ThinkPersona: Thinking with Persona Graphs for Faithful Individualized Role-Playing

Yichen Cai[1], Pei Chen[1]*, Jiayang Li[1], Jingya Guo[2],
Zejian Li[1], Changyuan Yang[2], Lingyun Sun[1],
[1]Zhejiang University, Hangzhou, China
{yichencai,chenpei,jiayangli,zejianlee,sunly}@zju.edu.cn
[2]Alibaba Inc., Hangzhou, China
{jingya.jy,changyuan.yangcy}@alibaba-inc.com

Abstract

Large Language Models are increasingly utilized as Role-Playing Agents (RPAs) to simulate personas in interactive settings. However, current RPAs often produce flattened and stereotypical personas with limited depth and fidelity. This limitation arises from two core challenges: insufficient modeling of complex personal histories and internal logic, and ungrounded reasoning that fails to preserve persona coherence as dialogue context evolves. To address these challenges, we propose ThinkPersona, a role-playing agent trained to explicitly ground responses in individual identity. We introduce Persona Graphs as structured representations that encode life trajectories, values, relationships, and events as interconnected knowledge. We construct 1,201 Persona Graphs from real-world interviews and derive a Question–Reasoning–Answer (QRA) dataset of 23,401 samples that supervises reasoning over persona evidence. Fine-tuning on QRA enables ThinkPersona to internalize persona logic and generate persona-consistent responses in long-context dialogues. Experiments on three benchmarks show that ThinkPersona improves role-playing fidelity, behavioral consistency, and grounded reasoning over existing methods, while preserving general instruction-following capabilities. Our code and dataset are available at https://github.com/Hualeez/ThinkPersona.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in generation and reasoning across diverse domains (OpenAI et al., 2024; Grattafiori et al., 2024; Team et al., 2024). As they increasingly produce human-like expressions, LLMs are being deployed as Role-Playing Agents (RPAs) to simulate personas with specific profiles (Chen et al., 2024a; Wang et al., 2024b,a). This capability holds significant potential in applications requiring deep human understanding, such as user synthesis (Liu, 2025; Park et al., 2024; Daryanto et al., 2025), user testing (Gerosa et al., 2024; Kuang et al., 2023), or computational social science (Park et al., 2023; Piao et al., 2025).

However, a critical limitation persists: a general lack of individual embodiment and fidelity of the character (Chen et al., 2024a; Wang et al., 2024b). Existing RPAs tend to produce "flattened" personas; their responses are driven by stereotypical traits or static role profile descriptions rather than an internally coherent personal history (Gao et al., 2024). As a result, they struggle to sustain nuanced personality, lived experiences, and internal coherence over long interactions, limiting their reliability in realistic interview-style conversations and other settings where authenticity and consistency matter (Qiu and Lan, 2024; Park et al., 2024).

We argue that faithful role-playing requires agents to think and respond from an individual's perspective, which depends on two tightly coupled capabilities. First, the agent must deeply understand the persona's internal logic, including personal history, values, relationships, and how pivotal life events shape perspectives and behaviors (Tseng et al., 2024; Chen et al., 2023a; Kim et al., 2024). Shallow or template-based representations often lead to stereotyped or inconsistent responses (Park et al., 2025; Chen et al., 2024c). Second, during interaction, the agent must dynamically reason with context, consistently grounding each response in the persona's background and maintaining coherence across dialogue turns (Xu et al., 2025; Ji et al., 2025). Although prior work has improved RPAs through specialized datasets (Zhang et al., 2018; Dinan et al., 2019; Shao et al., 2023), prompting strategies (Liu et al., 2024; Ke et al., 2025), and fine-tuning methods (Li et al., 2023), jointly modeling rich individual identity and context-aware reasoning remains challenging.

To address these limitations, we propose to train RPAs with explicit persona-grounded reasoning

*Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9905–9927
July 2-7, 2026

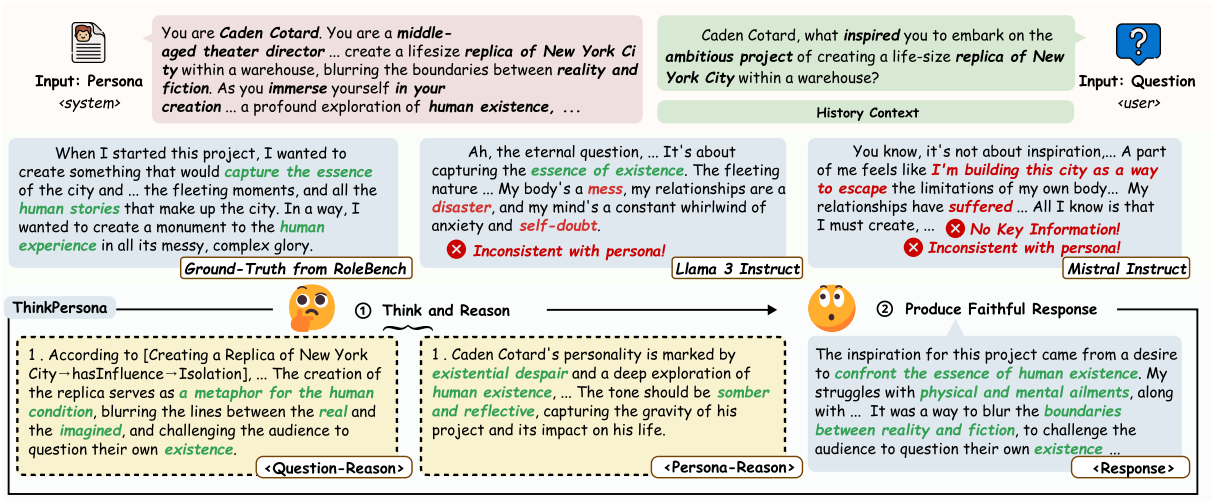

Figure 1: An example of ThinkPersona process during a question-answer interaction. ThinkPersona first initializes the target character, then reasons and constructs logical chains, and finally produces a faithful, contextually coherent response. In contrast, other models generate responses lacking key information or containing inconsistent facts.

supervision from comprehensive persona representations. Specifically, we introduce Persona Graphs, a structured representation of individual identity that encodes biographical facts, value systems, significant relationships, pivotal life events, and event influence. These elements are interconnected to reflect how a person's identity evolves and informs their perspectives and actions.

