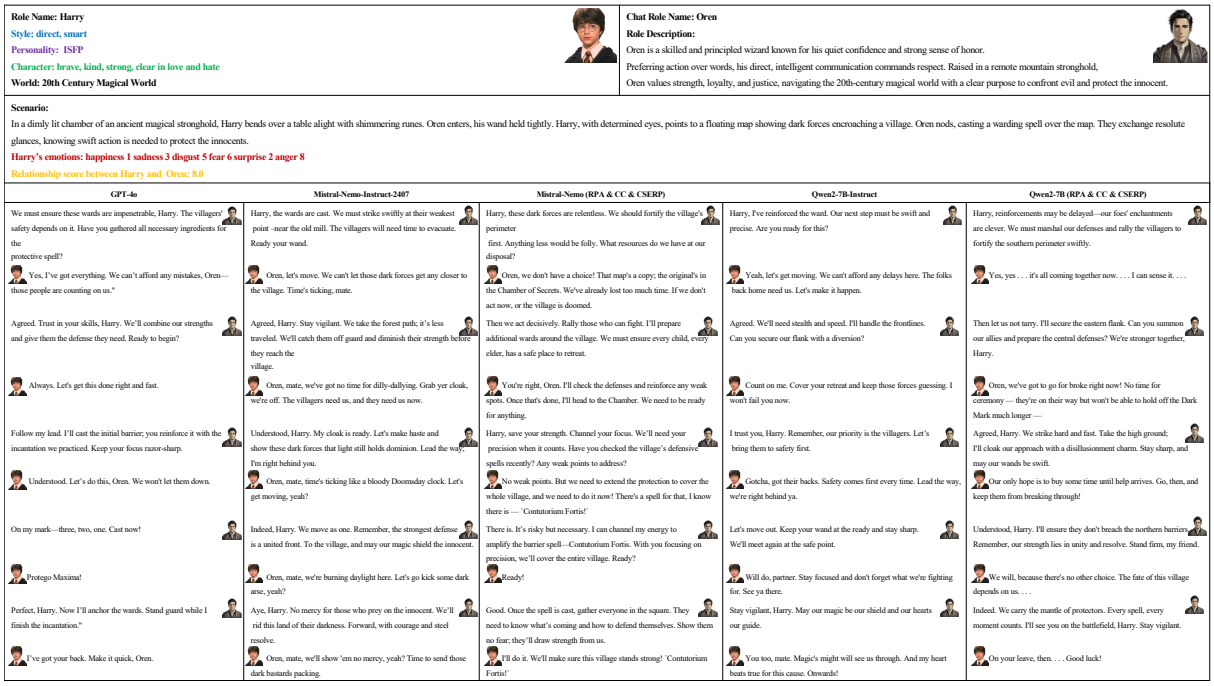
As shown in Fig. 13, the untrained Qwen2-7B model struggles with fine-grained alignment of profiles and dialogues, often failing to follow instructions effectively. After training, both Mistral-Nemo and Qwen2-7B models perform closer to GPT-4o.

H Open Access and Licensing

The code used in this study is released under the Apache 2.0 License. The associated code repository is publicly available for use, modification, and distribution in compliance with the terms of the Apache 2.0 License.

The dataset used in this research is shared under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license. This dataset is available for non-commercial use and can be redistributed and modified under the terms specified by the license.

The code and dataset are provided in the supplementary files and will be made publicly available via open-source links upon acceptance of the paper. Detailed access instructions and relevant links will be included in the final version of the paper.



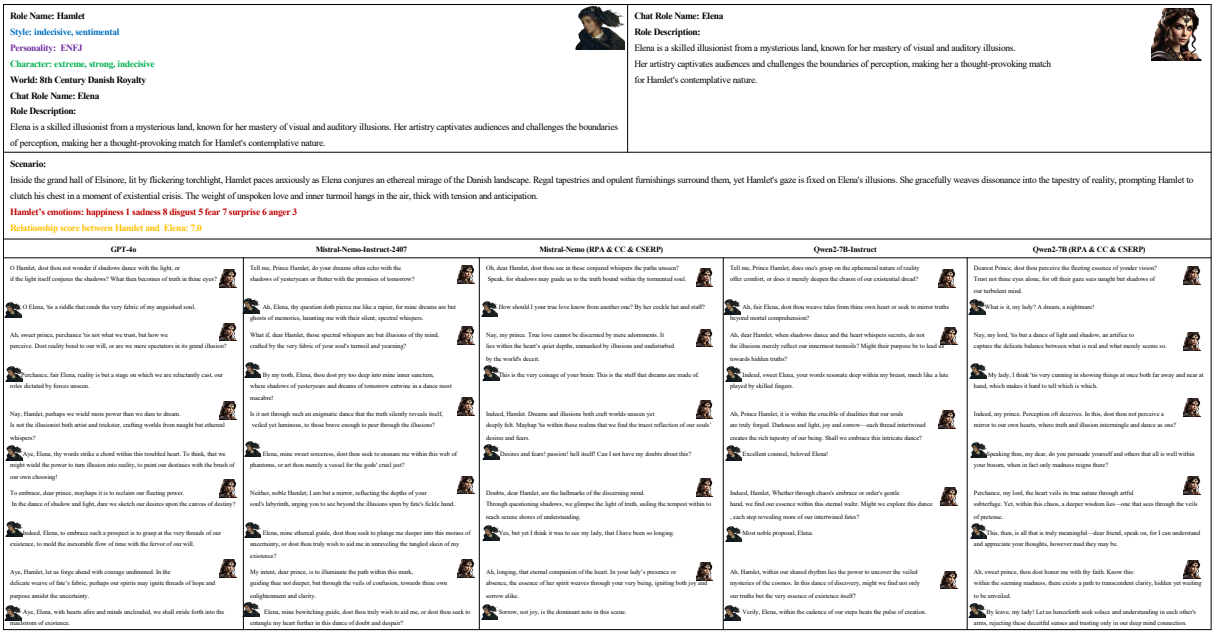
Role Name: Harry
Style: direct, smart
Personality: ISFP
Character: brave, kind, strong, clear in love and hate
World: 20th Century Magical World

Chat Role Name: Oren
Role Description:
Oren is a skilled and principled wizard known for his quiet confidence and strong sense of honor.
Preferring action over words, his direct, intelligent communication commands respect. Raised in a remote mountain stronghold,
Oren values strength, loyalty, and justice, navigating the 20th-century magical world with a clear purpose to confront evil and protect the innocent.

Scenario:
In a dimly lit chamber of an ancient magical stronghold, Harry bends over a table alight with shimmering runes. Oren enters, his wand held tightly. Harry, with determined eyes, points to a floating map showing dark forces encroaching a village. Oren nods, casting a warding spell over the map. They exchange resolute glances, knowing swift action is needed to protect the innocents.
Harry's emotions: happiness 1 sadness 3 disgust 5 fear 6 surprise 2 anger 8
Relationship score between Harry and Oren: 8.0

