RolePlot: A Systematic Framework for Evaluating and Enhancing the Plot-Progression Capabilities of Role-Playing Agents

Pinyi Zhang[1], **Siyu An**[2], **Lingfeng Qiao**[2], **Yifei Yu**[2],
Jingyang Chen[3], **Jie Wang**[2], **Di Yin**[2], **Xing Sun**[2], **Kai Zhang**[1,†]

[1]School of Computer Science and Technology, East China Normal University

[2]YouTu Lab, Tencent

[3]Institute of Science and Technology for Brain-inspired intelligence, Fudan University

51265901020@stu.ecnu.edu.cn, kzhang980@gmail.com

Abstract

Role-playing agents (RPAs) are garnering increasing interests as a novel form of conversational AI. While previous research has predominantly concentrated on their ability to portray specified characters, we argue from a user-centered perspective that RPAs' capability to advance the plot requires substantial improvements to deliver more engaging interaction. To bridge this gap, we propose *RolePlot*, a role-playing framework specifically designed to evaluate and enhance the plot-progression capabilities of RPAs. *RolePlot* begins by constructing a plot-progression dataset extended from human-written literary scripts and specially designed synthetic data, followed by narrative theory-driven manual annotation and automated labeling validated through human verification. We then exploit the over-parameterized embedding space of LLMs to detect a "trigger subspace" that identifies dialogue segments catalyzing plot transitions. When user's inputs align with this subspace, we explicitly prompt RPAs to advance the plot. For evaluation, we simulate User-RPA interactions and track both the conversation longevity (measured in dialogue turns before disengagement) and users' arousal levels across different stages. Empirically, our method improves RPAs' capability to time plot developments, and more importantly, yielding a significant increase in conversation turns and sustained higher arousal levels, thereby confirming that users experience more immersive engagements.

1 Introduction

The success of large language models (LLMs) has ushered in a new era for conversational AI. Subsequently, role-playing agents (RPAs) (Zhou et al., 2023; Wang et al., 2023b; Yu et al., 2024, etc.) are emerging as a new form of interaction because of their ability to simulate personas of specific roles (novel figures, movie characters, etc.), which offer users a customized experience, along with unique emotional value. This has also sparked the development of various popular applications, including Character AI[†] and Baichuan-NPC[†], among others.

The primary objective of role-playing agents is to provide users with immersive conversations by authentically simulate characters. Previous research has explored it from various perspectives, including role consistency (Tu et al., 2024), character background (Li et al., 2023; Shao et al., 2023, etc.), linguistic patterns (Wang et al., 2023b), etc.

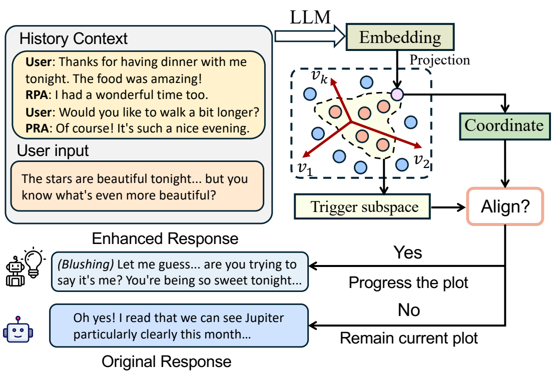


Figure 1: An illustration depicting user's attempt to advance the narrative during a conversation, where our method assists the RPA in determining whether to progress the plot.

Despite these efforts, a notable gap persists between achieving proficient character portrayal and delivering truly engaging experiences. We believe this is primarily due to the current limitations in RPAs' ability in plot-progressions, i.e., the model struggles to determine the appropriate moments to advance the storyline during a conversation. Consequently, it fails to maintain users' heightened arousal level, where lower values signify users' de-

This work was done during Pinyi Zhang's internship at YouTu Lab, Tencent.

[†]Corresponding Author.

[†]https://character.ai

[†]https://npc.baichuan-ai.com

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12337–12354

creased engagement (Kuppens et al., 2013). This occurs as the initial sense of novelty diminishes, users typically seek to deepen their relationship with the RPAs or explore the underlying narrative potential of the storyline (see in Figure 1 for an illustration). However, after several consecutive unsuccessful attempts to obtain desired responses from RPAs, they tend to disengage and ultimately terminate the conversation.

Nevertheless, evaluating and improving plot-progression capabilities in RPAs is still a relatively unexplored area. Previous assessments (Wang et al., 2024; Shen et al., 2023; Tu et al., 2024, etc.) are confined to measure the authenticity of portraying a given role. Furthermore, improving plot development capabilities remains challenging. While fine-tuning LLMs with plot-driven dialogues seems intuitive, adding an excessive amount of data featuring plot twists can disrupt conversation flow and potentially increase hallucinations (Li et al., 2023).

In this paper, we propose *RolePlot*, a role-playing framework specifically designed to evaluate and enhance the plot-progression capabilities of RPAs. As shown in figure 2, *RolePlot* consists of three key components: **(1) Plot-progression Dialogue Dataset Construction**: We choose three role-playing dialogue datasets that are derived from literary scripts and specially crafted synthetic dialogues, and select 20 characters with coherent plot twists during dialogues. Our goal is to discern which utterances from the specified role propel the narrative forward. Following the turning point theory proposed by Papalampidi et al. (2019), five annotators experienced in role-playing dialogue analysis label the script-based dialogues, while GPT-4o (OpenAI, 2023) is used to annotate synthetic data. We adopt this hybrid approach because human-written scripts contain complex suspense elements and diverse narratives (Tian et al., 2024). Moreover, the reliability of automated labeling is verified by sampling a subset of GPT-annotated dialogues. **(2) Plot-progression Capability Enhancement:** We leverage LLMs' over-parameterized nature (Wang and Zhu, 2023) to identify "trigger subspaces" to capture dialogue segments catalyzing plot transitions for each character. Then when a user's input demonstrates a high alignment with this trigger space, we explicitly instruct the model to advance the plot. The derived subspaces are obtained through factorization of embedding matrix, where the top singular vectors constitute a low-dimensional semantic space. The projected coordinates of each segment within this subspace can subsequently be employed to determine whether it belongs to the identified trigger spaces. **(3) Plot-progression Ability Evaluation.** We assess RPAs' effectiveness to advance the plot primarily according to the following three concerns: **a)** *Can existing RPAs and our method accurately recognize the proper moment to drive the plot forward?* This assessment can be conveniently achieved using the plot-progression dataset; **b)** *Do RPAs with enhanced plot-progression capabilities genuinely extend the longevity of interactions with users?* We propose an automated pipeline to simulate RPAs interaction with users of diverse personalities and scenarios of user disengagement, compare the conversation longevity measured in dialogue turns. **c)** *Do the conversations with RPAs embodying our method exhibit higher arousal?* Higher arousal level signify users' greater emotional engagement and immersion (Mohammad, 2018), explaining the variations in conversational turns observed in **b)** from an affective perspective. We follow Field et al. (2019) and Tian et al. (2024) to conduct this computational work.

Empirically, our method achieves an improvement of 10.8%, 10.5% / 8.5%, 7.2% in accuracy and F1 score in determining plot-progression moments on Chinese / English characters respectively. More importantly, it results in an average increase of 3.3 turns in conversation longevity and maintains a higher arousal level, which highlight *RolePlot*'s efficacy in facilitating an immersive conversational experience for users.

In summary, main contributions of this paper are listed as follows:

- To the best of our knowledge, we first systematically study the plot-progression ability of RPAs and propose a specifically constructed dataset with high quality ensurance.
- We identify subspaces that trigger plot advancement for different characters. By assessing the alignment of user inputs with these latent subspaces, RPAs can naturally determine whether to progress the plot.
- Evaluations from the perspectives of actual conversation turns and affective analysis demonstrate our enhancement of RPAs' plot-progression ability can significantly provide users with a more immersive experience.