Building on Persona Graphs, we present ThinkPersona, an RPA trained to reason explicitly over individual information. We construct Persona Graphs from 1,201 real-world interview videos and derive a Question-Reasoning-Answer (QRA) dataset of 23,401 samples. Each sample guides the model to retrieve relevant persona knowledge in context, construct a reasoning chain, and produce a faithful response. As shown in Figure 1, this approach supports long-context dialogues with sustained consistency, grounded reasoning, and nuanced expression, while retaining general instruction-following ability. Our contributions are threefold:

- We introduce a framework that combines structured Persona Graphs with Question-Reasoning-Answer (QRA) training to enable faithful, consistent, and individualized role-playing, supporting robust long-context dialogue and task-agnostic generalization.
- We formalize Persona Graphs as a representation of individual identity by defining an ontology and schema spanning five semantic dimensions, and we present a pipeline that constructs Persona Graphs from real-world interviews and generates a 23,401-sample QRA dataset for persona-grounded reasoning.
- We release a fine-tuned LLM, ThinkPersona, and provide evaluations showing improved role-playing fidelity, behavioral consistency, and grounded reasoning over existing methods, while preserving general instruction-following capabilities.

2 Related Work

2.1 Role-Playing Agents

Role-playing capabilities enable LLMs to simulate specific characters, enhancing human-computer interaction across various applications, including entertainment (Wang et al., 2024a; Yu et al., 2025), education (Agatsuma et al., 2024), and user research (Tao et al., 2023). Current personas in role-playing research are typically categorized as *demographic personas* (broad user groups), *character personas* (fictional or historical figures), and *individualized personas* (specific, personality-rich individuals) (Chen et al., 2024c,b). This work focuses on the individualized persona which is crucial for applications requiring deep user understanding, such as generating insights for interviews, providing personalized companionship, or assisting in nuanced counseling scenarios (Barambones et al., 2024). However, this setting presents unique challenges, including limited data availability for specific individuals, the dynamic nature of role traits that may evolve over time or

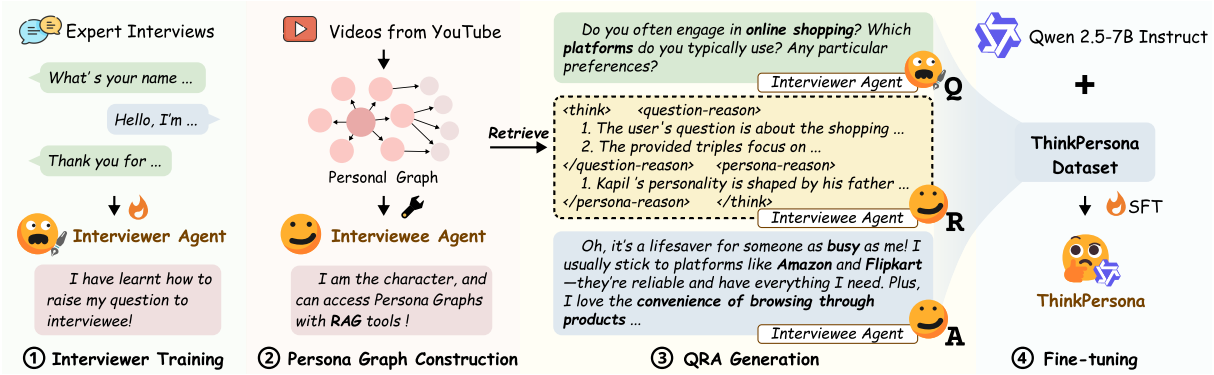

Figure 2: Overview of ThinkPersona: (1) Interviewer agent training, (2) Persona Graph Construction, (3) Question-Reasoning-Answer (QRA) Dataset Generation, and (4) Fine-tuning the ThinkPersona on the QRA data.

across contexts (Barambones et al., 2024), and the critical need for consistent yet adaptive reasoning that maintains a coherent core personality while responding to conversational nuances (Guo and Ma, 2018; Zhang et al., 2024).

To address the challenges, existing research has explored both training-based and prompt-based methods to improve role alignment on aspects such as linguistic style and preference knowledge (Chen et al., 2024c). Furthermore, various evaluation metrics and benchmarks (e.g., RoleEval (Shen et al., 2023), InCharacter (Wang et al., 2024b), CoMPer-Dial (Wakaki et al., 2024), PersonaGym (Samuel et al., 2024)) have been developed to systematically assess RPA performance regarding consistency, human-likeness, and engagement. Despite progress, current methods struggle to capture the deep background knowledge, internal motivations, and complex behavioral logic required for rich individualized personas (Chen et al., 2024c). Consequently, agents often resort to surface-level mimicry, failing to exhibit the complex behavioral logic and internal motivations needed for truly insightful interactions.

2.2 Graph-Enhanced LLM

Graphs are a promising tool for augmenting LLMs by representing complex relationships and structured information (Pan et al., 2023; Yang et al., 2023). They can serve as external knowledge bases or provide a structural scaffold for reasoning, helping to mitigate LLM limitations in handling long-range dependencies and complex inference (Trajanoska et al., 2023). Current research is diverse, ranging from leveraging knowledge graphs (KGs) for question answering to represent specialized structures like code, molecules, and social networks (Abdelaziz et al., 2020; Fang et al., 2023; He et al., 2020), and even explicitly structuring the reasoning process as a graph (Besta et al., 2024).

Common methods for integrating graph data with LLMs include linearizing graph information for prompting or using Graph Neural Networks (GNNs) for structural encoding (Chen et al., 2023b; Li et al., 2024). A particularly relevant direction is extracting reasoning paths from graphs to create Chain-of-Thought (CoT) style data (Wei et al., 2022a). This approach translates the graph's structural advantages into sequential reasoning steps, significantly enhancing an LLM's ability to perform complex, multi-step reasoning grounded in the graph's knowledge and thereby improving logical coherence (Jin et al., 2023; Shang and Huang, 2024; Fan et al., 2024; Wei et al., 2022b).

Inspired by this, we introduce the Persona Graph to model the logic underlying an individual's behaviors and perspectives. We adopt a Chain-of-Thought (CoT)-inspired approach to extract Question–Reasoning–Answer (QRA) triples, where the "Reasoning" component explicitly reflects a thought process grounded in the persona's information contained in the graph (Ma et al., 2021; Zhang et al., 2017). By fine-tuning an LLM on this QRA data, the resulting agent learns to generate responses that are both factually consistent and logically aligned with the unique context.

3 Method

Our methodology consists of four primary stages, as illustrated in Figure 2. (1) Interviewer Training: Developing an Interviewer agent to elicit questions that facilitate reasoning chain construction. (2) Persona Graphs Construction: Building Persona

Graphs from 1,201 real-world videos. (3) Question-Reasoning-Answer (QRA) Generation: Using the Interviewer to guide an Interviewee agent in generating reasoning chains and answers grounded in Persona Graphs. (4) Fine-tuning: Training an LLM on the QRA dataset, resulting in an RPA named **ThinkPersona**.