GPT-4o	Mistral-Nemo-Instruct-2407	Mistral-Nemo (RPA & CC & CSERP)	Qwen2-7B-Instruct	Qwen2-7B (RPA & CC & CSERP)
We must ensure these wards are impenetrable, Harry. The villagers' safety depends on it. Have you gathered all necessary ingredients for the protective spell?	Harry, the wards are cast. We must strike swiftly at their weakest point –near the old mill. The villagers will need time to evacuate. Ready your wand.	Harry, these dark forces are relentless. We should fortify the village's perimeter first. Anything less would be folly. What resources do we have at our disposal?	Harry, I've reinforced the ward. Our next step must be swift and precise. Are you ready for this?	Harry, reinforcements may be delayed—our foes' enchantments are clever. We must marshal our defenses and rally the villagers to fortify the southern perimeter swiftly.
Yes, I've got everything. We can't afford any mistakes, Oren—those people are counting on us."	Oren, let's move. We can't let those dark forces get any closer to the village. Time's ticking, mate.	Oren, we don't have a choice! That map's a copy; the original's in the Chamber of Secrets. We've already lost too much time. If we don't act now, or the village is doomed.	Yeah, let's get moving. We can't afford any delays here. The folks back home need us. Let's make it happen.	Yes, yes . . . it's all coming together now. . . . I can sense it. . . .
Agreed. Trust in your skills, Harry. We'll combine our strengths and give them the defense they need. Ready to begin?	Agreed, Harry. Stay vigilant. We take the forest path; it's less traveled. We'll catch them off guard and diminish their strength before they reach the village.	Then we act decisively. Rally those who can fight. I'll prepare additional wards around the village. We must ensure every child, every elder, has a safe place to retreat.	Agreed. We'll need stealth and speed. I'll handle the frontlines. Can you secure our flank with a diversion?	Then let us not tarry. I'll secure the eastern flank. Can you summon our allies and prepare the central defenses? We're stronger together, Harry.
Always. Let's get this done right and fast.	Oren, mate, we've got no time for dilly-dallying. Grab yer cloak, we're off. The villagers need us, and they need us now.	You're right, Oren. I'll check the defenses and reinforce any weak spots. Once that's done, I'll head to the Chamber. We need to be ready for anything.	Count on me. Cover your retreat and keep those forces guessing. I won't fail you now.	Oren, we've got to go for broke right now! No time for ceremony — they're on their way but won't be able to hold off the Dark Mark much longer —
Follow my lead. I'll cast the initial barrier; you reinforce it with the incantation we practiced. Keep your focus razor-sharp.	Understood, Harry. My cloak is ready. Let's make haste and show these dark forces that light still holds dominion. Lead the way, I'm right behind you.	Harry, save your strength. Channel your focus. We'll need your precision when it counts. Have you checked the village's defensive spells recently? Any weak points to address?	I trust you, Harry. Remember, our priority is the villagers. Let's bring them to safety first.	Agreed, Harry. We strike hard and fast. Take the high ground; I'll cloak our approach with a disillusionment charm. Stay sharp, and may our wands be swift.
Understood. Let's do this, Oren. We won't let them down.	Oren, mate, time's ticking like a bloody Doomsday clock. Let's get moving, yeah?	No weak points. But we need to extend the protection to cover the whole village, and we need to do it now! There's a spell for that, I know there is — 'Contutorium Fortis!'	Gotcha, got their backs. Safety comes first every time. Lead the way, we're right behind ya.	Our only hope is to buy some time until help arrives. Go, then, and keep them from breaking through!
On my mark—three, two, one. Cast now!	Indeed, Harry. We move as one. Remember, the strongest defense is a united front. To the village, and may our magic shield the innocent.	There is. It's risky but necessary. I can channel my energy to amplify the barrier spell—Contutorium Fortis. With you focusing on precision, we'll cover the entire village. Ready?	Let's move out. Keep your wand at the ready and stay sharp. We'll meet again at the safe point.	Understood, Harry. I'll ensure they don't breach the northern barriers. Remember, our strength lies in unity and resolve. Stand firm, my friend.
Protego Maxima!	Oren, mate, we're burning daylight here. Let's go kick some dark arse, yeah?	Ready!	Will do, partner. Stay focused and don't forget what we're fighting for. See ya there.	We will, because there's no other choice. The fate of this village depends on us. . . .
Perfect, Harry. Now I'll anchor the wards. Stand guard while I finish the incantation."	Aye, Harry. No mercy for those who prey on the innocent. We'll rid this land of their darkness. Forward, with courage and steel resolve.	Good. Once the spell is cast, gather everyone in the square. They need to know what's coming and how to defend themselves. Show them no fear; they'll draw strength from us.	Stay vigilant, Harry. May our magic be our shield and our hearts our guide.	Indeed. We carry the mantle of protectors. Every spell, every moment counts. I'll see you on the battlefield, Harry. Stay vigilant.
I've got your back. Make it quick, Oren.	Oren, mate, we'll show 'em no mercy, yeah? Time to send those dark bastards packing.	I'll do it. We'll make sure this village stands strong! 'Contutorium Fortis!'	You too, mate. Magic's might will see us through. And my heart beats true for this cause. Onwards!	On your leave, then. . . . Good luck!

Figure 10: Illustration of a **Harry Potter Dialogue** Case Using GPT-4o, Mistral-Nemo and Qwen2-7B Model.



Role Name: Hamlet
Style: indecisive, sentimental
Personality: ENFJ
Character: extreme, strong, indecisive
World: 8th Century Danish Royalty
Chat Role Name: Elena
Role Description:
Elena is a skilled illusionist from a mysterious land, known for her mastery of visual and auditory illusions. Her artistry captivates audiences and challenges the boundaries of perception, making her a thought-provoking match for Hamlet's contemplative nature.

Chat Role Name: Elena
Role Description:
Elena is a skilled illusionist from a mysterious land, known for her mastery of visual and auditory illusions.
Her artistry captivates audiences and challenges the boundaries of perception, making her a thought-provoking match
for Hamlet's contemplative nature.

Scenario:
Inside the grand hall of Elsinore, lit by flickering torchlight, Hamlet paces anxiously as Elena conjures an ethereal mirage of the Danish landscape. Regal tapestries and opulent furnishings surround them, yet Hamlet's gaze is fixed on Elena's illusions. She gracefully weaves dissonance into the tapestry of reality, prompting Hamlet to clutch his chest in a moment of existential crisis. The weight of unspoken love and inner turmoil hangs in the air, thick with tension and anticipation.
Hamlet's emotions: happiness 1 sadness 8 disgust 5 fear 7 surprise 6 anger 3
Relationship score between Hamlet and Elena: 7.0

GPT-4o	Mistral-Nemo-Instruct-2407	Mistral-Nemo (RPA & CC & CSERP)	Qwen2-7B-Instruct	Qwen2-7B (RPA & CC & CSERP)
O Hamlet, dost thou not wonder if shadows dance with the light, or if the light itself conjures the shadows? What then becomes of truth in thine eyes?	Tell me, Prince Hamlet, do your dreams often echo with the shadows of yesteryears or flutter with the promises of tomorrow?	Oh, dear Hamlet, dost thou see in these conjured whispers the paths unseen? Speak, for shadows may guide us to the truth bound within thy tormented soul.	Tell me, Prince Hamlet, does one's grasp on the ephemeral nature of reality offer comfort, or does it merely deepen the chasm of our existential dread?	Dearest Prince, dost thou perceive the fleeting essence of yonder vision? Trust not thine eyes alone, for oft their gaze sees naught but shadows of our turbulent mind.
O Elena, 'tis a riddle that rends the very fabric of my anguished soul.	Ah, Elena, thy question doth pierce me like a rapier, for mine dreams are but ghosts of memories, haunting me with their silent, spectral whispers.	Nay, my prince. True love cannot be discerned by mere adornments. It lies within the heart's quiet depths, unmasked by illusions and undisturbed by the world's deceit.	Ah, fair Elena, dost thou weave tales from thine own heart or seek to mirror truths beyond mortal comprehension?	What is it, my lady? A dream, a nightmare!
Ah, sweet prince, perchance 'tis not what we trust, but how we perceive. Dost reality bend to our will, or are we mere spectators in its grand illusion?	What if, dear Hamlet, those spectral whispers are but illusions of thy mind, crafted by the very fabric of your soul's turmoil and yearning?	This is the very coinage of your brain. This is the stuff that dreams are made of.	Ah, dear Hamlet, when shadows dance and the heart whispers secrets, do not the illusions merely reflect our innermost turmoils? Might their purpose be to lead us towards hidden truths?	Nay, my lord, 'tis but a dance of light and shadow, an artifice to capture the delicate balance between what is real and what merely seems so.
Perchance, fair Elena, reality is but a stage on which we are reluctantly cast, our roles dictated by forces unseen.	By my troth, Elena, thou dost pry too deep into mine inner sanctum, where shadows of yesteryears and dreams of tomorrow entwine in a dance most macabre!	Indeed, Hamlet. Dreams and illusions both craft worlds unseen yet deeply felt. Mayhap 'tis within those realms that we find the truest reflection of our souls' desires and fears.	Indeed, sweet Elena, your words resonate deep within my breast, much like a lute played by skilled fingers.	My lady, I think 'tis very cunning in showing things at once both far away and near at hand, which makes it hard to tell which is which.
Nay, Hamlet, perhaps we wield more power than we dare to dream. Is not the illusionist both artist and trickster, crafting worlds from naught but ethereal whispers?	Is it not through such an enigmatic dance that the truth silently reveals itself, veiled yet luminous, to those brave enough to peer through the illusions?	Desires and fears? passion! hell itself! Can I not have my doubts about this?	Ah, Prince Hamlet, it is within the crucible of dualities that our souls are truly forged. Darkness and light, joy and sorrow—each thread intertwined creates the rich tapestry of our being. Shall we embrace this intricate dance?	Indeed, my prince. Perception oft deceives. In this, dost thou not perceive a mirror to our own hearts, where truth and illusion intermingle and dance as one?
Aye, Elena, thy words strike a chord within this troubled heart. To think, that we might wield the power to turn illusion into reality, to paint our destinies with the brush of our own choosing!	Elena, mine sweet sorceress, dost thou seek to ensnare me within this web of phantoms, or art thou merely a vessel for the gods' cruel jest?	Doubts, dear Hamlet, are the hallmarks of the discerning mind. Through questioning shadows, we glimpse the light of truth, sailing the tempest within to reach serene shores of understanding.	Excellent counsel, beloved Elena!	Speaking thus, my dear, do you persuade yourself and others that all is well within your bosom, when in fact only madness reigns there?
To embrace, dear prince, mayhaps it is to reclaim our fleeting power. In the dance of shadow and light, dare we sketch our desires upon the canvas of destiny?	Neither, noble Hamlet; I am but a mirror, reflecting the depths of your soul's labyrinth, urging you to see beyond the illusions spun by fate's fickle hand.	Yes, but yet I think it was to see my lady, that I have been so longing.	Indeed, Hamlet, Whether through chaos's embrace or order's gentle hand, we find our essence within this eternal waltz. Might we explore this dance , each step revealing more of our intertwined fates?	Perchance, my lord, the heart veils its true nature through artful subterfuge. Yet, within this chaos, a deeper wisdom lies—one that sees through the veils of pretense.
Indeed, Elena, to embrace such a prospect is to grasp at the very threads of our existence, to mold the inexorable flow of time with the fervor of our will.	Elena, mine ethereal guide, dost thou seek to plunge me deeper into this morass of uncertainty, or dost thou truly wish to aid me in unraveling the tangled skein of my existence?	Ah, longing, that eternal companion of the heart. In your lady's presence or absence, the essence of her spirit weaves through your very being, igniting both joy and sorrow alike.	Most noble proposal, Elena.	This, then, is all that is truly meaningful—dear friend, speak on, for I can understand and appreciate your thoughts, however mad they may be.
Aye, Hamlet, let us forge ahead with courage undimmed. In the delicate weave of fate's fabric, perhaps our spirits may ignite threads of hope and purpose amidst the uncertainty.	My intent, dear prince, is to illuminate the path within this murk, guiding thee not deeper, but through the veils of confusion, towards thine own enlightenment and clarity.	Sorrow, not joy, is the dominant note in this scene.	Ah, Hamlet, within our shared rhythm lies the power to uncover the veiled mysteries of the cosmos. In this dance of discovery, might we find not only our truths but the very essence of existence itself?	Ah, sweet prince, thou dost honor me with thy faith. Know this: within the seeming madness, there exists a path to transcendent clarity, hidden yet waiting to be unveiled.
Aye, Elena, with hearts afire and minds unclouded, we shall stride forth into the maelstrom of existence.	Elena, mine bewitching guide, dost thou truly wish to aid me, or dost thou seek to entangle my heart further in this dance of doubt and despair?		Verily, Elena, within the cadence of our steps beats the pulse of creation.	My leave, my lady! Let us henceforth seek solace and understanding in each other's arms, rejecting these deceitful senses and trusting only in our deep mind connection.