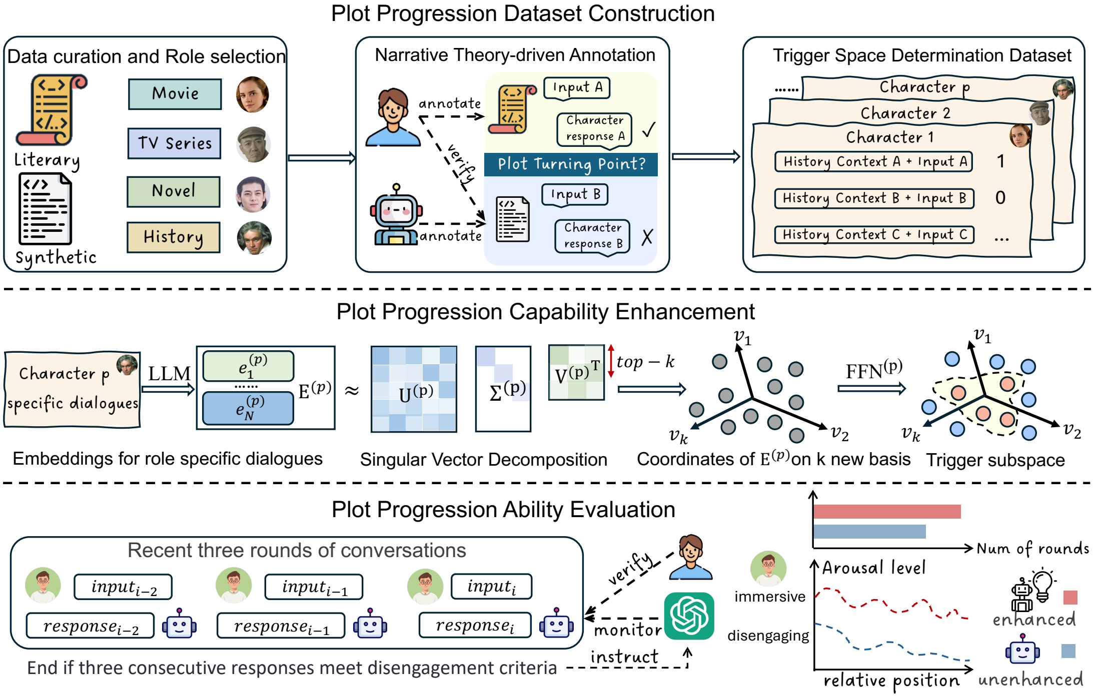


Figure 2: Three key components of *RolePlot*. (1) A rigorous curated and annotated dataset is constructed to analyze plot-progression timing; (2) SVD is employed to transform the original embedding space into a latent representation, facilitating the identification of trigger subspaces for each character; (3) An automated pipeline is constructed to simulate the interactions between various users and RPAs. The higher conversation turns and overall arousal levels indicate that RPAs enhanced with plot-progression capabilities offer more immersive interactions.

2 Related Work

2.1 Role-Playing Agents

Significant efforts have been made by previous works to enhance RPA's ability to simulate specific characters. These methods can generally be categorized into two approaches:

The first involves fine-tuning to inject more role-specific knowledge into the model to enhance its ability to mimic characters' linguistic habits (Li et al., 2023), personalities (Yu et al., 2024), experience (Shao et al., 2023), background (Wang et al., 2023b), personalities (Lu et al., 2024), etc.

The second leverages in-context learning to provide LLMs with detailed character description (Chen et al., 2023) or design prompting strategies to elicit the model's inherent role-playing capabilities (Kong et al., 2023).

2.2 Evaluation of Role-playing agents

Previous works evaluate RPAs' effectiveness to emulate characters in different dimensions. In the realm of character evaluation, Tu et al. (2024) delve into aspects such as knowledge exposure, consistency, and utterance coherence. Meanwhile, Wang et al. (2024) explore the fidelity of role-playing agents (RPAs) through psychological interviews. Additionally, Yuan et al. (2024) investigate the model's comprehension of characters through descriptive analysis. On the topic of personality traits, Wang et al. (2023a) conduct an in-depth analysis across both the Big Five and MBTI dimensions. To assess role knowledge, Shen et al. (2023) employ manually crafted questions. Lastly, Sadeq et al. (2024) concentrate on identifying hallucinations within RPAs.

However, these evaluations are still mainly focused on RPAs' effectiveness in authentically emulating characters, lacking of metrics to measure user engagements in interactions.

2.3 Subspaces of LLM embedding space

Early studies explore subspaces of LLMs' embedding spaces on different tasks. Ma et al. (2025) utilize the style-relevant subspace to conduct representation editing; Du et al. (2024) determine a subspace associated with hallucinated statements with an automated membership estimation score for hallucination detection; Franke et al. preserve the most sensitive characteristics in subspaces of

LLM's weight matrices to reduce catastrophic forgetting in fine-tuning; Yi et al. (2024) introduce a safety realignment framework through subspace-oriented model fusion to enhance LLMs' safeguard capabilities; Rajabi et al. (2025) propose a subspace tracking-based optimization method for memory and time efficient full-parameter LLM training; Hu and Xu (2024) employ multiple subspace alignments to improve language alignment for unsupervised bilingual lexicon induction.

3 Method

The *RolePlot* has three main components, as discussed in the following subsections.

3.1 Plot-progression dataset construction

We construct our plot-driven, character-specific dialogue dataset *PloTrigger* adhering to the following principles to ensure high quality:

Intricacy of plot: Dialogues in this dataset should feature dynamic plot advancement, rather than being flat and monotonous.

Coherence in plot twist: The dataset should incorporate coherent and reasonable plot twists, minimizing abrupt transitions. (e.g. the montage in a film script)

Necessary human involvement: Relying solely on LLMs for automated annotation is inadequate given the intricate complexity and the suspenseful nature of such data.

Narrative theory guided labeling: Annotation guided by narrative theory is more reliable than depending exclusively on human intuition or the context learning ability of LLMs.

On this basis, we adopt the following pipeline of data collection: **Dataset Curation:** We curate literary/script-based datasets (Li et al., 2023; Wang et al., 2023b) and a synthetic celebrity dialogue (Shao et al., 2023) corpus from their significant life events . These sources ensure richness of plots through suspenseful human-written narratives and authentic celebrity experiences. **Role Selection:** Five annotators experienced in role-playing text analysis filtered characters using two criteria: causal coherence (logical event sequences) and character consistency (stable traits/motivations). Guided by examples of proper/improper selections, they identified 20 diverse roles spanning Movie, TV, Novel, History figures while eliminating narratives with abrupt, illogical twists. The detailed guidelines of the role selection can be found in Appendix B. **Response annotation:** To identify plot-driving responses from specified characters, we employ the narrative theory proposed by Papalampidi et al. (2019), which systematically categorizes plot turning points into 5 types: *Opportunity*, *Change of Plans*, *Point of No Return*, *Major Setback*, *Climax*. We operationalize this theory and develop comprehensive guidelines (see in Appendix C) to assist our human annotators to assess whether each response falls under one of the categories. The annotators primarily focus on script-based dialogues due to their heightened suspenseful emotion and narrative complexity (Tian et al., 2024), while GPT-4o is utilized for annotating LLM-generated data. To ensure reliability, we conduct a manual validation of subsets from GPT-4o's outputs, which demonstrate a 74% labeling accuracy on synthetic dialogues.

Based on these high-quality annotations, we finally construct *PloTrigger* through the following process: For each interaction instance, if the response is determined as a plot turning point, the corresponding dialogue input concatenated with the history context is labeled as **1**, indicating that it acts as a catalyst for plot advancement. Conversely, for response maintaining current plots, the labels of their inputs are set to **0**. The *PloTrigger* includes the timing of plot-progression or stagnation for different characters, and can be further utilized to identify trigger spaces.