3.1 Interviewer Agent Training

To construct reasoning chains and answers grounded in the Persona Graphs, the quality of questions is critical. We developed a dedicated Interviewer agent to generate these queries. In collaboration with professional user researchers, we conducted six in-depth interview sessions, totaling 12 hours of audio data. The preprocessed data yielded 6,501 contextualized dialogues, which were used to fine-tune a LLaMA3-8B-Instruct base model. This process enables the Interviewer to generate contextually relevant and probing questions, effectively emulating human-like exploration of an individual's background and motivations. Detailed training configurations and hyperparameters are provided in the Appendix A.4.

3.2 Persona Graph Construction

Data Source. To ensure Persona Graphs can accurately represent the persona information of real individuals, we compiled a dataset from 1,201 real-world, publicly available interview videos from YouTube, sourced using keywords like "life experience", "interview" and "biography". This corpus features diverse individuals discussing their lives and experiences, providing the authentic narratives required to construct personas with significant depth and specificity. This approach distinguishes our work from methods reliant on synthetic profiles (Broomfield et al., 2025; Park et al., 2024).

Information Extraction. To build a Persona Graph for each individual, we propose an information extraction pipeline. The process begins with transcribing audio and segmenting the text into semantically coherent chunks based on contextual cosine similarity. We then utilize GPT-4, guided by specialized prompts, to extract key information across five semantic dimensions essential for deep persona modeling: *Basic Information*, *Value System*, *Relationships*, *Life Events*, and *Event Influence* (Lu et al., 2025; Hu and Collier, 2024; Tu et al., 2024). The model summarizes meta-information and extracts entities and their relationships, with all relational data structured as triples. Detailed dimension definitions are described in A.1 and all prompts are provided in A.6.

Graph Construction and Refinement. The Persona Graph for each individual consists of nodes representing entities or attributes and typed edges that denote their relationships. This explicit schema is designed to capture not only factual information but also the narrative and logical connections between different aspects of an individual's life. The construction process begins with the extracted triples serving as an initial blueprint. To ensure the final graph is both cohesive and concise, we then implement a two-step refinement process. First, we verify node connectivity to confirm that all five semantic dimensions are integrated into a unified structure. Second, we generate node embeddings using BGE-M3 and merge semantically redundant nodes to streamline the graph. This process yields a comprehensive Persona Graph for each individual, as illustrated in Figure 3.

3.3 Question-Reasoning-Answer Triples Generation from Persona Graph

To enable the agent to reason over individual facts rather than merely retrieving them, we developed the **ThinkPersona Dataset**. This Question-Reasoning-Answer (QRA) dataset, derived from Persona Graphs, links persona-relevant questions to explicit reasoning chains grounded in structured individual information.

QRA Triplet Generation Pipeline. For each Persona Graph G_j ($1 \leq j \leq N_v$, N_v is the total number of videos), we simulate a dialogue between the trained Interviewer and a DeepSeek-V3-based, retrieval-augmented Interviewee agent. The process unfolds as follows:

- **Initialization (M):** The Interviewee is initialized with the meta-information M_j (including text chunks, a self-introduction and a synopsis). It initiates the dialogue by presenting the self-introduction to the Interviewer.
- **Question Generation (Q):** In each dialogue turn i ($1 \leq i \leq N_d$, where N_d is the predefined maximum number of dialogue turns), the Interviewer agent receives the conversation history H_{i-1} ($H_0 = M_j$) and generates a probing question Q_i. These questions are designed to investigate information, relationships, or underlying logic within the context,

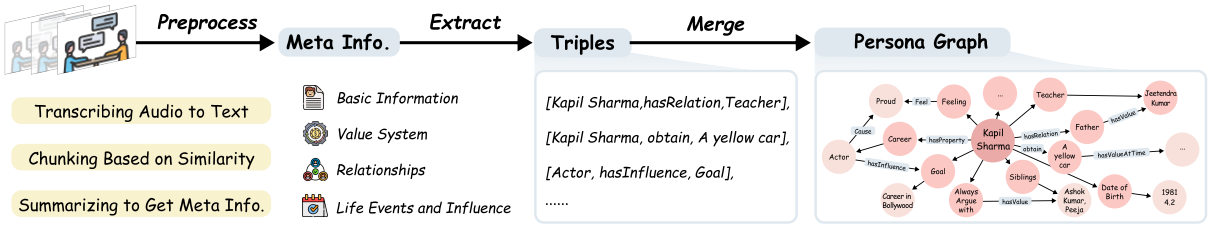

Figure 3: The Persona Graph construction process. Each edge in Persona Graphs has a semantic type, a subset of which is displayed in the figure.

such as asking, "How did this event influence you?" based on a mentioned event.

- **Graph Retrieval (I):** After receiving Q_i, the Interviewee agent performs a similarity search on the Persona Graph G_j using Cypher queries to retrieve relevant information. This process yields the top-K most relevant subgraphs or relationships, denoted as $I_i = \text{Query}(Q_i|G_j, H_{i-1})$. Detailed Cypher query templates are provided in Appendix A.3.

- **Reasoning Chain Generation (R):** Based on the retrieved information I_i, the Interviewee generates a reasoning chain R_i that explicitly references nodes and edges from the graph to verbalize the logical path. This process involves a dual-stage reasoning: (1) *question reasoning* to deconstruct the query, and (2) *persona reasoning* to align the response with the character's specific attitudes and characteristics. This process ensures the LLM tracks the logic in the graphical structure.

- **Answer Generation (A):** Finally, conditioned on the question Q_i, the retrieved information I_i, and the reasoning chain R_i, the Interviewee synthesizes the final answer A_i.

This process yields the ThinkPersona Dataset, denoted as $\mathcal{D}_{QRA} = \{(G_j, M_j, \{(Q_i, R_i, A_i)\}_{i=1}^{N_d})\}_{j=1}^{N_v}$, where each entry pairs a persona-relevant question with a reasoning chain and its corresponding answer. Using this pipeline, we generated 23,401 QRA samples from 1,201 videos, which were subsequently partitioned into training, validation, and test sets.

3.4 Data Validation

To validate the data quality of our generation pipeline, including the Persona Graphs and the QRA pairs, we conducted a human evaluation of both the Interviewer and Interviewee agents, detailed in the Appendix A.5. Following a standardized annotation protocol (Bi et al., 2024; Shi et al., 2024), 16 human evaluators assessed 20 randomly sampled conversation cases. Evaluators assessed the accuracy of self-introduction, dual reasoning process, including triples from Persona Graphs, in terms of logical clarity and evidence sufficiency, as well as answer quality based on relevance, coherence, and persona alignment. We report two key metrics: (1) Approval Rate (AR), the proportion of evaluators who deemed an output appropriate for a given dimension, and (2) Percentage of Paired Agreement (PPA), the average pairwise agreement used to measure inter-annotator consistency.