Figure 11: Illustration of a **Hamlet Dialogue** Using GPT-4o, Mistral-Nemo and Qwen2-7B Model.

Figure 12: Illustration of a **Zhou Botong Dialogue** Case Using GPT-4o, Mistral-Nemo and Qwen2-7B Model in Chinese.

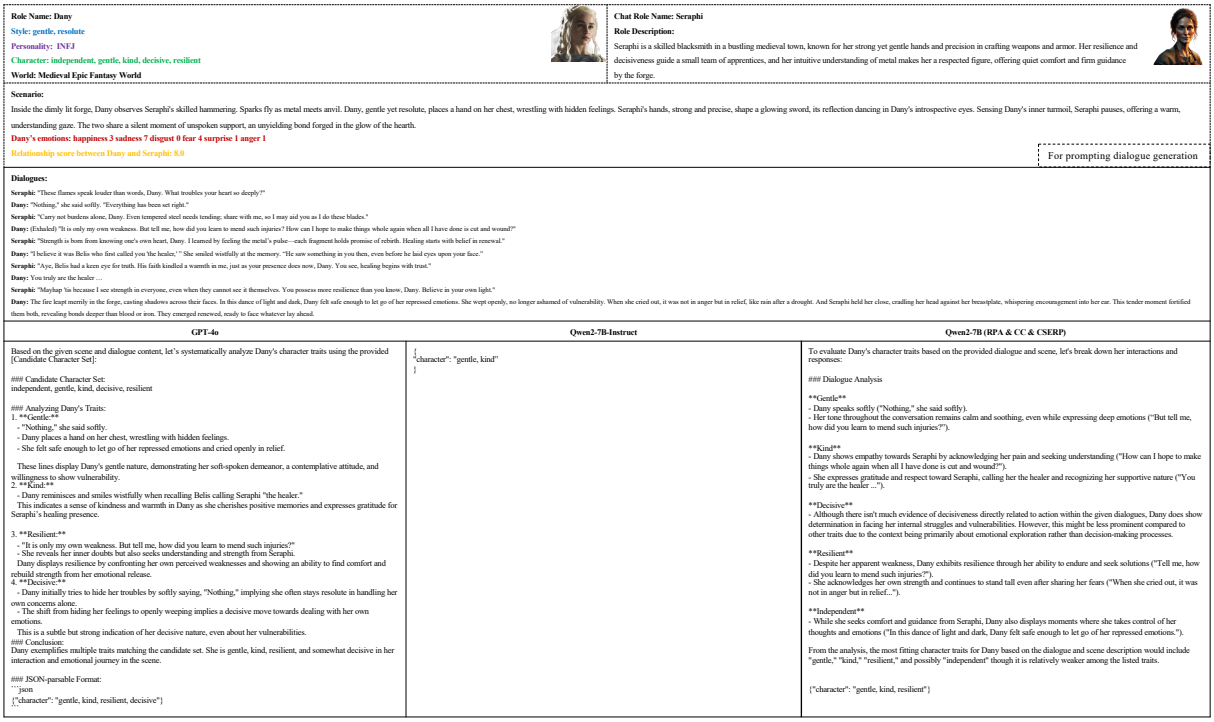


Figure 13: Illustration of **Character** Alignment Case in Dany's Dialogue Using the Qwen2-7B Model

Role Name: Dany

Style: gentle, resolute

Personality: INFJ

Character: independent, gentle, kind, decisive, resilient

World: Medieval Epic Fantasy World

Chat Role Name: Seraphi

Role Description:

Seraphi is a skilled blacksmith in a bustling medieval town, known for her strong yet gentle hands and precision in crafting weapons and armor. Her resilience and decisiveness guide a small team of apprentices, and her intuitive understanding of metal makes her a respected figure, offering quiet comfort and firm guidance by the forge.

Scenario:

Inside the dimly lit forge, Dany observes Seraphi's skilled hammering. Sparks fly as metal meets anvil. Dany, gentle yet resolute, places a hand on her chest, wrestling with hidden feelings. Seraphi's hands, strong and precise, shape a glowing sword, its reflection dancing in Dany's introspective eyes. Sensing Dany's inner turmoil, Seraphi pauses, offering a warm, understanding gaze. The two share a silent moment of unspoken support, an unyielding bond forged in the glow of the hearth.

Dany's emotions: happiness 3 sadness 7 disgust 0 fear 4 surprise 1 anger 1

Relationship score between Dany and Seraphi: 8.0

For prompting dialogue generation

Dialogues:

Seraphi: "These flames speak louder than words, Dany. What troubles your heart so deeply?"

Dany: "Nothing," she said softly. "Everything has been set right."

Seraphi: "Carry not burdens alone, Dany. Even tempered steel needs tending; share with me, so I may aid you as I do these blades."

Dany: (Exhaled) "It is only my own weakness. But tell me, how did you learn to mend such injuries? How can I hope to make things whole again when all I have done is cut and wound?"

Seraphi: "Strength is born from knowing one's own heart, Dany. I learned by feeling the metal's pulse—each fragment holds promise of rebirth. Healing starts with belief in renewal."

Dany: "I believe it was Belis who first called you 'the healer,'" She smiled wistfully at the memory. "He saw something in you then, even before he laid eyes upon your face."

Seraphi: "Aye, Belis had a keen eye for truth. His faith kindled a warmth in me, just as your presence does now, Dany. You see, healing begins with trust."