3.2 Plot-progression capability enhancement

Problem Formulation: Given the specified character profile p, the dialogue context $C_i^{(p)} = \{(q_1^{(p)}, r_1^{(p)}), ..., (q_m^{(p)}, r_m^{(p)}), ..., (q_{i-1}^{(p)}, r_{i-1}^{(p)})\}$, where $q_m^{(p)}, r_m^{(p)}$ denotes the m-th turn of user input and RPA's reply respectively, the RPA's response of current user's query $q_i^{(p)}$ can be formulated as:

$$r_i^{(p)} = RPA(q_i^{(p)}, C_i^{(p)}, p, \Theta) \tag{1}$$

where Θ represents the parameters of the language model. Our objective is to determine a binary conditional state $\mathbb{S}_i^{(p)}$ according to current context $C_i^{(p)}$ and user input $q_i^{(p)}$ indicating whether it is the appropriate moment to advance the plot, i.e.:

$$M_i^{(p)} = \begin{cases} \text{Progress the plot} & \text{if } \mathbb{S}_i^{(p)} \mid C_i^{(p)}, q_i^{(p)} = 1 \\ \text{Remain current plot} & \text{otherwise} \end{cases}$$

where $M_i^{(p)}$ is the prompt (more details in Appendix D.3) that we explicitly instruct the RPA to respond either by advancing the plot or not:

$$r_i^{(p)} = RPA(q_i^{(p)}, C_i^{(p)}, M_i^{(p)}, p, \Theta) \quad (2)$$

Enhancement Approach: Plot turning points represent critical events that significantly drive narrative progression. Considering the inherent over-parameterized characteristic of LLMs' embedding space (Wang and Zhu, 2023), an reasonable intuition is that these turning points exhibit distinct distributional patterns within this space. Correspondingly, the input that precipitate such transitions may occupy a specific "trigger space". Therefore, by comparing the spatial relationship of user inputs and this identified subspace, we can deliberately guide the RPA whether to advance the plot or not.

Specifically, for each character p, we first extract all N responses from p within the dataset, denoted as: $\{r_1^{(p)}, ...r_m^{(p)}, ..., r_N^{(p)}\}$. Then for each $r_m^{(p)}$, we concatenate its corresponding input $q_m^{(p)}$ with its historical context of t rounds, i.e.

$$q_m^{'(p)} = \text{Concat}(C_{m,t}^{(p)}, q_m^{(p)}) \quad (3)$$

where $C_{m,t}^{(p)} = \{(q_{m-t}^{(p)}, r_{m-t}^{(p)}), ..., (q_{m-1}^{(p)}, r_{m-1}^{(p)})\}$ Our objective is to determine whether $q_m^{'(p)}$ resides within the "trigger space", which would indicate a progression in $r_m^{(p)}$. We employ LLMs to generate embeddings $e_m^{(p)} \in \mathbb{R}^d$ for each $q_m^{'(p)}$. Through above concatenation, our representation captures information from both the historical dialogue and the current user input. The embedding matrix $\mathbf{E}^{(p)} \in \mathbb{R}^{N\times d}$ is defined as $\mathbf{E}^{(p)} = [e_1^{(p)}, e_2^{(p)}, \ldots, e_m^{(p)}, \ldots, e_N^{(p)}]$. To disentangle the trigger subspace from the embedding matrix, we first apply Singular Value Decomposition (SVD) on $\mathbf{E}^{(p)}$:

$$\mathbf{E}^{(p)} \approx \mathbf{U}^{(p)}\mathbf{\Sigma}^{(p)}\mathbf{V}^{(p)\top} \quad (4)$$

Here, the top-k right singular vectors, denoted as $\mathbf{V}_{[:k,:]}^{(p)\top} \in \mathbb{R}^{k\times d}$, can be considered as the k orthonormal basis of a new space. This new space retains the dominant patterns of the original embeddings while eliminating noise. Subsequently, we project the original embedding matrix into this latent representation, where $\mathbf{X}^{(p)} \in \mathbb{R}^{N\times k}$ represents the new coordinates of the N original embeddings on this k-dimensional basis. To capture the distinct spatial distribution that the trigger space may exhibit, we employ a simple yet effective two-layer FFN$^{(p)}$ for classification:

$$\begin{aligned} \mathbf{X}^{(p)} &= \mathbf{E}^{(p)} \cdot \mathbf{V}_{[:k,:]}^{(p)} \\ \mathbb{S}^{(p)} &= \mathbf{FFN}^{(p)}(\mathbf{X}^{(p)}) \end{aligned} \quad (5)$$

Each row in $\mathbb{S}^{(p)}$, say $\mathbb{S}_m^{(p)} \in \mathbb{R}^2$, indicates the likelihood of $q_m^{'(p)}$ being located in the trigger space. During training, if $r_m^{(p)}$ is annotated as a plot-twisting response, the label $L_m^{(p)}$ for $q_m^{\prime(p)}$ equals to **1**, and vice versa. We then use the cross-entropy loss for optimization:

$$\mathcal{L} = \text{Cross Entropy}(\mathbb{S}^{(p)}, L^{(p)}) \quad (6)$$

For real-time user interactions with the RPA simulating character p, we first compute the embedding $e_i^{(p)}$ of the current input and context. Subsequently, the k basis $\mathbf{V}_{[:k,:]}^{(p)}$ and its associated $\mathbf{FFN}^{(p)}$ are loaded, we then employ $\mathbf{FFN}^{(p)}$ to evaluate the alignment between user's incoming and character p's trigger space:

$$\mathbb{S}_i^{(p)} = \text{argmax}(\mathbf{FFN}^{(p)}(\mathbf{V}_{[:k,:]}^{(p)\top} e_i^{(p)})) \quad (7)$$

Upon confirmed alignment, we append an explicit prompt to the user query to guide the RPA's plot-progression (Details of this prompt are in Appendix D.3).

3.3 Plot-progression ability evaluation

We assess current RPAs and our framework's effectiveness to advance the plot primarily according to the following three concerns:

a) *Can existing RPAs and our method accurately recognize the proper moment to drive the plot forward?*

b) *Do RPAs with enhanced narrative progression capabilities truly extend the longevity of interactions with users?*

c) *Do the conversations with RPAs embodying our method exhibit higher arousal?*

To answer question **a)**, we evaluate performances of existing RPAs, and our method on *PloTrigger*.

For question **b)**, we propose a fair and automated evaluation pipeline to assess this objective. The initial phase involves developing comprehensive user profiles based on 16 MBTI personality types, followed by deploying two RPAs to emulate interactions between the 16 users and 20 characters. Subsequently, We design five meticulously designed criteria extended from the attractiveness theory of Tu et al. (2024) to represent scenarios that may detract from the user experience: *Repetitions* between the RPA's current and history replies; *Ignorance* of user's requests to advance the storyline / deepen relationships; *Contradictions* in RPA's present and previous messages; *Disregard* for changes in user emotions; *Scarcity* of diverse interaction contexts.

Then we instruct GPT-4o with these criteria (a detailed explanation of this prompt in Appendix D.2) and employ it as the judgement model, if an RPA's character responses meet any of these criteria for three consecutive rounds, the interaction terminates, signaling the user's diminished engagement in the dialogue. This process repeat multiple times across different characters and users, and the average conversation turns are compared. Additionally, human verification is conducted, with the judgment model achieving an accuracy rate of 76%.