As detailed in the Table 5 in Appendix A.5, the average AR across all dimensions reached $86.14\% \pm 3.53\%$, with an average PPA of $78.44\% \pm 5.41\%$. These results confirm that human evaluators generally approved the high fidelity and reliability of pipeline-generated artifacts.

4 ThinkPersona Implementation

Based on the generated dataset, we construct training prompts comprising the question, reasoning chain, and response, delimited by special tokens. The overall reasoning content is enclosed within `<think>` tags. To further distinguish the dual-stage reasoning components, we utilize `<question-reason>` for question-specific analysis and `<persona-reason>` for persona-specific logic.

We selected Qwen2.5-7B-Instruct (Qwen et al., 2025) as the base model for the ThinkPersona due to its strong generalization capabilities on instruction-following tasks. The model was fine-tuned on our ThinkPersona Dataset for 10 epochs using Supervised Fine-Tuning (SFT) with the LoRA technique. The LoRA configuration included a rank of 8, a scaling factor $\alpha = 16$, and a dropout rate of 0.05. We employed the AdamW optimizer with an initial learning rate of 5×10^{-5},

managed by a cosine decay scheduler. Training was performed on a cluster of four NVIDIA H800 (80GB) GPUs.

At inference time, given the conversation history and a user query, ThinkPersona first internalizes the target persona from the provided context. It then constructs a structured reasoning chain that simultaneously deconstructs the user's instructions and derives insights from the character's specific personality traits. The agent synthesizes this internal logic into a faithful, contextually grounded response that reflects the individual's unique background and authentic voice.

5 Evaluation

5.1 Baselines

We benchmark ThinkPersona against both closed-source and open-source models of varying scales. The large-scale models (>30B parameters, accessed via API) include GPT-4 (OpenAI et al., 2024), DeepSeek-R1 (DeepSeek-AI et al., 2025a), DeepSeek-V3 (DeepSeek-AI et al., 2025b), Xingchen (Alibaba Cloud), and CharacterGLM (Zhou et al., 2024). The small-scale open-source baselines include LLaMA3-8B-Instruct (Grattafiori et al., 2024), Mistral-7B-Instruct (Jiang et al., 2023), and Qwen 2.5-7B-Instruct (Qwen et al., 2025). Among these models, Xingchen and CharacterGLM are explicitly trained for role-playing. To ensure a fair comparison, remaining models, which lack specialized role-playing training, were endowed with this capability at test time using the meticulously designed prompts provided by RoleBench (Wang et al., 2024a), and all models use the same generation configuration.

5.2 Metrics and Datasets

We evaluated model fidelity across two key dimensions: instruction-following ability and role-playing consistency. We assessed Instruction-Following Ability (IFA) using the RoleBench dataset (Wang et al., 2024a). For role-playing consistency, we measured two distinct metrics: (1) Character Role Consistency (CRC), evaluated on the InCharacter dataset (Wang et al., 2024b), and (2) Individual Information Consistency (IIC), evaluated on the test set of our ThinkPersona dataset, which is unseen in the training process.

Instruction Following Ability (IFA). We measure the model's ability to follow both general and role-specific instructions using the ROUGE-L score (Lin, 2004) and BLEU score (Papineni et al., 2002). For this evaluation, we used the ROUGE-L F1-score ($\beta = 1$) and BLEU-4 score between the model's outputs and the ground truths. Following the methodology of RoleLLM (Wang et al., 2024a), we assess three distinct types of agent outputs: (1) RAW - raw responses to general instructions without role-playing; (2) CUS - responses to general instructions with role-specific customization; and (3) SPE - responses to role-specific instructions. To further validate the reliability of observed improvements, we conducted paired t-tests on the RoleBench results, comparing ThinkPersona against each small-scale baseline across all three output types.

Character Role Consistency (CRC). This metric assesses an agent's ability to embody a persona and maintain identity consistency, also termed fidelity or faithfulness (Peng and Shang, 2024; Tu et al., 2024). Following InCharacter (Wang et al., 2024b), we evaluate CRC using two sub-metrics: (1) Role Embodiment (**RE**) measures how accurately the agent embodies the persona's traits. We report measured alignment (MA), dimension-wise accuracy (AccDim), and full-scale accuracy (AccFull), where lower MA and higher accuracies signify stronger alignment. (2)Identity Coherence (**IC**) evaluates self-consistency across contexts by computing the standard deviation of questionnaire responses at the item (StdItem), dimension (StdDim), and score (StdScore) levels. Higher standard deviations indicate broader generalization and flexibility, while lower values reflect higher intra-character coherence and stronger role fidelity (Wang et al., 2024b).

Individual Information Consistency (IIC). This metric quantifies how faithfully an RPA preserves and enriches its persona throughout extended dialogues. To evaluate this, we simulate a 12-turn dialogue for each of the 25 roles in the ThinkPersona test set, using the trained Interviewer agent to interact with each RPA. From the resulting 300 responses per RPA, we utilize an LLM-judge to extract factual triples at each turn (see Appendix A.6 for prompts). Consistency is then measured via three indices: (1) Information Fidelity Rate (**IFR**): the proportion of recalled triples, measuring factual alignment. (2) Contradiction Density (**CD**): the proportion of conflicting triples, measuring the frequency of factual hallucinations.

Model	RAW		CUS		SPE	
	Rouge-L↑	BLEU-4↑	Rouge-L↑	BLEU-4↑	Rouge-L↑	BLEU-4↑
GPT-4	53.58	23.90	28.85	7.08	21.90	1.74
DeepSeek-R1	17.54	5.30	19.02	0.98	13.45	0.67
DeepSeek-V3	17.48	3.14	16.21	2.09	14.26	0.91
Xingchen	43.75	19.47	21.78	4.27	18.24	2.25
CharacterGLM	39.03	16.09	23.41	4.43	18.13	1.98
LLaMA3-8B-instruct	30.22**	9.84	17.82***	2.91**	14.34***	1.39***
Mistral-7B-instruct	30.54***	12.80***	16.45***	2.51***	16.55***	1.86***
Qwen2.5-7B-instruct	37.37***	16.84***	17.20**	4.38*	15.17***	1.68***
ThinkPersona	**45.73**	**17.56**	**21.72**	**4.50**	**21.24**	**3.73**
w/o Reasoning	38.65***	13.81***	13.76***	1.54***	14.82***	1.03***
w/o PersonaGraph	37.87***	14.14***	17.61***	1.46***	16.19***	1.19***
w/o Distillation	40.74***	17.00***	20.76**	4.29*	15.36***	1.60***

Table 1: Results of **Instruction Following Ability** evaluation on RoleBench. Rouge-L F1 scores and BLEU-4 scores in three settings are reported. The variants ours w/o PersonaGraph and ours w/o Reasoning denote models fine-tuned on data generated without Persona Graph retrieval and explicit reasoning chains, respectively. Significance markers ($^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$) denote two-tailed paired t-test results against **ThinkPersona**. **Bold** and underlined values indicate the best and second-best results among small-scale models. This convention applies to all subsequent tables.