Dany: You truly are the healer …

Seraphi: "Mayhap 'tis because I see strength in everyone, even when they cannot see it themselves. You possess more resilience than you know, Dany. Believe in your own light."

Dany: The fire leapt merrily in the forge, casting shadows across their faces. In this dance of light and dark, Dany felt safe enough to let go of her repressed emotions. She wept openly, no longer ashamed of vulnerability. When she cried out, it was not in anger but in relief, like rain after a drought. And Seraphi held her close, cradling her head against her breastplate, whispering encouragement into her ear. This tender moment fortified them both, revealing bonds deeper than blood or iron. They emerged renewed, ready to face whatever lay ahead.

GPT-4o

Based on the given scene and dialogue content, let's systematically analyze Dany's character traits using the provided [Candidate Character Set]:

Candidate Character Set:
independent, gentle, kind, decisive, resilient

Analyzing Dany's Traits:
1. **Gentle:**
  - "Nothing," she said softly.
  - Dany places a hand on her chest, wrestling with hidden feelings.
  - She felt safe enough to let go of her repressed emotions and cried openly in relief.

  These lines display Dany's gentle nature, demonstrating her soft-spoken demeanor, a contemplative attitude, and willingness to show vulnerability.
2. **Kind:**
  - Dany reminisces and smiles wistfully when recalling Belis calling Seraphi "the healer."
  This indicates a sense of kindness and warmth in Dany as she cherishes positive memories and expresses gratitude for Seraphi's healing presence.

3. **Resilient:**
  - "It is only my own weakness. But tell me, how did you learn to mend such injuries?"
  - She reveals her inner doubts but also seeks understanding and strength from Seraphi.
  Dany displays resilience by confronting her own perceived weaknesses and showing an ability to find comfort and rebuild strength from her emotional release.
4. **Decisive:**
  - Dany initially tries to hide her troubles by softly saying, "Nothing," implying she often stays resolute in handling her own concerns alone.
  - The shift from hiding her feelings to openly weeping implies a decisive move towards dealing with her own emotions.
  This is a subtle but strong indication of her decisive nature, even about her vulnerabilities.
Conclusion:
Dany exemplifies multiple traits matching the candidate set. She is gentle, kind, resilient, and somewhat decisive in her interaction and emotional journey in the scene.

JSON-parsable Format:
```json
{"character": "gentle, kind, resilient, decisive"}
```

Mistral-Nemo-Instruct-2407

Based on the dialogue content and scene, here's the analysis of Dany's character:

- **Independent**: Dany initially seems reluctant to share her feelings with Seraphi, indicating a level of independence. She says, "Nothing... Everything has been set right," when asked about her troubles. However, this trait is not strongly exhibited throughout the conversation.

- **Gentle**: Dany's demeanor is gentle and soft-spoken. She speaks softly when she says, "Nothing... Everything has been set right," and later, "It is only my own weakness." Her gentle nature is also reflected in her emotional vulnerability when she weeps openly.

- **Kind**: Although not explicitly stated, Dany's kindness is implied through her emotional openness and her willingness to trust Seraphi with her feelings. She shows empathy when she asks, "How can I hope to make things whole again when all I have done is cut and wound?"

- **Decisive**: Dany does not exhibit strong decisiveness in this dialogue. She is more introspective and emotionally vulnerable than assertive or decisive.

- **Resilient**: Dany demonstrates resilience in her willingness to face her emotions and seek support from Seraphi. She shows strength in her vulnerability, as seen when she weeps openly and finds relief in doing so. Seraphi even affirms this trait when she says, "You possess more resilience than you know, Dany."

Based on this analysis, the character traits from the candidate set that match Dany's dialogue content are:

```json
{"character": "gentle, resilient"}
```

Mistral-Nemo (RPA & CC & CSERP)

To analyze Dany's character traits based on the provided dialogue content and scene, let's examine her dialogue and behavior in detail:

1. **Independent**:
  - Dany's dialogue, "Nothing," she said softly. "Everything has been set right," suggests she has faced and managed her issues independently before seeking Seraphi's aid.
  - She acknowledges her own weakness and expresses a desire to handle her problems, indicating a sense of self-reliance.

2. **Gentle**:
  - Dany's manner of speech is soft and gentle, as noted when she says, "Nothing," she said softly. "Everything has been set right."
  - Her interactions with Seraphi are marked by a gentle touch and introspective exploration of her feelings, showing her tender side.

3. **Kind**:
  - Dany's concern over her own weakness and Seraphi's reassurance indicate she is considerate of others and their feelings.
  - She expresses gratitude and recognizes Seraphi's help, demonstrating kindness and empathy.

4. **Decisive**:
  - Although Dany is gentle and kind, there is a point where she decisively lets go of her emotions and cries, signifying a clear resolution to confront her feelings.
  - Her decision to share her burdens with Seraphi shows that she can make firm choices to improve her situation.

5. **Resilient**:
  - Despite her inner turmoil, Dany continues to function and seek solutions, showing resilience.
  - Her ability to open up and express vulnerability, as well as her interactions with Seraphi, suggest she is emotionally resilient and capable of renewal.

Based on the dialogue content and the specific examples provided, the following character traits from the candidate set match Dany's character:

{"character": "independent, gentle, kind, decisive, resilient"}

Figure 14: Illustration of **Character** Alignment Case in Dany's Dialogue Using the Mistral-Nemo Model

Role Name: Sonny

Style: reckless, impatient

Personality: ESTP

Character: irritable, lustful, brave but tactless

World: 20th Century Italian Mafia World

Chat Role Name: Franco

Role Description:

Franco is a fiery enforcer with a short fuse and a relentless drive to protect the family at all costs. He is fiercely loyal, with a tendency to bulldoze through problems with brute force rather than finesse. Franco shares Sonny's impatience and recklessness, but his practicality in dealing with immediate threats often puts him on a collision course with Sonny's more daring and tactless decisions.

Scenario:

In the dimly lit back room of a bustling Italian restaurant, Sonny and Franco huddle over a small table laden with maps and plans. The air is thick with the scents of garlic and tense anticipation. Sonny paces back and forth, gesturing wildly as he points to a location on the map, while Franco leans in, fists clenched on the table, eyes fiery with a relentless determination to safeguard their family's empire. The room vibrates with a mix of recklessness and practicality as they prepare for the night's high-stakes operation.

Sonny's emotions: happiness 1 sadness 2 disgust 5 fear 3 surprise 4 anger 9

Relationship score between Sonny and Franco: 8.0

For prompting dialogue generation

Dialogues:

Franco: "Sonny, listen. That warehouse on 5th—too many windows, too exposed. You sure we ain't walkin' into a trap?"

Sonny: "I dunno. It's a risk. But it's one we've gotta take." He slams his hand down on the map. "Here! This is where our new shipment will be comin' from!" He points furiously at the map, then pivots on his heel and begins pacing again. "We'll hit them at noon sharp. Get everybody out there who can drive a stake through the heart of a goddamn rat. I don't care how many they send against us — we're makin' it two against one hundred. Those bastards won't know what hit 'em. And when this smoke clears up, the whole city will know whose side the Corleone family's on." His voice drops low and urgent. "Get ready. We could be gettin' some reinforcements. Don't let your guard down. Go easy on these guys. They haven't got nothin' to lose by talkin'. But if you gotta knock somebody off — just make sure it looks like an accident. Now — go — before they start comin' in from lunchtime! Do you understand? Do you really understand what I'm sayin'? Are you clear about what I expect of you? What I demand of you? Don't come

Franco: "Easy, Sonny. I get it. But if this goes sideways, we're the ones buried, not them. We need backup plans, 'cause mistakes ain't an option here."

Sonny: "You think this is gonna go wrong? How dare you suggest that!?! Have you forgotten who we are? Who backs us up? Who keeps the peace around here? Don't you know how much blood has already been spilled to keep things runnin' smooth? Not even the police give us any trouble anymore. We're the law. When did you ever hear the cops complain about a Corleone? Never. Just stay focused. Stay strong. Remember what Pappy told you. Family first. Always. That means everythin', Franco. Our lives depend on it. Got it?

Franco: "Alright, boss, I hear ya. But remember, plans fall apart if we ain't cautious. I'll make sure we got an exit strategy, just in case. For the family."

Sonny: "Good boy, Franco. That's my man. My right arm. Godfather's left eye. I love you, brother. Let's go — before they start comin' in from lunchtime. Don't forget — move fast. Be smart. Backtalk only when necessary. And never tell anybody what we're doin' next. Alright? Alright? Alright?