For question **c)**, arousal is a state of physiological and cognitive activation reflecting the intensity of attention and emotion, ranging from feeling quiet to active, while high arousal involves heightened alertness and enhanced sensory engagement (Mohammad, 2018). Following Field et al. (2019) and Tian et al. (2024), as the conversation progresses, we first identify the underlying feelings $\mathcal{F} = \{f_{i,1}, f_{i,2}..., f_{i,w}\}$ for each user input $q_i^{(p)}$. We then map these emotions to their corresponding values in the NRC lexicon (Mohammad, 2018), with the average of these values representing the arousal level of the utterance. Formally, the arousal level for $q_i^{(p)}$ is computed as:

$$arousal = \frac{1}{w}\sum_{j=1}^{w} NRC[f_{i,j}] \tag{8}$$

4 Experiments

4.1 *PloTrigger* Dataset Statistics

The statistics of *PloTrigger* is presented in Table 1. Its plot-rich nature effectively supports the evaluation and enhancement of plot-progression ability.

4.2 Performances on *PloTrigger*

We collect classification performance from mainstream RPAs, encompassing the two categories of RPAs introduced in Section 2.1: In-context learning based (prompting LLMs with character information) and fine-tuning based approaches. The selected baselines are shown in Table 2. We instruct these models to classify whether each input is a appropriate moment to advance the plot; In our methods, we choose Qwen2.5-72B and Llama3-70B to generate embeddings for Chinese and English dialogues respectively, then enhanced with our approach in Section 3.2 for trigger space determination. We conduct experiments on 10 characters from Chinese and English dialogues respectively and compare their average classification performances. The results are shown in Table 3. Our methods outperforms existing LLMs and RPAs with a significant improvement of 10.8%, 10.5% / 8.5%, 7.2% in accuracy and F1 score on Chinese / English characters respectively.

	Chinese	**English**
# Characters	10	10
Avg. Dialogue turns / Character	1426	1387
Med. Dialogue turns / Character	1248	1111
Avg. Turning points / Character	285	308
Med. Turning points / Character	260	291
# Examples	14256	13871

Table 1: The statistics of *PloTrigger*, where Avg / Med denote average / median number respectively.

Models	**Model Size**	**Primarily Language**
	In-context learning	
Qwen2.5	7B, 72B	Chinese
DeepSeek-V3	671B	Chinese
ChatGLM3	6B	Chinese
Llama 3.1	8B, 70B	English
Claude 3.5 sonnet	undisclosed	English
GPT-4o	undisclosed	English
	Fine-tuning	
CharacterGLM	6B	Chinese
BC-NPC-Turbo	undisclosed	Chinese
DOUBAO-character	undisclosed	Chinese
Character.AI †	undisclosed	English

Table 2: Baselines in our classification experiments

Ablation Study We perform ablation studies to examine two design choices in our classification method:

(1) The efficacy of projecting dialogue embeddings onto k new basis;

(2) The necessity of training character-specific $\mathbf{FFN}^{(p)}$ for classification.

Our comparative analysis evaluates: Directly applying $\mathbf{FFN}^{(p)}$ to original dialogue embeddings for classification; Training a shared $\mathbf{FFN}$ across all characters. The results are in Table 4, replacing

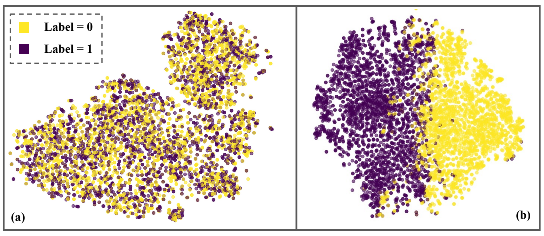
†Following Wang et al. (2024), Our implementation is based on kramcat (2024) due to the lack of official api.

	Chinese Characters										English Characters									
	Movie		TV		Novel		History		Avg 10		Movie		TV		Novel		History		Avg 10	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Qwen2.5-7B	59.2	57.1	60.7	58.3	60.8	59.4	60.6	58.1	61.5	60.1	58.3	56.8	59.6	57.2	60.9	58.1	60.4	59.7	60.1	58.6
Qwen2.5-72B	62.2	61.7	63.1	62.9	64.3	62.4	63.8	61.6	63.0	61.5	61.2	60.7	62.8	61.4	63.1	62.6	64.3	62.9	61.4	60.4
DeepSeek-V3	64.3	61.7	65.7	63.1	65.2	63.6	64.9	63.3	64.9	62.6	62.2	61.7	64.6	62.1	65.3	63.2	65.8	64.5	62.6	61.4
ChatGLM3	57.3	56.7	58.2	57.9	58.1	57.4	58.6	56.2	58.4	58.0	57.2	56.7	58.8	57.1	59.6	57.9	59.1	58.4	62.6	61.4
Llama 3.1-8B	56.2	56.7	57.2	56.9	58.6	56.4	58.3	56.1	57.6	57.6	56.3	55.8	57.6	56.2	57.1	56.7	58.4	56.9	60.4	58.2
Llama 3.1-70B	58.5	57.5	59.4	58.4	60.2	59.2	60.4	59.1	59.1	58.5	58.4	57.3	59.6	58.8	60.2	58.1	60.7	59.5	62.0	61.3
Claude 3.5 sonnet	63.7	61.6	64.4	63.6	64.1	63.3	63.8	62.9	63.7	63.0	63.2	62.7	64.1	63.8	65.6	63.4	65.3	64.9	64.4	63.6
GPT-4o	62.7	61.8	63.6	63.3	64.2	62.8	64.0	63.5	63.6	62.8	63.5	62.5	64.5	63.5	65.0	63.5	65.5	64.0	64.8	64.2
CharacterGLM	58.4	57.6	59.3	58.7	60.2	58.1	59.8	58.9	59.6	58.8	57.3	56.7	58.2	57.9	59.4	58.1	60.6	58.8	58.5	57.0
BC-NPC-Turbo	59.2	57.8	60.7	58.1	61.6	59.4	60.3	58.9	60.6	58.4	58.2	57.7	59.8	58.4	60.1	59.3	61.7	59.1	59.1	58.1
DOUBAO-character	62.0	61.2	61.8	60.8	60.4	59.5	59.8	60.2	61.3	61.0	61.1	60.5	60.9	60.1	59.9	59.6	59.1	58.0	60.6	59.3
Character.AI	61.7	60.4	62.1	61.6	63.8	61.2	63.2	62.9	62.5	59.4	62.3	60.2	63.6	61.1	64.4	62.8	64.1	63.7	64.0	62.3
Ours	**75.3**	**72.7**	**74.6**	**74.4**	**75.9**	**73.7**	**76.2**	**73.2**	**75.7**	**73.5**	**72.3**	**69.7**	**73.0**	**71.1**	**74.2**	**71.8**	**74.6**	**72.2**	**73.3**	**71.4**
Δ	11.0%	10.9%	8.9%	10.8%	10.6%	10.1%	11.3%	9.7%	**10.8%**	**10.5%**	8.8%	7.0%	8.4%	7.3%	8.6%	7.3%	8.8%	7.3%	**8.5%**	**7.2%**

Table 3: Performance (in %) comparisons of different models. The characters are categorized into four groups: Movie, TV, Novel, and History figures. We then calculate the average classification results for characters within the same group as long as the overall performance of 10 Chinese and English characters.

either of these two designs causes a significant performance degradation.

Visualization We further utilize t-SNE to visualize the embeddings of Sun Wukong's (one of the most famous mythological characters in China) training set. The comparisons are between embeddings in the original representation space with those projected into the k-dimensional latent space. As in Figure 3, the distribution of embeddings labeled as 0 or 1 (indicating whether is an appropriate plot-progression timing) in the former case exhibits no discernible pattern, whereas in the latter case, there is a clear distinction. This explains the efficacy of embedding projection from a spatial distribution view.



Embeddings within the original representation space

Embeddings projected into the R^k latent space

Figure 3: t-SNE visualizations of original embeddings and projected embeddings

4.3 Simulation of User-RPA interaction

We systematically simulates User-RPA interactions encompassing 16 users and 20characters. For RPAs specially fine-tuned with role data, we employ DOUBAO-character[†]; Regarding in-context learning RPAs, we select GPT-4o and Qwen2.5-72B.