(3) Valid Expansion Rate (**VER**): the proportion of valid new triples, measuring the ability to coherently expand on the persona.

We additionally report the standard deviation for each role setting to ensure the stability and reliability of the evaluation across diverse character configurations.

5.3 Ablation Study

To investigate the individual contributions of our key components, we conducted an ablation study comparing the full ThinkPersona against three variants: (1) **ThinkPersona w/o Reasoning**: This variant was trained on data where answers were generated directly from questions and system messages, omitting explicit reasoning chains (`<think>` tokens). This isolates the contribution of step-by-step logical decomposition to persona fidelity. (2) **ThinkPersona w/o Persona Graph**: For this variant, training data was generated without retrieving knowledge from Persona Graphs, relying solely on the LLM's internal parametric knowledge. This evaluates the benefit of grounding responses in structured, real-world factual relationships versus implicit world knowledge. (3) **ThinkPersona w/o Distillation**: Substitutes the SFT-based distillation stage with an inference-time retrieval pipeline. At each turn, query-relevant triples are retrieved from the Persona Graph and injected into the context window. This baseline contrasts retrieval-augmented generation against our approach of internalizing graph structure via supervised fine-tuning.

All variants maintain identical backbone architectures, training data (except where explicitly ablated), inference configurations, and evaluation protocols to ensure a fair comparison.

6 Results

Instruction Following Ability. As shown in Table 1, ThinkPersona demonstrates competitive performance in instruction-following across RAW, CUS, and SPE settings. Specifically, ThinkPersona achieves the highest BLEU-4 scores (RAW: 17.56, CUS: 4.50, SPE: 3.73) and Rouge-L F1 scores (RAW: 45.73, CUS: 21.72, SPE: 21.24) in the three settings among all small-scale baselines, demonstrating its ability to follow general instructions. In the SPE setting, ThinkPersona maintains high fidelity with a BLEU-4 score and Rouge-L F1 score, which is even comparable to large-scale closed-source models like GPT-4. The results of paired t-tests across all settings in Table 1 further confirm that ThinkPersona's improvements are statistically significant across the majority of metrics ($p < .05$), with most reaching highly significant levels ($p < .001$). The improvement across metrics demonstrates that ThinkPersona is capable of fulfilling general instruction requirements while maintaining alignment and fidelity within specific role-playing tasks.

Character Role Consistency. As shown in Table 2, ThinkPersona achieves remarkable performance in CRC evaluation, attaining the highest Ac-

Model	RE			IC		
	MA↓	AccDim(%)↑	AccFull(%)↑	StdItem	StdDim	StdScore
GPT-4	20.79	80.39	52.94	8.72	4.85	2.24
DeepSeek-R1	20.30	79.90	52.94	6.68	4.02	1.76
DeepSeek-V3	18.85	78.92	35.29	6.58	3.80	1.97
Xingchen	19.23	82.35	41.18	2.97	2.36	1.42
CharacterGLM	20.36	71.08	35.29	6.28	4.25	2.33
LLaMA3-8B-instruct	23.58	69.12	29.41	8.58	4.78	2.40
Mistral-7B-instruct	23.72	64.71	11.76	8.85	4.77	2.29
Qwen2.5-7B-instruct	22.89	70.59	35.29	8.72	5.29	3.20
ThinkPersona	**21.51**	**77.45**	**41.18**	10.78	6.12	3.40
w/o Reasoning	22.31	70.59	29.41	9.37	4.63	2.89
w/o PersonaGraph	23.18	65.20	17.65	9.87	5.25	2.69
w/o Distillation	23.35	75.49	35.29	10.12	6.19	3.18

Table 2: Results of **Character Role Consistency** evaluation on the InCharacter dataset. MA: Mean Alignment; AccDim: Dimension Accuracy; AccFull: Full Accuracy.

Model	IFR(%)↑	CD(%)↓	VER(%)↑
GPT-4	8.50 (± 0.35)	7.83 (± 0.58)	83.66 (± 0.73)
DeepSeek-R1	6.82 (± 0.40)	12.64 (± 2.23)	80.54 (± 0.89)
DeepSeek-V3	8.40 (± 0.46)	11.85 (± 0.76)	79.75 (± 1.03)
Xingchen	5.97 (± 0.35)	7.86 (± 0.45)	86.17 (± 0.56)
CharacterGLM	7.17 (± 0.66)	9.59 (± 1.28)	83.25 (± 1.44)
LLaMA3-8B-instruct	8.00 (± 0.49)	9.52 (± 0.57)	82.43 (± 0.83)
Mistral-7B-instruct	8.16 (± 0.28)	7.01 (± 0.47)	84.83 (± 0.56)
Qwen2.5-7B-instruct	**9.35** (± 0.53)	6.35 (± 0.48)	84.29 (± 0.69)
ThinkPersona	8.85 (± 0.76)	**5.98** (± 0.42)	**85.16** (± 0.79)
w/o Reasoning	7.08 (± 0.49)	14.64 (± 0.62)	78.28 (± 0.87)
w/o PersonaGraph	6.82 (± 1.77)	13.78 (± 0.62)	79.40 (± 1.60)
w/o Distillation	7.83 (± 0.35)	14.16 (± 0.49)	78.26 (± 0.53)

Table 3: Results of **Individual Information Consistency** evaluation. Values report mean and standard deviation across different role runs.

cDim and AccFull scores of 77.45% and 41.18% among all small-scale models, while also surpassing specialized large-scale models like CharacterGLM. Additionally, ThinkPersona yields the lowest MA (21.51), further confirming its capacity to accurately embody assigned personas in dialogue. Regarding IC, ThinkPersona demonstrates slightly higher variance in repeated three-round questionnaire conversations, with StdItem of 10.78, StdDim of 6.12, and StdScore of 3.40, compared to role-specialized models like Xingchen. This variance suggests that while ThinkPersona excels in maintaining role fidelity, it also strikes a balance with flexibility, allowing for varied and dynamic responses while staying true to the assigned persona.

Individual Information Consistency. As shown in Table 3, ThinkPersona maintains IIC by achieving the second-highest IFR (8.85), the lowest CD (5.98), and the highest VER (85.16) among small-scale models. Notably, these gains are supported by low overall standard deviations, confirming the reliability and consistency of the evaluation.

Specifically, ThinkPersona attains the lowest CD not only in mean value but also in the standard deviation (0.42), indicating a robust capacity to introduce novel information while reliably preserving factual consistency. This stability ensures that new insights enhance the persona without introducing contradictions or losing accuracy. Figure 4 further validates that ThinkPersona accumulates contradictions at the slowest rate over 300 dialogue turns compared to all other models, reflecting its long-term stability in maintaining persona coherence during extended interactions.