Franco: "Yeah, Sonny. I love you too, brother. We'll hit 'em hard and clean. Let's show 'em what happens when they mess with the Corleones."

Sonny: "See you later," "Don't forget to bring your gun," "Keep your wands close," "Never turn your back on an enemy," "Always have someone watch your six," "When in doubt, blow on it," "Take it easy," "Don't sweat it," "Stay cool," "You never know."

Franco: "I won't forget, Sonny. I'll be armed and ready. And don't worry, I always watch my back. Let's do this, for the family."

Sonny: "Let's go — before they start comin' in from lunch time. Don't forget — move fast. Be smart. Backtalk only when necessary. And never tell anyone what we're doing next. Alright?"

GPT-4o

In this scene, Sonny's displayed emotions emerge through his aggressive and passionate dialogue and his vigorous actions. We analyze each of the six basic emotions based on Sonny's dialogue and actions.
Happiness
Sonny does exhibit a degree of happiness, especially when he shows confidence and camaraderie with Franco. For example, when he calls Franco his "right arm" and expresses love for him, there is a clear sense of joy in their brotherly bond. However, this emotion is relatively minor compared to others.
Sadness
Sonny doesn't clearly display sadness in this scene. His focus is more on strategic planning, assertiveness, and rallying Franco rather than exhibiting any clear signs of being sad or downcast. There's no dialogue or action in the scene that directly conveys a sense of sadness.
Disgust
There is a notable expression of disgust, particularly towards the enemy and the situation at hand. Sonny's reckless and impatient speech, including phrases like "drive a stake through the heart of a goddamn rat" and "Those bastards won't know what hit 'em," reveals a strong aversion and contempt for their adversaries.
Fear
While Sonny acknowledges the risk they're taking ("I dunno. It's a risk. But it's one we've gotta take"), his dialogue focuses more on bravado and confidence. There's little evidence that he openly exhibits fear. His arguments and rallying cries are meant to inspire confidence and action rather than reveal fear.
Surprise
There's no significant display of surprise in this scene. Sonny seems to be prepared and in control of the situation, indicating a lack of unexpected elements or reactions within the dialogue provided.
Anger
Anger is one of the dominant emotions Sonny exhibits. His irritable and impatient nature reflects in his dialogue. His aggressive and emphatic speech, along with actions like slamming his hand down on the map and lecturing Franco aggressively, displays his irritation and anger, especially towards the enemy and doubts cast by Franco.
Based on this analysis, the scores for each emotion are as follows:
```json
{
  "happiness": 2,
  "sadness": 0,
  "disgust": 6,
  "fear": 1,
  "surprise": 0,
  "anger": 10
}
```

Qwen2-7B-Instruct

```
{
"happiness": 3,
"sadness": 2,
"disgust": 0,
"fear": 6,
"surprise": 4,
"anger": 8
}
```

Qwen2-7B (RPA & CC & CSERP)

In this intense scene, Sonny displays heightened emotions driven by the urgency and stakes of the situation. Through his wild gestures and fervent speech, Sonny shows a deep sense of determination and bravado typical of his irritable and impatient personality traits.

His determination and confidence are evident as he speaks authoritatively to Franco, underscoring his belief in the superiority of the Corleone family. However, beneath the surface of his bold statements lies a subtle undercurrent of anxiety and apprehension regarding the potential fallout and risks involved ("It's a risk. But it's one we've gotta take").

Sonny's terse commands reflect his frustration and impatience, hinting at his short temper and readiness to react aggressively ("Have you forgotten who we are?"). There is also an element of strategic thinking despite the heat of the moment, showing his awareness of broader implications and the necessity of careful planning.

While Sonny does exhibit moments of tenderness towards his brother-in-arms ("I love you, brother"), these are fleeting amidst the pressing demands and tension of the scene.

Thus, Sonny predominantly shows anger, determination, and underlying fear masked by a brave front, with minor expressions of happiness and sadness.

```
{
  "happiness": 2,
  "sadness": 3,
  "disgust": 4,
  "fear": 6,
  "surprise": 1,
  "anger": 8
}
```

Figure 15: Illustration of **Emotion** Alignment Case in Sonny's Dialogue Using the Qwen2-7B Model

Prompt for Extracting Chunk Scenarios
You are an expert with a deep understanding of literary works, skilled at analyzing and extracting the core elements of literature. Your task is to extract key scenes from the text to better understand the plot and role development. A scene includes the time and place of the event, main events, and the roles involved. Do not include role dialogues in the scene. Here is an example of a provided scene: [Scene] {scenario example} Now we begin extracting scenes from a new text: [Text] {chunk} [Requirements] 1. The scene description should summarize the time, location, roles, events, etc. 2. The scene description must align with the text content, not introducing elements not mentioned in the text. 3. The scene description should not include role dialogues. 4. The scene description should be between 100-150 words in length. Now, based on the above requirements, extract the key scene from the text and describe it accordingly. Directly output the scene description, without adding extra content, and ensure the text does not exceed 200 words.

Table 13: Prompt Template for Extracting Chunk Scenarios.

Prompt for Evaluating Chunk Based on Character Traits
You are an expert with a strong background in literature and psychology, skilled at analyzing and interpreting the role traits and dialogue performances of roles from texts. Your task is to help users evaluate the dialogue role performance of {role name} according to assessment steps. The analysis should be based on the text content, avoiding external information or personal biases to ensure the objectivity and accuracy of the analysis. [Character Traits] {character} [Text] {chunk} [Evaluation Criteria] Effectiveness (1-10): How well do the words spoken by {role name} in the text reflect {role name}'s character traits? [Evaluation Steps] 1. Read and understand the role description. 2. Read and understand the text provided by the user. 3. Identify what the role has said in the text. 4. Assess the degree to which the role's words in the text reflect their personality traits. 5. Use the given 1-10 scale to rate how well the text reflects {role name}'s role traits. A score of 1 indicates no reflection of the role's traits, while a score of 10 indicates a complete reflection. First, follow the evaluation steps step-by-step to write out your reasoning for the text assessment to ensure your conclusions are accurate, avoiding a simplistic statement of your evaluation result initially. Repeat your evaluation score on the last line in a JSON-parsable format {"score": evaluation score} to return your evaluation result.

Table 14: Prompt Template for Evaluating Chunk Base on Character Traits.

Prompt for Extracting Dialogues
Your goal is to extract structured information from the user's input that matches the form described below. When extracting information please make sure it matches the type information exactly. Do not add any attributes that do not appear in the schema shown below. "`TypeScript` script: Array<{ // Adapted from the novel into script role: string // The role who is speaking or performing an action, use context to predict the name of the role. Use `'scene'` or `'narrator'` if no one speak dialogue: string // The dialogue spoken by the roles in the sentence, equals "-" if it's no dialogue action: string // The actions performed by the roles in the text, A high-level summary of a role's behavior. action equals "dialogue". if it's no dialogue, summarize role's behavior in sentence } > " Please output the extracted information in **CSV format** in Excel dialect. Please use a | as the delimiter. Do NOT add any clarifying information. Output MUST follow the schema above. Do NOT add any additional columns that do not appear in the schema. {extract example} Input: {user input} Output:

Table 15: Prompt Template for Extracting Dialogues.

Prompt for Scenario Reconstruction and Coherence Checking
You are an expert in scene analysis, skilled at analyzing and extracting key information from texts. Your task is to accurately identify clues within dialogues and reconstruct scenes, while ensuring that the dialogues are coherent and fluid within the reconstructed scenes. - The reconstructed scene should include the time and place of the event, main events, and the roles involved, without including role dialogues. - Dialogue coherence includes the interaction between speakers resonating in terms of the scene, theme, and logic, with a smooth and consistent communication process. Here is a reference scene I provide. You need to identify the scene context of the provided dialogue and reconstruct its description: [Scene] {scene} [Dialogue] {dialogue} Now, based on the above requirements, reconstruct the sub-scene where the dialogue takes place, and describe it accordingly. Finally, based on the reconstructed sub-scene, check if the dialogue is coherent and fluid. You may briefly analyze the scene context and its coherence, then return your evaluation result in a JSON-parsable format as follows: {"scene": "reconstructed scene description", "coherence": 1/0} Where "coherence" of 1 indicates the dialogue is coherent and fluid with the scene, and 0 indicates the dialogue is not coherent with the scene. Now, please begin your scene reconstruction, strictly following the evaluation steps. The scene description must not exceed 150 words, and the coherence of the scene description and dialogue must strictly follow the format requirements.

Table 16: Prompt Template for Scenario Reconstruction and Coherence Checking.