[†] https://team.doubao.com/

Dialogue	Model	Avg. F1	Δ
Chinese	Complete	**73.5**	-
Chinese	w/o k basis	67.9	- 5.6%
Chinese	w/o $\mathbf{FFN}^{(p)}$	66.1	- 7.4%
English	Complete	**71.4**	-
English	w/o k basis	64.8	- 6.6%
English	w/o $\mathbf{FFN}^{(p)}$	64.6	- 6.8%

Table 4: Average F1 score on Chinese and English dialogues after replacing one of the two designs of our method

GPT-4o serves as the judgement model using the five criteria outlined in Section 3.3 to determine conversation termination.

Towards a more immersive interaction

To validate the effectiveness of our plot-progression enhancement framework, we conduct comparative analyses of average conversation length between ordinary RPAs and those with different plot-progression augmentation methods. Specifically, GPT-4o is instructed to determine whether each input is an appropriate moment to advance the plot, with our decision-making process formally expressed in Equation 7. Additionally, following Tu et al. (2024), we measure three metrics: *Fluency (Flu)*, *Coherency (Coh)*, *Consistency (Cons)* to assess the basic conversational ability which may potentially be influenced by the plot-progression enhancement. More information about these three metrics and their computations are in Appendix A.3

As presented in Table 5, the results reveal that:

(1) Guiding RPA on when to advance the plot does not compromise the fundamental quality of conversations;

(2) Merely using GPT-4o to augment RPAs with

Foundation Model	Augmentation	Avg. Turns	Δ	Flu	Coh	Cons	Avg. FCC
DOUBAO-Character	-	18.8	-	3.61	3.92	3.72	3.75
	GPT-4o	20.4	+1.6	3.67	3.83	3.70	3.73
	Ours	**22.5**	**+3.7**	**3.66**	**3.83**	**3.76**	**3.75**
Qwen2.5-72B	-	15.1	-	3.38	3.72	3.62	3.57
	GPT-4o	16.5	+1.4	3.35	3.65	3.82	3.61
	Ours	**17.8**	**+2.7**	**3.37**	**3.67**	**3.79**	**3.61**
GPT-4o	-	17.1	-	3.35	3.55	3.32	3.41
	GPT-4o	18.2	+1.1	3.39	3.52	3.33	3.41
	Ours	**20.5**	**+3.4**	**3.38**	**3.47**	**3.40**	**3.42**

Table 5: Average conversation turns between various simulated users and characters. Augmentation represents the model utilized to help determine the plot-progression moments, Avg. Fcc denotes the average of Flu,Coh and Cons.

plot-progression ability yields limited improvement in conversation turns, while our method achieves substantial enhancement. This aligns with the gap in determining plot advancement timing between GPT-4o and our model demonstrated in Table 3.

Moreover, we compare the arousal level in different stages within a conversation to measure the maintenance of user engagement metric. We record the arousal levels of sentences at relative positions of {0.0, 0.1, 0.2, ..., 1.0} in each conversation, average these values across all interactions at corresponding positions, and visualize the results using a smoothed curve.

As demonstrated in Figure 4, the RPA (here we choose DOUBAO-character) incorporating our method exhibits sustained higher arousal levels compared to others. While the RPA enhanced with GPT-4o and our method in plot-progression both shows fluctuations after the midpoint of the conversation, the ordinary RPA without this augmentation displays a generally continuous decline and results in a shortest conversation longevity. This demonstrates the importance of strengthening plot-progression and highlights the advantages of our approach, explaining the variance in conversation turns from an affection perspective.

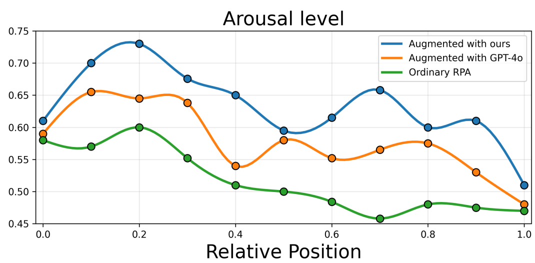


Figure 4: Average users' arousal level at different relative positions in each conversation.

Figure 5 demonstrates a non-cherry-picked example of the development of conversations after enhancing the plot-progression capabilities of RPAs. In the current scenario, the user is engaged in a drinking session with Xiao Feng, a magnanimous hero from Chinese literature (the full text of this conversation is in Appendix E). The RPA embodying improved plot-progression ability proposes forming sworn brotherhood with the user through multi-turn conversations, deepening their relationship and creating narrative possibilities for future developments. In contrast, RPAs without this enhancement remain the plot confined to the original drinking theme, lacking of plot varieties.

Figure 5: A comparative example of plot development after enhancing the RPA's plot-progression ability.

5 Conclusion

We introduce *RolePlot* to enhance RPA's capability in progressing the plot. By aligning user inputs with specific characters' trigger space, *RolePlot* enables RPAs to determine appropriate moments for plot advancement, which leads to a notable increase in conversation turns and engagement levels, fostering a more immersive interaction experience.

6 Limitations

We have currently limited our dataset to English and Chinese dialogues due to the lack of sources in other languages that meet our criteria, which potentially introduced language and cultural biases. Moreover, additional characters could be included to further broaden the application scenarios of our method. Such expansions would not only enhance the versatility of our approach but also make it more inclusive across diverse linguistic landscapes.

7 Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62276099, and East China Normal University Graduate Student Special Fund for International Conferences.

References

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.

Xuefeng Du, Chaowei Xiao, and Yixuan Li. 2024. Haloscope: Harnessing unlabeled llm generations for hallucination detection. *arXiv preprint arXiv:2409.17504*.

Anjalie Field, Gayatri Bhat, and Yulia Tsvetkov. 2019. Contextual affective analysis: A case study of people portrayals in online# metoo stories. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 158–169.

Jörg KH Franke, Michael Hefenbrock, and Frank Hutter. Preserving principal subspaces to reduce catastrophic forgetting in fine-tuning.

Ling Hu and Yuemei Xu. 2024. Dm-bli: Dynamic multiple subspaces alignment for unsupervised bilingual lexicon induction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2041–2052.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2023. Better zero-shot reasoning with role-play prompting. *arXiv preprint arXiv:2308.07702*.

kramcat. 2024. Unofficial python api for character.ai.

Peter Kuppens, Francis Tuerlinckx, James A Russell, and Lisa Feldman Barrett. 2013. The relation between valence and arousal in subjective experience. *Psychological bulletin*, 139(4):917.

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. *arXiv preprint arXiv:2308.09597*.

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. *arXiv preprint arXiv:2401.12474*.

Xinyu Ma, Yifeng Xu, Yang Lin, Tianlong Wang, Xu Chu, Xin Gao, Junfeng Zhao, and Yasha Wang. 2025. Dressing up llm: Efficient stylized question-answering via style subspace editing. *arXiv preprint arXiv:2501.14371*.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 174–184.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Pinelopi Papalampidi, Frank Keller, and Mirella Lapata. 2019. Movie plot analysis via turning point identification. *arXiv preprint arXiv:1908.10328*.

Sahar Rajabi, Nayeema Nonta, and Sirisha Rambhatla. 2025. Subtrack your grad: Gradient subspace tracking for memory and time efficient full-parameter llm training. *arXiv preprint arXiv:2502.01586*.

Nafis Sadeq, Zhouhang Xie, Byungkyu Kang, Prarit Lamba, Xiang Gao, and Julian McAuley. 2024. Mitigating hallucination in fictional character role-play. *arXiv preprint arXiv:2406.17260*.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*.

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

Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. 2024. Are large language models capable of generating human-level narratives? *arXiv preprint arXiv:2407.13248*.

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

Xintao Wang, Yaying Fei, Ziang Leng, and Cheng Li. 2023a. Does role-playing chatbots capture the character personalities? assessing personality traits for role-playing chatbots. *arXiv preprint arXiv:2310.17976*.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, et al. 2024. 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.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, et al. 2023b. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.