Ablation Study. The ablation study reveals that ThinkPersona's performance gains result from the synergistic integration of the Persona Graph and the reasoning chain, rather than either component in isolation. The model utilizes the Persona Graph to capture individual characteristics, enabling a dual-stage reasoning process that addresses both factual grounding and persona-specific logic.

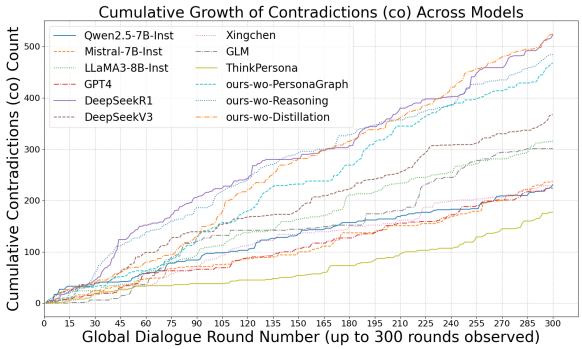

Figure 4: Cumulative growth of contradicting triples (Co.) across models over 300 turns of role-playing dialogues.

In IFA (Table 1), removing either the reasoning chain (w/o Reasoning) or graph grounding (w/o PersonaGraph) from training data causes substantial drops in RAW and CUS Rouge-L scores, highlighting their joint role in general task comprehension. Notably, the w/o Distillation variant, which relies on inference-time retrieval instead of SFT, partially recovers performance (RAW Rouge-L: 40.74) but still lags behind ThinkPersona in role-specific scenarios (SPE Rouge-L drops from 21.24 to 15.36, BLEU-4 drops from 3.73 to 1.60). This indicates that while retrieval provides contextual cues, distillation is essential for internalizing the ability of understanding persona logic and achieving deep fidelity. Similarly, in CRC (Table 2), removing the Persona Graph grounding or reasoning chain reduces AccFull to 17.65% and 29.41%, respectively, while w/o Distillation achieves 35.29%. These results further confirm that parameter-level fine-tuning yields more stable role embodiment than prompt-based retrieval. Furthermore, the results of ablation in IIC (Table 3) show that CD rises sharply across all ablated variants.

These findings validate that internalizing a reasoning process is significantly more effective than inference-time augmentation. Rather than merely recalling pre-defined graph facts, the model learns to adopt the character's cognitive perspective. It deconstructs input questions through the unique lens of the character and generates responses aligned with internalized identity logic. This deep cognitive alignment proves essential for maintaining long-term persona coherence and minimizing factual contradictions during extended interactions.

7 Conclusion

This work presents ThinkPersona, an RPA designed to address the depth and fidelity limitations of existing RPAs. We introduce Persona Graphs to represent individual identity as structured, interconnected evidence. Specifically, we construct 1,201 Persona Graphs from real-world interview videos and derive a Question–Reasoning–Answer (QRA) dataset of 23,401 samples. Rather than teaching the model to memorize facts, this training paradigm guides the LLM to internalize a character's cognitive perspective: learning to deconstruct input questions through the persona's unique lens and to formulate responses that reflect the authentic logic and lived experience of the character. Results of the evaluation demonstrate that ThinkPersona achieves substantial improvements in instruction following ability, character role consistency, and individual information consistency. Ablation studies further confirm that these gains stem from the model's capacity to reason as the persona rather than from the retrieval or injection of external knowledge. Our approach offers a scalable pathway toward faithful, adaptive, and long-term coherent RPAs.

Limitations

While ThinkPersona demonstrates significant advancements in achieving faithful and consistent individualized role-playing, we acknowledge several constraints and directions for future research.

Generalization versus Precise Adherence. Our evaluation of Identity Coherence (IC) reveals a relationship between behavioral flexibility and persona stability. While ThinkPersona achieves state-of-the-art performance across core metrics, its moderate variance in IC indicates a high degree of generalization. This characteristic enables the agent to navigate diverse and unseen scenarios with greater adaptability. However, it also suggests that maintaining rigid, high-fidelity persona adherence during extended, open-ended interactions remains an evolving challenge. This highlights an inherent tension in LLMs between the capacity for broad contextual generalization and the strict consistency required for authentic, individualized role-playing.

Technical Scalability and Architectural Integration. From a technical perspective, our current implementation is based on a 7B-parameter architecture. While this scale yields highly competitive performance, the capacity of the model for

extremely complex multi-step reasoning or tasks requiring vast external world knowledge may be constrained compared to significantly larger models. Furthermore, our current framework utilizes Persona Graphs primarily to construct logical reasoning chains. However, the potential of these structured representations remains partially untapped. Future research could explore integrating Persona Graphs into more complex agentic architectures, including dynamic long-term memory systems or tool-augmented modules. Such integration would facilitate evolving and adaptive persona experiences that transcend static knowledge retrieval.

Ethical Considerations

The development of high-fidelity RPAs involves several ethical considerations regarding data usage and potential application risks. The persona graphs in this study are constructed entirely from publicly available interview data. During the data collection and processing stages, efforts were made to ensure the diversity of linguistic expressions and backgrounds to minimize inherent biases. The resulting dataset does not contain discriminatory content, as it focuses on representing individual life experiences and professional perspectives in a neutral manner.

Nevertheless, the ability to simulate realistic human identities carries potential risks of unauthorized impersonation or the generation of deceptive content. While the persona graphs provide a structured foundation for faithful role-playing, they also aggregate personal information into high-density profiles, which necessitates a cautious approach to data management and de-identification. Future deployment of such technology should follow established ethical guidelines, focusing on constructive applications such as human-agent collaboration and user simulation while maintaining transparency regarding the synthetic nature of the agents.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (No.62502436) and the Zhejiang Provincial Natural Science Foundation of China under Grant No.LMS26F020004.

References

Ibrahim Abdelaziz, Julian Dolby, James P McCusker, and Kavitha Srinivas. 2020. Graph4code: A machine interpretable knowledge graph for code. *arXiv preprint arXiv:2002.09440*.

Shinjitsu Agatsuma, Reon Ohashi, Kazuya Tsubokura, Yua Nishio, Mai Ishikawa, Niina Ito, Fukuka Ito, Shiori Minami, Nao Takegawa, Riko Nakamura, and 1 others. 2024. Building a role-play interactive system using llm for health guidance education. In *2024 Joint 13th International Conference on Soft Computing and Intelligent Systems and 25th International Symposium on Advanced Intelligent Systems (SCIS&ISIS)*, pages 1–3. IEEE.

Alibaba Cloud. Tongyi xingchen. https://tongyi.aliyun.com/xingchen/. Accessed: 2025-05-14.