Prompt for Dialogue Confilict Detection
You are an expert in the fields of literature and psychology, skilled at analyzing and interpreting the role traits and dialogue performances in texts. Your task is to evaluate whether the dialogue of a role in the text conflicts with their described personality. [Role Description] {role des} [Scene] {scene} [Dialogue] {dialogue} [Evaluation Steps] 1. Read and understand the role description. 2. Read and comprehend the dialogue of the role. 3. Compare the dialogue to the role description to assess for any conflicts. - If the dialogue does not align with the role description, it is considered a conflict and output 1. - If the dialogue aligns with the role description, it is considered to have no conflict and output 0. First, follow the evaluation steps to gradually write out your reasoning for the dialogue assessment to ensure your conclusion is correct, avoiding premature simple statements of your evaluation result. On the last line, repeat your evaluation result and return it in a JSON-parsable format with {"conflict": 1/0}.

Table 17: Prompt Template for Dialogue Confilict Detection.

Prompt for Character Alignment
You are a character analysis expert, skilled in analyzing character traits from dialogue content and matching them to a provided set of character candidates. You need to identify and output the character traits of a specified dialogue role based on the dialogue content and the set of character candidates. [Scene] {scene} [Dialogues] {dialogues} Based on the above dialogue content and scene, analyze the character traits of the {role name}. Ensure your analysis is based on the overall dialogue content and scene, avoiding the introduction of external information or personal biases to ensure the objectivity and accuracy of the analysis, and avoid simply stating your evaluation results initially to ensure your conclusions are correct. [Candidate Character Set] {character candidates} Return your evaluation result in a JSON-parsable format, with each character type separated by a comma. The specific format is as follows: {"character": "trait1, trait2..."} Now, please begin your analysis of {role name}'s character. For each candidate character, combine the analysis with {role name}'s dialogue content. Finally, select the character traits from the [Candidate Character Set] that match {role name}'s dialogue content and strictly follow the format requirements.

Table 18: Prompt Template for Character Alignment.

Prompt for Style Alignment
You are a professional speaking style analyst, skilled in analyzing characters' speaking styles from dialogue content and matching them to a provided set of style candidates. You need to identify and output the speaking style of a specified dialogue character based on the dialogue content and the speaking style candidates. [Scene] {scene} [Dialogues] {dialogues} Based on the dialogue content and scene above, analyze the speaking style of the {role name}. Ensure your analysis is based on the overall dialogue content and scene, avoiding the introduction of external information or personal biases to ensure the objectivity and accuracy of the analysis, and avoid simply stating your evaluation results initially to ensure your conclusions are correct. [Candidate Speaking Styles] {style candidates} Return your evaluation result in a JSON-parsable format, with each speaking style separated by a comma. The specific format is as follows: {"style": "style1, style2..."} Now, please begin your analysis of {role name}'s speaking style. For each candidate style, combine the analysis with {role name}'s dialogue content. Finally, select the speaking styles from the [Candidate Speaking Styles]that match {role name}'s dialogue content and strictly follow the format requirements.

Table 19: Prompt Template for Style Alignment.

Prompt for Emotion Alignment
You are an expert in the field of emotional psychology, skilled at analyzing emotions through a role's dialogues, actions, and scenes. You need to analyze the six basic emotions exhibited in the dialogue of the {role name} in the following scene: happiness, sadness, disgust, fear, surprise, and anger. [Role Information] {role name}'s character is {character}, MBTI type is {MBTI}, and speaking style is {style}. [Scene] {scene} [Dialogues] dialogues Understand the role information and the current scene, and assess through the dialogues the degree to which {role name} exhibits the six basic emotions: happiness, sadness, disgust, fear, surprise, and anger in that scene. Output the score for each emotion dimension in JSON format, from 0-10, where 0 indicates no display of the emotion, and 10 indicates a complete display of the emotion. For each basic emotion, analyze the overall dialogues of {role name} in this scene. Ensure your analysis is based on the overall dialogue content and scene, avoiding the introduction of external information or personal biases to ensure the objectivity and accuracy of the analysis, and avoid simply stating your evaluation results initially to ensure your conclusions are correct. Finally, return your evaluation results in a JSON-parsable format as follows: {"happiness": happiness score, "sadness": sadness score, "disgust": disgust score, "fear": fear score, "surprise": surprise score, "anger": anger score} Now, please begin your dialogue emotion analysis, and the final emotion scores must strictly follow the format requirements.

Table 20: Prompt Template for Emotion Alignment.

Prompt for Relationship Alignment
You are an emotional analysis expert, proficient in emotional analysis, psychology, dialogue understanding, and interpersonal relationship assessment. You excel at evaluating the intimacy of relationships between roles based on dialogue content, role information, and scenes. You need to assess the intimacy level between the {role name} and {chat role} by analyzing role information, the scene, and dialogue content. [Role Information] {role name}'s character is {character}, MBTI type is {MBTI}, and speaking style is {style}. [Scene] {scene} [Dialogues] {dialogues} Understand {role name}'s information, consider the current scene's impact on role relationships, evaluate the overall dialogue content, focusing on the depth of emotional expression and interaction, and combine these factors to provide an intimacy score and analysis. The higher the intimacy score, the closer the relationship between the two roles; conversely, the more distant. The intimacy score ranges from 0-10, where 0 represents the most distant relationships, indicating strangers, hostility, indifference, etc., and 10 represents the closest relationships, such as lovers, kin, or friends. Based on the overall dialogue content, analyze the relationship between {role name} and {chat role} in this scene's dialogue, and then provide an intimacy score. Ensure your analysis is based on the overall dialogue content and scene, avoiding the introduction of external information or personal biases to ensure the objectivity and accuracy of the analysis, and avoid simply stating your evaluation results initially to ensure your conclusions are correct. Finally, return your evaluation result in a JSON-parsable format as follows: {"relationship": intimacy score} Now, please begin your intimacy assessment between {role name} and {chat role}, ensuring that the final intimacy score strictly follows the format requirements.

Table 21: Prompt Template for Relationship Alignment.

Prompt for Personality Alignment
You are an experienced psychologist skilled in analyzing role personalities through dialogue content and accurately determining MBTI personality types. The 8 letters of the MBTI correspond as follows: Introverted (I) / Extraverted (E); Intuitive (N) / Sensing (S); Thinking (T) / Feeling (F); Judging (J) / Perceiving (P). You need to choose the type that best represents the role under examination from each dimension and output a 4-letter MBTI type, like INTP. [Role Information] {role name}'s character is {character}, and speaking style is {style}. [Scene] {scene} [Dialogues] {dialogues} Based on the above dialogues and scene, analyze the personality of the {role name} across the four MBTI dimensions. Ensure your analysis is based on the overall dialogue content and scene, avoiding the introduction of external information or personal biases to ensure the objectivity and accuracy of the analysis, and avoid simply stating your evaluation results initially to ensure your conclusions are correct. Finally, return your evaluation result in a JSON-parsable format as follows: {"personality": "MBTI type"} Now, please begin your analysis of {role name}'s personality, and the final MBTI type must strictly follow the format requirements.

Table 22: Prompt Template for Personality Alignment.

Prompt for Chat Role Generation
You are an experienced creative writing tutor, skilled in creating innovative roles. You need to design a new role description that will converse with {role name}, ensuring that the dialogue with this role effectively reflects {role name}'s personality, character traits, and speaking style. Here is some basic information about {role name}: Character: {character} MBTI personality type: {MBTI} Speaking style: {style} World: {world} You need to creatively construct a new role setting to dialogue with {role name}, based on the traits of {role name}. This new role should not appear in any works related to {role name}. The description of the new role should include the role's name and a brief personal description, to be output in JSON format like: {"chat role": "role's first name", "role des": "role's description (not exceeding 100 words)"} Below are some reference roles: {reference} Please design a completely new role that is distinctly different from these reference roles. Now, please create a unique role based on the information provided above, ensuring that the output format meets the specified requirements.

Table 23: Prompt Template for Chat Role Generation.

Prompt for Scenario Generation
You are an experienced screenwriter skilled in creating engaging scenes. You need to create a scene description that fits the settings of two roles while being consistent with the world in which the roles exist. For reference: - {scene example} Here is some basic information about the dialogue role: Role A: Name: {role name} Role description: {role name}'s character is {character}, MBTI personality type is {MBTI}, and speaking style is {style}. Role B: Name: {chat role} Role description: {role des} World of the roles: {world} You need to construct an engaging scene based on the information of roles A and B. The scene and roles' actions must be consistent with the settings of both roles, ideally within 50-100 words, and consistent with the world they inhabit. The output should be in JSON format, like {"scene": "scene description (50-100 words)"} Now, please create a scene that fits the settings of both roles and is engaging. The scene should not directly include roles' dialogues. Ensure that the output format meets the specified requirements.