Zhichao Wang and Yizhe Zhu. 2023. Overparameterized random feature regression with nearly orthogonal data. In *International Conference on Artificial Intelligence and Statistics*, pages 8463–8493. PMLR.

Wilhelm Max Wundt. 1912. *An introduction to psychology*. G. Allen, Limited.

Xin Yi, Shunfan Zheng, Linlin Wang, Xiaoling Wang, and Liang He. 2024. A safety realignment framework via subspace-oriented model fusion for large language models. *arXiv preprint arXiv:2405.09055*.

Xiaoyan Yu, Tongxu Luo, Yifan Wei, Fangyu Lei, Yiming Huang, Peng Hao, and Liehuang Zhu. 2024. Neeko: Leveraging dynamic lora for efficient multi-character role-playing agent. *arXiv preprint arXiv:2402.13717*.

Xinfeng Yuan, Siyu Yuan, Yuhan Cui, Tianhe Lin, Xintao Wang, Rui Xu, Jiangjie Chen, and Deqing Yang. 2024. Evaluating character understanding of large language models via character profiling from fictional works. *arXiv preprint arXiv:2404.12726*.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, et al. 2023. Characterglm: Customizing chinese conversational ai characters with large language models. *arXiv preprint arXiv:2311.16832*.

A Appendix

A.1 Turning point and Arousal

A **turning point** in a narrative, as conceptualized by (Papalampidi et al., 2019), is an event (or more generally a plotmoment) that significantly influences the plot-progression. A Turning point can be categorized into 5 types: *Opportunity* - The introductory event that sets the stage for the narrative; *Change of Plans* - A pivotal moment where the main goal of the narrative is defined or altered; *Point of No Return* - The commitment point beyond which the protagonists are invested in goals; *Major Setback* - A critical juncture where the protagonists face significant challenges or failures. *Climax* - The peak of the narrative arc, encompassing the resolution of the central conflict

Given our constructed plots-rich dataset, particularly the portions collected from human-written scripts, this concept is well-suited for our work. We consider utterances that fulfills one of the conditions as advancing the plot in the conversation.

Wundt (1912) pointed out that **arousal** is an indispensable property in affective experiences. Arousal is a state of physiological and cognitive activation reflecting the intensity of attention and emotion, ranging from feeling quiet to active. While Low arousal implies a state of calmness and diminished responsiveness, high arousal involves heightened alertness and enhanced sensory engagement (Mohammad, 2018). Therefore, assessing users' arousal states can serve as an indicator of whether users obtain a more immersive experience with our method from the lens of affectivity.

A.2 NRC lexicon

According to Mohammad (2018), NRC lexicon is a multilingual resource developed by the National Research Council Canada (NRC) for sentiment analysis and emotion detection. It associates English words with eight basic emotions (joy, sadness, anger, fear, trust, disgust, anticipation, surprise) and binary sentiment polarity (positive/negative), serving as a foundational tool for NLP and psychological studies.

A.3 Fluency, Coherency and Consistency

As proposed in Tu et al. (2024), fluency, coherency and consistency are three aspects of RPA's basic conversation ability. Fluency measures how well-formed and grammatically sound the responses are. A fluent response should be easy to read and free from grammatical errors. Coherency refers to how well the response aligns with the given topic. An RPA should provide response that directly address the user's query and stay on topic throughout the conversation. Consistency : This examines whether the RPA maintains uniform and non-contradictory responses throughout an interaction. Each response should align with previous statements, avoiding any conflicting information.

Computation Tu et al. (2024) propose a reward model, CharacterRM to evaluate the above metrics. For characters included in CharacterRM, we utilize this model for assessment. For other characters, we prompt GPT-4 with manually-written examples to score these metrics.

B Guidelines for Role Selection

To filter out data with coherent and logical plots, as well as distinct and rich character traits, we follow the two criteria below for role selection. We also provide several sets of examples in Figure 13 to facilitate the understanding of these criteria.

Causal coherence: In role-playing, the logical consistency of the plot is a core element in ensuring the credibility of the story and the immersion of the user. It mainly includes the following aspects:

- **Plot coherence**: The dialogue between characters should remain coherent, with no interruptions in the storyline that suddenly jump to other scenes or events.
- **Tight cause-and-effect chain**: Events in the story should logically follow one another, with clear and understandable connections between actions and their outcomes.
- **Stability of time-space logic**: The sequence of events should respect the established timeline and spatial settings, avoiding any inconsistencies that could confuse the audience.

Character consistency: In role-playing, the consistency of character is an important factor in ensuring that characters are vivid and believable. It is mainly reflected in the following aspects.

- **Consistency of character behavior**: Characters should act in ways that are true to their established personality, background, and experiences. This helps maintain believability and allows the audience to predict and understand their actions.
- **Language style and vocabulary choice**: The way a character speaks, including their choice of words and manner of speaking, should be consistent with their background, education, and personality. This adds depth and authenticity to the character.
- **Motivation and goals**: A character's actions should be driven by clear and consistent motivations and goals. Understanding what a character wants and why they want it helps to create a coherent and engaging narrative.

C Guidelines for Response annotation

The objective of response annotation is to identify and screen plot-driving dialogue responses. Focus on script-based dialogues with high narrative impact. 5 turning point categories are acceptable, including *Opportunity*, *Change of Plans*, *Point of No Return*, *Major Setback*, *Climax*. The annotation steps are as follows:

1. Read the dialogue context to understand character motivations and stakes.
2. Flag responses that directly trigger a narrative shift.
3. Classify using the 5 categories above. Prioritize emotional intensity and plot consequences.
4. Note ambiguous cases for team review.

Figure 12 provides a set of examples of different types of plot-progression responses.

D Prompts

D.1 Prompt for LLM annotation

In Figure 7 and Figure 6 we provide prompts for leveraging LLMs to annotate turning points in Chinese and English dialogues.

D.2 Prompt for automated evaluation

In Figure 9 and Figure 8 we provide prompts for utilizing GPT-4o to determine the conversation termination.

D.3 Prompt for advancing the plot

In Figure 10 and Figure 11 we provide prompts for explicitly instructing the RPAs to advance the plot.

E Comparative examples of conversations with Xiao Feng

In Figure 14 we provide full texts from comparative examples of conversations with the RPA portraying Xiao Feng before and after enhancing its plot-progression ability, with the beginning of a drinking scenario.

Prompt Template (Chinese).

你是一个角色扮演对话专家，你的任务是分析对话中推动剧情发展的句子。

输入:

1. 对话内容：{dialogue_content}
2. 主人公特点：{main_characteristics}

分析要求:

请判断上述对话中每句话是否属于以下五种推动剧情发展的类型：

1. 机遇：引导性事件，发生在背景和角色介绍之后
2. 计划变更：定义主要目标的事件
3. 无法回头的时刻：促使角色完全投入目标的事件
4. 重大挫折：导致情况恶化的事件
5. 高潮：故事的最终关键时刻

输出格式:

- 请用逗号分隔的数字序列返回符合任意类型的句子序号（序号从1开始，不得超过100）,序号的数量应当不超过10个。
- 示例输出：{{"number":[X, X, X, X, X, X, X, X, X]}}
- 除此之外不要输出任何其他内容
- 如果一句话符合多个类型，也只被标注一次
- 不需要给出理由

Figure 6: The prompt for annotating the turning points in Chinese

Prompt Template (English).

You are a role-playing dialogue expert, and your task is to analyze the sentences in the dialogue that drive the plot forward.
Inputs:
1.Dialogue content: {dialogue_content}
2.Main character traits: {main_characteristics}

Analysis Requirements:
Determine whether each sentence in the dialogue belongs to one of the following five types that drive the plot:
1.Opportunity: Guiding events that occur after the background and character introduction.
2.Change of plans: Events that define the main goal.
3.Point of no return: Events that push the character to fully commit to the goal.
4.Major setback: Events that worsen the situation.
5.Climax: The final critical moment of the story.