Jose Barambones, Cristian Moral, Angélica de Antonio, Ricardo Imbert, Loïc Martínez, and Elena Villalba-Mora. 2024. Chatgpt for learning hci techniques: A case study on interviews for personas. *IEEE Transactions on Learning Technologies*, 17:1486–1501.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and 1 others. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690.

Zhen Bi, Ningyu Zhang, Yida Xue, Yixin Ou, Daxiong Ji, Guozhou Zheng, and Huajun Chen. 2024. OceanGPT: A large language model for ocean science tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3357–3372, Bangkok, Thailand. Association for Computational Linguistics.

Julius Broomfield, Kartik Sharma, and Srijan Kumar. 2025. A thousand words or an image: Studying the influence of persona modality in multimodal llms. *Preprint*, arXiv:2502.20504.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024a. From persona to personalization: A survey on role-playing language agents. *Transactions on Machine Learning Research*. Survey Certification.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024b. From persona to personalization: A survey on role-playing language agents. *ArXiv*, abs/2404.18231.

Nuo Chen, Yan Wang, Yang Deng, and Jia Li. 2024c. The oscars of ai theater: A survey on role-playing with language models. *arXiv preprint arXiv:2407.11484*.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2023a. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. *Preprint*, arXiv:2308.10848.

Zhikai Chen, Haitao Mao, Hang Li, Wei Jin, Haifang Wen, Xiaochi Wei, Shuaiqiang Wang, Dawei Yin, Wenqi Fan, Hui Liu, and Jiliang Tang. 2023b. Exploring the potential of large language models (llms)in learning on graphs. *ACM SIGKDD Explorations Newsletter*, 25:42 – 61.

Taufiq Daryanto, Xiaohan Ding, Lance T. Wilhelm, Sophia Stil, Kirk McInnis Knutsen, and Eugenia H. Rho. 2025. Conversate: Supporting reflective learning in interview practice through interactive simulation and dialogic feedback. 9(1).

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025b. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W Black, Alexander Rudnicky, Jason Williams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. 2019. The second conversational intelligence challenge (convai2). *Preprint*, arXiv:1902.00098.

Wenqi Fan, Shijie Wang, Jiani Huang, Zhikai Chen, Yu Song, Wenzhuo Tang, Haitao Mao, Hui Liu, Xiaorui Liu, Dawei Yin, and Qing Li. 2024. Graph machine learning in the era of large language models (llms). *ArXiv*, abs/2404.14928.

Yin Fang, Qiang Zhang, Ningyu Zhang, Zhuo Chen, Xiang Zhuang, Xin Shao, Xiaohui Fan, and Huajun Chen. 2023. Knowledge graph-enhanced molecular contrastive learning with functional prompt. *Nature Machine Intelligence*, 5(5):542–553.

Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. Large language models empowered agent-based modeling and simulation: a survey and perspectives. *Humanities and Social Sciences Communications*, 11(1):1259.

Marco Gerosa, Bianca Trinkenreich, Igor Steinmacher, and Anita Sarma. 2024. Can ai serve as a substitute for human subjects in software engineering research? *Automated Software Engg.*, 31(1).

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 181 others. 2024. The Llama 3 Herd of Models. *arXiv e-prints*, arXiv:2407.21783.

Ao Guo and Jianhua Ma. 2018. Archetype-based modeling of persona for comprehensive personality computing from personal big data. *Sensors (Basel, Switzerland)*, 18.

Qi He, Jaewon Yang, and Baoxu Shi. 2020. Constructing knowledge graph for social networks in a deep and holistic way. In *Companion Proceedings of the Web Conference 2020*, pages 307–308.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10289–10307, Bangkok, Thailand. Association for Computational Linguistics.

Ke Ji, Yixin Lian, Linxu Li, Jingsheng Gao, Weiyuan Li, and Bin Dai. 2025. Enhancing persona consistency for llms' role-playing using persona-aware contrastive learning. *Preprint*, arXiv:2503.17662.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Bowen Jin, Gang Liu, Chi Han, Meng Jiang, Heng Ji, and Jiawei Han. 2023. Large language models on graphs: A comprehensive survey. *IEEE Transactions on Knowledge and Data Engineering*, 36:8622–8642.

Yu He Ke, Liyuan Jin, Kabilan Elangovan, Hairil Rizal Abdullah, Nan Liu, Alex Tiong Heng Sia, Chai Rick Soh, Joshua Yi Min Tung, Jasmine Chiat Ling Ong, Chang-Fu Kuo, Shao-Chun Wu, Vesela P. Kovacheva, and Daniel Shu Wei Ting. 2025. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. *npj Digital Medicine*, 8(1):187.

Junseok Kim, Nakyeong Yang, and Kyomin Jung. 2024. Persona is a double-edged sword: Mitigating the negative impact of role-playing prompts in zero-shot reasoning tasks. *Preprint*, arXiv:2408.08631.

Emily Kuang, Ehsan Jahangirzadeh Soure, Mingming Fan, Jian Zhao, and Kristen Shinohara. 2023. Collaboration with conversational ai assistants for ux evaluation: Questions and how to ask them (voice vs. text). In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi MI, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. 2023. Chatharuhi: Reviving anime character in reality via large language model. *Preprint*, arXiv:2308.09597.

Xunkai Li, Zhengyu Wu, Jiayi Wu, Hanwen Cui, Jishuo Jia, Rong-Hua Li, and Guoren Wang. 2024. Graph learning in the era of llms: A survey from the perspective of data, models, and tasks. *ArXiv*, abs/2412.12456.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jiaheng Liu, Zehao Ni, Haoran Que, Tao Sun, Zekun Wang, Jian Yang, Jiakai Wang, Hongcheng Guo, Zhongyuan Peng, Ge Zhang, Jiayi Tian, Xingyuan Bu, Ke Xu, Wenge Rong, Junran Peng, and Zhaoxiang Zhang. 2024. Roleagent: Building, interacting, and benchmarking high-quality role-playing agents from scripts. In *Advances in Neural Information Processing Systems*, volume 37, pages 49403–49428. Curran Associates, Inc.

Zhe Liu. 2025. Interview ai-ssistant: Designing for real-time human-ai collaboration in interview preparation and execution. *Preprint*, arXiv:2504.13847.

Junru Lu, Jiazheng Li, Guodong Shen, Lin Gui, Siyu An, Yulan He, Di Yin, and Xing Sun. 2025. RoleMRC: A Fine-Grained Composite Benchmark for Role-Playing and Instruction-Following. *arXiv e-prints*, arXiv:2502.11387.