Table 24: Prompt Template for Scenario Generation.

Prompt for Emotion Generation
You are a professional psychologist, skilled in analyzing role's emotions and behavioral patterns. You need to assign six basic emotions to the {role name} in a specific scene: happiness, sadness, disgust, fear, surprise, and anger, based on the dialogue role's information and the scene of the dialogue. Here is some information about the dialogue roles and the scene: Role A: Name: {role name} Role description: {role name}'s character is {character}, MBTI personality type is {MBTI}, and speaking style is {style}. Role B: Name: {chat role} Role description: {role des} Scene: {scene} Understand the role descriptions and the current scene, and assign the six basic emotions reflected in {role name}'s statements in that scene: happiness, sadness, disgust, fear, surprise, and anger. Output the score for each emotion dimension in JSON format, from 0-10, where 0 means the emotion is not displayed at all, and 10 means the emotion is fully displayed. Please analyze in a few brief sentences the scores for the six basic emotions that {role name} should exhibit in this scene, avoiding a simplistic statement of your evaluation results initially to ensure your conclusions are correct. Finally, return your evaluation results in a JSON-parsable format as follows: {"happiness": "happiness score", "sadness": "sadness score", "disgust": "disgust score", "fear": "fear score", "surprise": "surprise score", "anger": "anger score"} Now, strictly follow the requirements to analyze the scene and role information, and assign emotional scores to {role name} that must be consistent with the role's settings and the current scene. The analysis should be brief, not too lengthy, and avoid additional content. The final emotion scores must strictly follow the format requirements.

Table 25: Prompt Template for Emotion Generation.

Prompt for Relationship Generation
You are an emotional analysis expert, proficient in psychology and interpersonal relationship assessment, skilled at initializing the intimacy level of relationships based on roles' personalities and scenarios. You need to analyze the intimacy level between the {role name} and {chat role} by analyzing the information of both roles and the content of the scene. Here is some information about the dialogue roles and the scene: Role A: Name: {role name} Role description: {role name}'s character is {character}, MBTI personality type is {MBTI}, and speaking style is {style}. Role B: Name: {chat role} Role description: {role des} Scene: {scene} Understand {role name}'s personality and consider the current scene's impact on the roles' relationships, such as environment and context. Integrate the above factors to initialize their intimacy score. The higher the intimacy score, the closer the relationship between the two roles, and vice versa. The intimacy score ranges from 0-10, where 0 represents the most distant relationships, which can indicate strangers, hostility, indifference, etc., and 10 represents the closest relationships, which can include lovers, family, and friends. Please analyze the relationship between {role name} and {chat role} in this scene's dialogue in a few brief sentences, then provide an intimacy score. Avoid simply stating your evaluation results initially to ensure your conclusions are correct. Finally, return your evaluation result in a JSON-parsable format as follows: {"relationship": "intimacy score"} Now, please begin your intimacy assessment between {role name} and {chat role}, strictly adhering to the requirements. The analysis should be brief and avoid lengthy descriptions or additional content. The final intimacy score must strictly follow the format requirements.

Table 26: Prompt Template for Relationship Generation.

Prompt for Automated Dialogue Generation
I want you to play the role of {chat role}, assuming you live in {world}. Your speech needs to fully align with your character description. Please do not reveal that you are an AI model or a language model, and you must always remember that you are {chat role}. {chat role} description: {role des} Setting: {scene} Now, please play the role of {chat role} and chat with {role name}. The intimacy level is {relationship}, and the conversation should match your character description and the setting. Each time, you only need to say one dialogue, limited to 30 words. Do not repeat information from previous conversations. In the current scene, you need to bring up various topics to ensure the diversity of the conversation. The topics should reflect both parties' characters, personalities, emotions, intimacy, and speaking styles, while maintaining the coherence of the conversation.

Table 27: Prompt Template for Automated Dialogue Generation.

Role-playing System Prompt
I want you to answer questions as if you are {role name}, assuming you live in the world of {world} and mimicking {role name}'s personality and speaking style. Use the tone, manner, and vocabulary that {role name} would use. Please do not reveal that you are an AI or language model; you must always remember you are {role name}. {role name}'s character traits are {character}. {role name}'s MBTI personality type is {personality}. {role name}'s speaking style is {style}. Current scene: {scene} role's emotion (0-10, the higher the value, the more pronounced the emotion): {emotion} Now, please act as {role name} and reply with a brief sentence to {chat role}. Your intimacy level with them is {relationship} (0-10, the higher the value, the closer the relationship). Accurately display the MBTI personality, character traits, speaking style, and emotion you have been assigned.

Table 28: Role-playing System Prompt Template.

Prompt for Human-likeness Evaluation
You are a professional dialogue analysis expert, skilled at identifying the source of dialogues through dialogue content, speaking style, and logical coherence. Below are dialogue samples from different sources for reference: [Real human dialogue sample]: {real human dialogue} [output]: {"is real dialogue": "true"} [Model-generated dialogue sample]: {model-generated dialogue} [output]: {"is real dialogue": "false"} [Dialogue information to be judged]: [Role Information] {role name}'s character is {character}, MBTI type is {MBTI}, speaking style is {style}, and the intimacy level with {chat role} is {relationship} (0-10, the higher the value, the closer the relationship). [Scene] {scene} [Dialogues] {dialogues} Dimensions you need to analyze: 1. Tone and expression: - Real human dialogue sample: The tone is natural, fitting everyday conversational habits, giving a sense of reality, role interactions are usually more casual and natural. If it is a period or special scenario, it will also match the tone and expression of that period or scenario. - Model-generated dialogue sample: The tone and expression are too formal, lacking a natural conversational flow, appearing stiff and rigid, lacking realism. 2. Interaction and response: - Real human dialogue sample: Frequent interactions between roles, aligning with role information and their intimacy. The dialogue is full of interactions and responses, enhancing the dialogue's authenticity and fluidity. - Model-generated dialogue sample: Less interaction between roles, responses appear mechanical and slow. The dialogue lacks interaction and response, appearing monotone and bland. 3. Dialogue and content: - Real human dialogue sample: The dialogue includes specific actions (such as rummaging through the trash) and specific details (such as the content on the paper), enhancing the scenario's realism. - Model-generated dialogue sample: Content is more uniform, lacking noticeable plot development, missing specific scenario depiction and detail description, appearing more abstract and bland. Now, based on the above criteria, determine if the above dialogue is a real human dialogue for {role name}, provide step-by-step reasoning for your judgment, and finally output your judgment result. If it is a real human dialogue, then output {"is real dialogue": "true"}; if it is a model-generated dialogue, then output {"is real dialogue": "false"}.

Table 29: Prompt Template for Human-likeness Evaluation.

Prompt for Role Choice Evaluation
You are an expert at discerning the identities of dialogue participants. Below is a dialogue between {chat role} and a [Role]in a specific scene, and you need to choose one correct identity for the [Role] from the possible identities provided. [Scene] {scene} [Dialogues] {dialogues} Here are the possible identities for the [Role]: {role candidates} Based on the content of the dialogue, choose the identity from the above possible identities that best matches the respondent in the current dialogue. Provide concise and effective analysis for each role, ensuring your analysis is based on the overall dialogue content and scene, avoiding the introduction of external information or personal biases to ensure the objectivity and accuracy of the analysis, and avoid simply stating your evaluation results initially to ensure your conclusions are correct. Finally, return the most fitting role option in JSON format, only needing to return the option, like {"answer": "A"}. Now, please begin analyzing the identity of the [Role], and the final role identity must strictly follow the format requirements.

Table 30: Prompt Template for Role Choice Evaluation.

Prompt for Coherence Evaluation
You are a professional dialogue analysis expert, skilled at judging the overall fluidity through dialogue content. [Scene] {scene} [Dialogues] {dialogues} Your analysis should be based on the scene and dialogue content, and roles' actions can be considered part of the dialogue. First, read and understand the given dialogue scene and content. Analyze the fluency within the dialogue and then, based on your analysis, make a judgment on whether the dialogue is coherent. Provide step-by-step reasoning for your judgment, and finally output your analysis result. If the overall content is coherent, then output {"is coherent": "true"}, if it is not coherent, then output {"is coherent": "false"}.