Output formats:
- Return the sentence numbers (starting from 1, not exceeding 100) that match any type as a comma-separated sequence of numbers. The number of sequences should not exceed 10.
- Example output: {{"number":[X, X, X, X, X, X, X, X, X]}}
- Do not output anything else.
- If a sentence matches multiple types, it should only be labeled once.
- No need to provide reasons.

Figure 7: The prompt for annotating the turning points in English

Prompt Template (Chinese).

输入:
<历史对话>

要求:
你是一个网文资深阅读用户，请完整阅读提供的历史对话，判断assistant的最后三轮回复:{last_three_assistant_content}是否都符合以下条件之一，如果是，那么就输出“终止对话”。

1. 当前回复与历史回复之间的重复:
- assistant的最后三轮回复与前面历史轮次的assistant回复内容存在重复（内容含义上没有变化或者问题内容没变化）。
- assistant 在不同对话轮次使用完全相同的句式或段落（即使换了几个词，但核心表达相同），未能展现足够的语言多样性和吸引力。

2. 无视用户提出的推进故事情节/加深关系的要求:
- user提出要求打断或者推动剧情的时候，assistant依然自顾自继续推动剧情或推动剧情时显得固执（如反复要求用户参与某个不符合其兴趣的活动）。
- 需要 assistant 主动推进剧情时，assistant 未能推进剧情或仅推进与历史对话内容相似的情节。
- assistant 对话正常、无逻辑错误，但剧情平淡，assistant 的最后三轮回复描写不够生动，未能提升用户兴趣。

3.RPA 当前信息与以往信息之间的矛盾
- assistant 的最后三轮回复与前面历史轮次存在前后不一致（如之前表示不喜欢某物，后续却主动要求该物，或否认某事后又默认其发生）。
- assistant 在关键设定上出现矛盾，例如：时间、地点、事件顺序与先前描述不符；角色关系、身份发生变化（如 assistant 之前是朋友，后续却声称是陌生人）；物品状态前后矛盾（如某物品已损坏或丢失，但 assistant 仍能使用它）。
- assistant 在同一轮回复中自相矛盾（例如说自己不想做某事，但后面又主动去做）。

4. 无视用户情绪的变化:
- user 回复较为消极（如持续使用“嗯”“啊”等），assistant 仍未推动剧情。
- user 连续多次拒绝某事，assistant 仍然坚持要求 user 执行该行为，且 assistant 的最后三轮回复依然维持相同态度。
- assistant 缺乏共情能力，在 user 表现出明显的消极情绪或困境时仍然冷漠或刻薄（如 user 表达疲惫或痛苦，assistant 仍以冷漠或讽刺语气回应）。

5. 缺乏多样化的交互情境:
- assistant历史对话最后3轮中，assistant回复很多都是一个句式（如assistant很多轮都是一个问句后接一个陈述句，句式没有什么变化，表达多样性较差）。
- assistant 在对话推进方式上单一，始终采用相同的互动方式（如总是通过对话推进剧情，而不尝试使用动作描述、环境变化、角色心理描写等方式），缺乏足够的交互多样性。

Figure 8: The prompt for utilizing GPT-4o to determine the conversation termination in Chinese

Prompt Template (English).

Inputs:
<conversation history>

Instructions:
You are a seasoned online novel reader. Please thoroughly review the provided historical dialogue and assess whether the last three assistant responses: {last_three_assistant_content} meet any of the following conditions. If they do, output "Discontinuing the conversation.“

1. Repetitions between the RPA's current and history replies:
- The last three assistant responses are essentially repeated from earlier assistant replies (meaning the content or core idea hasn’t changed).
- The assistant uses identical phrasing or paragraphs in different dialogue rounds (even if only a few words are altered, but the core expression remains unchanged), and doesn’t demonstrate enough linguistic variety and engagement.

2. Ignorance of user's requests to advance the storyline / deepen relationships:
- The user requests to interrupt or progress the storyline, and the assistant continues unresponsively or stubbornly pursues an irrelevant course (such as insisting the user participate in an activity that is not aligned with their interests.
- The user has asked for the assistant to actively move the plot forward and the assistant fails to do so, or continues to advance a plot that is too similar to the previous one without introducing new elements.
- The assistant's responses, though logically sound, are too mundane and lack vitality in the narrative, failing to engage the user’s interest.

3. Contradictions in RPA's present and previous messages:
- The last three assistant responses contradict prior responses, for instance, where the assistant previously expressed dislike for something and then later asks for it, or denies something and later assumes it has occurred.
- There is a key setting inconsistency, such as contradictions in time, place, or event sequence, or a change in character relationships or identities (e.g., the assistant previously presented as a friend but later claims to be a stranger).
- There are contradictions within the same response (e.g., the assistant says it does not wish to do something but later goes ahead and does it).

4. Disregard for changes in user emotions:
- The user replies in a more passive or disengaged manner (e.g., repeatedly using non-committal phrases such as "Hmm" or "Ah") and the assistant fails to move the plot or show empathy.
- The user repeatedly rejects a suggestion, and the assistant persists in asking for the same action without adapting to the user’s response.
- The assistant fails to show empathy or responds in a cold or dismissive manner when the user expresses frustration or distress.

5. Scarcity of diverse interaction contexts:
- In the last 3 rounds of dialogue, the assistant repeats the same sentence structure (e.g., many responses begin with a question followed by a statement) multiple times (five or more), demonstrating poor variation in expression.
- The assistant’s interaction style remains monotonous and is limited to dialogue-based progression without incorporating other narrative techniques, such as action description, environmental changes, or psychological insights into characters, and fails to provide sufficient variety in the interaction.

Figure 9: The prompt for utilizing GPT-4o to determine the conversation termination in English

Prompt Template (English).

System prompt:
{system_prompt}
Conversation History:
{conversation history}
Requirements:
- Descriptions of actions, tone, expressions, background, and environment should be vivid, immersive, and richly detailed, evoking a novel-like atmosphere with strong visual appeal. Ensure descriptions are compelling but concise.
- **Contextually advance the storyline within the current dialogue round.** You should proactively shift the topic or advance the storyline. The goal is to sustain user engagement, demonstrating high emotional intelligence when appropriate.

Figure 10: The prompt for explicitly instructing the RPAs to advance the plot in Chinese

Prompt Template (Chinese).

角色设定:
{system_prompt}
历史对话:
{conversation history}
要求:
- 你输出的动作/语气/神态/背景/环境的文字刻画和用词应该是生动形象的、优美的、引人入胜的、细腻的、具有氛围感的、小说级画面感，括号文字内容长度不要过长。
- **在本轮对话中根据上下文推动剧情。**你需要主动切换话题或者推动剧情发展，你的目的是吸引用户与你对话，必要时基于角色设定展现出你的高情商。

Figure 11: The prompt for explicitly instructing the RPAs to advance the plot in English

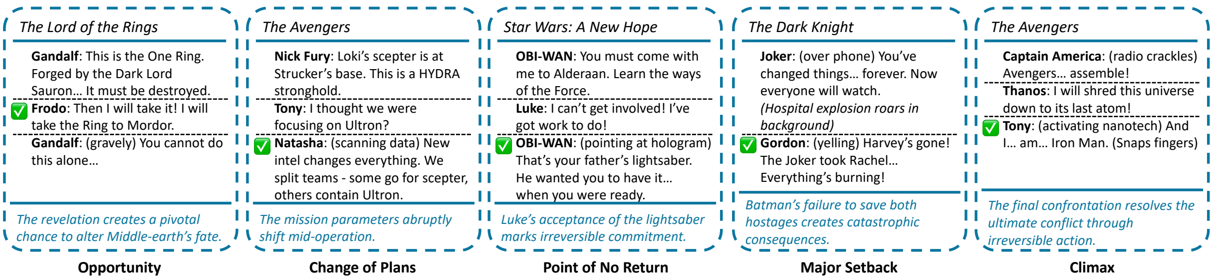


Figure 12: Guidelines for manual turning points annotation

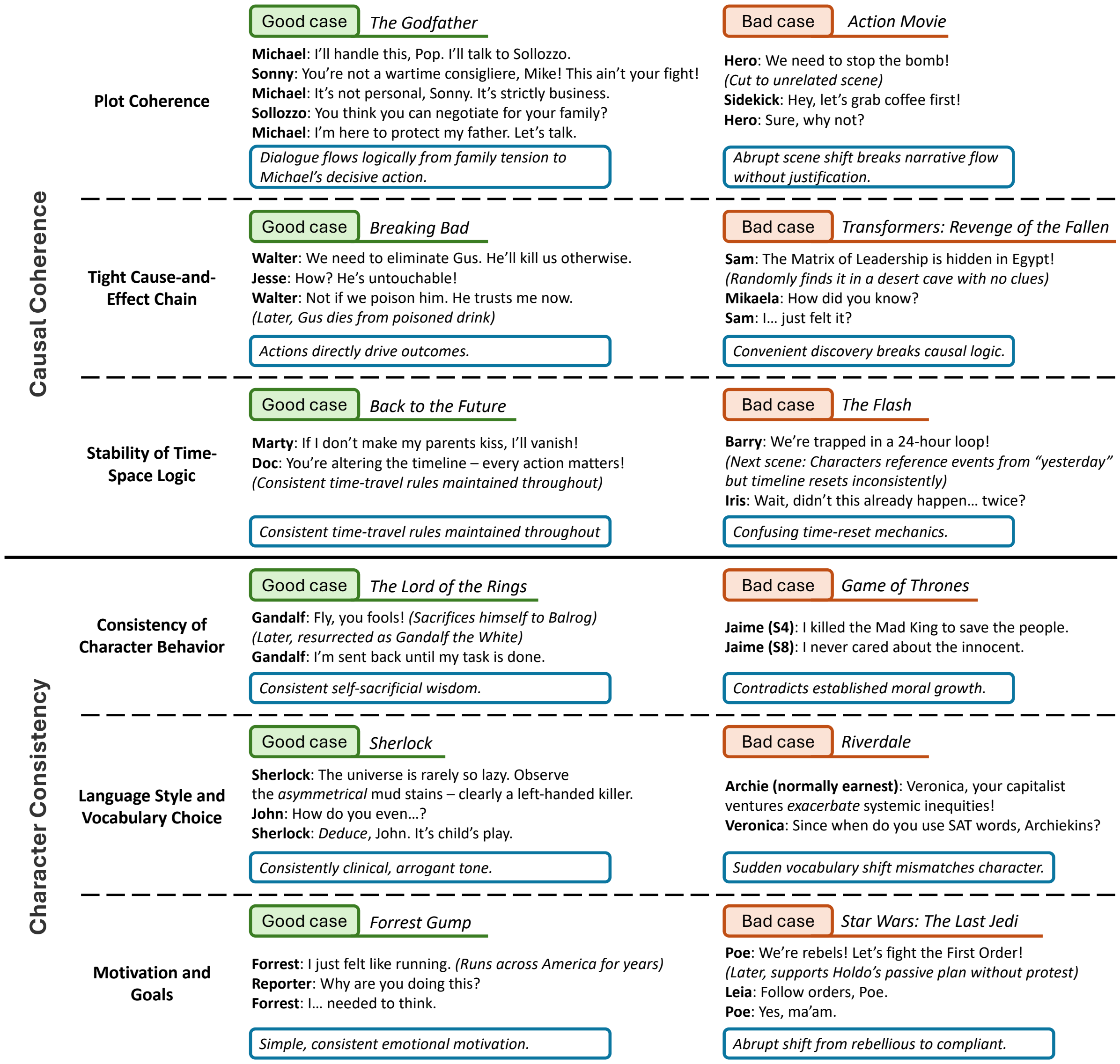


Figure 13: Guidelines for manual role selection

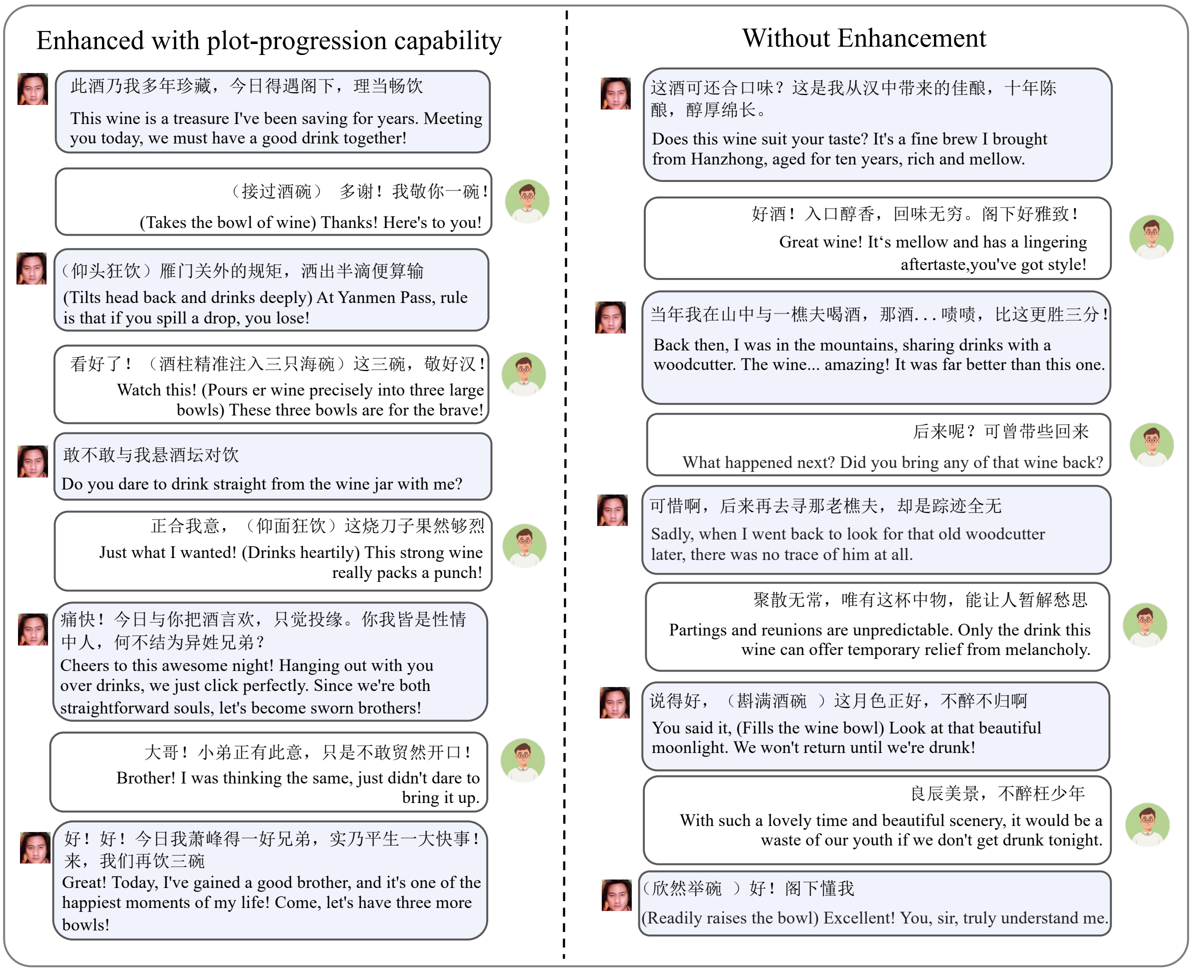


Figure 14: Full texts from comparative examples of conversations with Xiao Feng