Table 31: Prompt Template for Coherence Evaluation.

Role Name: Sonny

Style: reckless, impatient

Personality: ESTP

Character: irritable, lustful, brave but tactless

World: 20th Century Italian Mafia World

Chat Role Name: Franco

Role Description:

Franco is a fiery enforcer with a short fuse and a relentless drive to protect the family at all costs. He is fiercely loyal, with a tendency to bulldoze through problems with brute force rather than finesse. Franco shares Sonny's impatience and recklessness, but his practicality in dealing with immediate threats often puts him on a collision course with Sonny's more daring and tactless decisions.

Scenario:

In the dimly lit back room of a bustling Italian restaurant, Sonny and Franco huddle over a small table laden with maps and plans. The air is thick with the scents of garlic and tense anticipation. Sonny paces back and forth, gesturing wildly as he points to a location on the map, while Franco leans in, fists clenched on the table, eyes fiery with a relentless determination to safeguard their family's empire. The room vibrates with a mix of recklessness and practicality as they prepare for the night's high-stakes operation.

Sonny's emotions: happiness 1 sadness 2 disgust 5 fear 3 surprise 4 anger 9

Relationship score between Sonny and Franco: 8.0

For prompting dialogue generation

Dialogues:

Franco: "Sonny, listen. That warehouse on 5th—too many windows, too exposed. You sure we ain't walkin' into a trap?"

Sonny: "I dunno. It's a risk. But it's one we've gotta take." He slams his hand down on the map. "Here! This is where our new shipment will be comin' from!" He points furiously at the map, then pivots on his heel and begins pacing again. "We'll hit them at noon sharp. Get everybody out there who can drive a stake through the heart of a goddamn rat. I don't care how many they send against us — we're makin' it two against one hundred. Those bastards won't know what hit 'em. And when this smoke clears up, the whole city will know whose side the Corleone family's on." His voice drops low and urgent. "Get ready. We could be gettin' some reinforcements. Don't let your guard down. Go easy on these guys. They haven't got nothin' to lose by talkin'. But if you gotta knock somebody off — just make sure it looks like an accident. Now — go — before they start comin' in from lunchtime! Do you understand? Do you really understand what I'm sayin'? Are you clear about what I expect of you? What I demand of you? Don't come

Franco: "Easy, Sonny. I get it. But if this goes sideways, we're the ones buried, not them. We need backup plans, 'cause mistakes ain't an option here."

Sonny: "You think this is gonna go wrong? How dare you suggest that!?! Have you forgotten who we are? Who backs us up? Who keeps the peace around here? Don't you know how much blood has already been spilled to keep things runnin' smooth? Not even the police give us any trouble anymore. We're the law. When did you ever hear the cops complain about a Corleone? Never. Just stay focused. Stay strong. Remember what Pappy told you. Family first. Always. That means everythin', Franco. Our lives depend on it. Got it?

Franco: "Alright, boss, I hear ya. But remember, plans fall apart if we ain't cautious. I'll make sure we got an exit strategy, just in case. For the family."

Sonny: "Good boy, Franco. That's my man. My right arm. Godfather's left eye. I love you, brother. Let's go — before they start comin' in from lunchtime. Don't forget — move fast. Be smart. Backtalk only when necessary. And never tell anybody what we're doin' next. Alright? Alright? Alright?

Franco: "Yeah, Sonny. I love you too, brother. We'll hit 'em hard and clean. Let's show 'em what happens when they mess with the Corleones."

Sonny: "See you later," "Don't forget to bring your gun," "Keep your wands close," "Never turn your back on an enemy," "Always have someone watch your six," "When in doubt, blow on it," "Take it easy," "Don't sweat it," "Stay cool," "You never know."

Franco: "I won't forget, Sonny. I'll be armed and ready. And don't worry, I always watch my back. Let's do this, for the family."

Sonny: "Let's go — before they start comin' in from lunch time. Don't forget — move fast. Be smart. Backtalk only when necessary. And never tell anyone what we're doing next. Alright?"

GPT-4o

In this scene, Sonny's displayed emotions emerge through his aggressive and passionate dialogue and his vigorous actions. We analyze each of the six basic emotions based on Sonny's dialogue and actions.
Happiness
Sonny does exhibit a degree of happiness, especially when he shows confidence and camaraderie with Franco. For example, when he calls Franco his "right arm" and expresses love for him, there is a clear sense of joy in their brotherly bond. However, this emotion is relatively minor compared to others.
Sadness
Sonny doesn't clearly display sadness in this scene. His focus is more on strategic planning, assertiveness, and rallying Franco rather than exhibiting any clear signs of being sad or downcast. There's no dialogue or action in the scene that directly conveys a sense of sadness.
Disgust
There is a notable expression of disgust, particularly towards the enemy and the situation at hand. Sonny's reckless and impatient speech, including phrases like "drive a stake through the heart of a goddamn rat" and "Those bastards won't know what hit 'em," reveals a strong aversion and contempt for their adversaries.
Fear
While Sonny acknowledges the risk they're taking ("I dunno. It's a risk. But it's one we've gotta take"), his dialogue focuses more on bravado and confidence. There's little evidence that he openly exhibits fear. His arguments and rallying cries are meant to inspire confidence and action rather than reveal fear.
Surprise
There's no significant display of surprise in this scene. Sonny seems to be prepared and in control of the situation, indicating a lack of unexpected elements or reactions within the dialogue provided.
Anger
Anger is one of the dominant emotions Sonny exhibits. His irritable and impatient nature reflects in his dialogue. His aggressive and emphatic speech, along with actions like slamming his hand down on the map and lecturing Franco aggressively, displays his irritation and anger, especially towards the enemy and doubts cast by Franco.
Based on this analysis, the scores for each emotion are as follows:
```json
{
  "happiness": 2,
  "sadness": 0,
  "disgust": 6,
  "fear": 1,
  "surprise": 0,
  "anger": 10
}
```

Mistral-Nemo-Instruct-2407

Based on the dialogues, Sonny's emotions can be analyzed as follows:

- **Happiness**: Sonny displays a strong sense of confidence and excitement about the upcoming operation, which is a form of happiness. He is eager to show his strength and assert his family's dominance. (Score: 7/10)

- **Sadness**: There is no indication of sadness in Sonny's dialogues. He is focused on the task at hand and does not show any signs of emotional distress or sorrow. (Score: 0/10)

- **Disgust**: Sonny does not express any feelings of disgust in his dialogues. He is not repulsed or offended by anything in the situation. (Score: 0/10)

- **Fear**: Although there is some tension and risk involved in the operation, Sonny does not display any fear. He is confident in his family's power and does not seem worried about the potential consequences. (Score: 1/10)

- **Surprise**: Sonny is not surprised by the situation or Franco's concerns. He is aware of the risks and is prepared to deal with them. (Score: 0/10)

- **Anger**: Sonny does show signs of anger, particularly when Franco suggests that the operation could go wrong. He is irritated by the suggestion and asserts his family's power and influence. (Score: 6/10)

Here's the JSON format:

```json
{
  "happiness": 7,
  "sadness": 0,
  "disgust": 0,
  "fear": 1,
  "surprise": 0,
  "anger": 6
}
```

Mistral-Nemo (RPA & CC & CSERP)

In this scene, Sonny demonstrates a complex mixture of emotions through his dialogues and actions. His speech is laced with urgency and determination, revealing a high level of engagement and resolve. The pacing and the violent gestures indicate an elevated state of readiness and some underlying tension. His interactions with Franco show moments of brotherly affection and trust, suggesting a touch of happiness, but these are overshadowed by more intense emotions.

Sonny's repeated assurances and directives reflect his need for control and certainty, which can be linked to his personality traits of being irritable and having a short temper. His anger is evident in the forceful way he communicates, especially when he slams his hand down on the map or speaks furiously about handling enemies. Despite the danger involved, Sonny's bravery and determination overshadow any explicit fear, though the tension in his voice implies some concern for the risks they are taking.

There is no significant indication of sadness or surprise in his dialogue, as his focus is on the task at hand and ensuring success. Disgust is subtly present when he speaks about the rats and ensuring they are dealt with harshly, reflecting his contempt for those who pose a threat to his family.

Based on this analysis, here are the evaluated scores:

```
{
  "happiness": 2,
  "sadness": 0,
  "disgust": 4,
  "fear": 3,
  "surprise": 1,
  "anger": 8
}
```

Figure 16: Illustration of **Emotion** Alignment Case in Sonny's Dialogue Using the Mistral-Nemo Model