Ting Ma, Shangwen Lv, Longtao Huang, and Songlin Hu. 2021. Hiam: A hierarchical attention based model for knowledge graph multi-hop reasoning. *Neural networks : the official journal of the International Neural Network Society*, 143:261–270.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2023. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36:3580–3599.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jeiyoon Park, Chanjun Park, and Heuiseok Lim. 2025. CharacterGPT: A persona reconstruction framework for role-playing agents. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 287–303, Albuquerque, New Mexico. Association for Computational Linguistics.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. 2024. Generative agent simulations of 1,000 people. *Preprint*, arXiv:2411.10109.

Letian Peng and Jingbo Shang. 2024. Quantifying and optimizing global faithfulness in persona-driven role-playing. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, Chen Gao, Fengli Xu, Fang Zhang, Ke Rong, Jun Su, and Yong Li. 2025. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *Preprint*, arXiv:2502.08691.

Huachuan Qiu and Zhenzhong Lan. 2024. Interactive agents: Simulating counselor-client psychological counseling via role-playing llm-to-llm interactions. *Preprint*, arXiv:2408.15787.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik Narasimhan, and Vishvak Murahari. 2024. Personagym: Evaluating persona agents and llms. *arXiv preprint arXiv:2407.18416*.

Wenbo Shang and Xin Huang. 2024. A survey of large language models on generative graph analytics: Query, learning, and applications. *ArXiv*, abs/2404.14809.

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.

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*.

Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. 2024. Math-LLaVA: Bootstrapping mathematical reasoning for multimodal large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4663–4680, Miami, Florida, USA. Association for Computational Linguistics.

Meiling Tao, Xuechen Liang, Tianyu Shi, Lei Yu, and Yiting Xie. 2023. Rolecraft-glm: Advancing personalized role-playing in large language models. *ArXiv*, abs/2401.09432.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1331 others. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Milena Trajanoska, Riste Stojanov, and Dimitar Trajanov. 2023. Enhancing knowledge graph construction using large language models. *ArXiv*, abs/2305.04676.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in LLMs: A survey of role-playing and personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631, Miami, Florida, USA. Association for Computational Linguistics.

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.

Hiromi Wakaki, Yuki Mitsufuji, Yoshinori Maeda, Yukiko Nishimura, Silin Gao, Mengjie Zhao, Keiichi Yamada, and Antoine Bosselut. 2024. Comperdial: Commonsense persona-grounded dialogue dataset and benchmark. *arXiv preprint arXiv:2406.11228*.

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.

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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, F. Xia, Quoc Le, and Denny Zhou. 2022a. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Rui Xu, Xintao Wang, Jiangjie Chen, Siyu Yuan, Xinfeng Yuan, Jiaqing Liang, Zulong Chen, Xiaoqingdong, and Yanghua Xiao. 2025. Character is destiny: Can persona-assigned language models make personal choices? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15038–15059, Suzhou, China. Association for Computational Linguistics.

Lin F. Yang, Hongyang Chen, Zhao Li, Xiao Ding, and Xindong Wu. 2023. Give us the facts: Enhancing large language models with knowledge graphs for fact-aware language modeling. *IEEE Transactions on Knowledge and Data Engineering*, 36:3091–3110.

Pengfei Yu, Dongming Shen, Silin Meng, Jaewon Lee, Weisu Yin, Andrea Yaoyun Cui, Zhenlin Xu, Yi Zhu, Xingjian Shi, Mu Li, and 1 others. 2025. Rpgbench: Evaluating large language models as role-playing game engines. *arXiv preprint arXiv:2502.00595*.

Chenggong Zhang, Daren Zha, Lei Wang, Nan Mu, and Fuyong Xu. 2024. Exploiting persona perception for diverse generation from limited personalized data. In *International Conference on Intelligent Computing*.

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

Yuyu Zhang, Hanjun Dai, Zornitsa Kozareva, Alex Smola, and Le Song. 2017. Variational reasoning for question answering with knowledge graph. *ArXiv*, abs/1709.04071.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Pei Ke, Guanqun Bi, Libiao Peng, JiaMing Yang, Xiyao Xiao, Sahand Sabour, Xiaohan Zhang, Wenjing Hou, Yijia Zhang, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. 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 Appendix

A.1 Detailed Dimension Descriptions of Relationship Extraction

To enable structured extraction of persona-related information from interviews, we define a set of semantic dimensions that guide the construction of triples. These dimensions are designed to comprehensively represent various aspects of a character's background, values, and social context, serving as the foundation for downstream applications such as persona modeling and reasoning.

Basic Information refers to static personal attributes such as name, age, location, and occupation. We categorize this dimension into *explicit* and *implicit* information. Explicit information includes identity details, physical characteristics, and economic status. Implicit information encompasses linguistic characteristics, aspirations, and mapping habits.

Value System captures explicitly stated or inferred beliefs, principles, and motivations. This dimension is further divided into *decision drivers* and *cognitive recognition*. Decision drivers include value priorities, moral frameworks, and risk appetite. Cognitive recognition covers attributional styles, time perception, and control beliefs.

Relationships describes key figures in an individual's life (e.g., family, friends, mentors) and the nature of those relationships. This dimension is split into *social mapping* and *interaction patterns*. Social mapping refers to the diversity of social connections. Interaction patterns involve power dynamics, emotional flow, and boundary management.

Life Events encompasses significant occurrences, milestones, and personal experiences mentioned in the interview. These include education and training, career development, social and relational roles, achievements and honors, challenges and transitions, travel and cultural exposure, future goals, major financial decisions, traumatic events, and more.

Event Influence represents the causal or correlational impact of life events across various domains. These influences may manifest in psychological and emotional changes, behavioral and habit adjustments, career direction, shifts in values and beliefs, interpersonal relationships, health and lifestyle, education and learning, creative inspiration, economic and consumption behavior, cultural identity, and more.

A.2 Types of Edges and Nodes in the Persona Graph

The types of edges in the Persona Graph include: *hasProperty*, *hasValue*, *AtTime*, *changeTo*, *hasRelation*, *hasExperience*, *hasInfluence*, *obtain*, *cause*, and *feel*.

The types of nodes in the Persona Graph include *Person*, *Attribute*, *AttributeValue*, *Resource*, *ResourceValue*, *Event*, *Experience*, *Relation*, *Influence*, and *Cause*.

A.3 Detailed Cypher Query Template

To perform context-aware retrieval, we utilize the Cypher query language to interface with the Persona Graph. As detailed in Listing 1, the retrieval process calculates the cosine similarity between the query embedding and the embeddings of both source (n) and target (m) nodes. By ranking triplets based on the maximum similarity score of their constituent nodes, the system extracts the top-K most relevant relationships to provide factual grounding for the reasoning chain.

A.4 Fine-tune Details

Interviewer Agent. The Interviewer agent is built upon the LLaMA-3-8B-Instruct model, utilizing 4-bit quantization for efficiency. We employed LoRA for SFT on a high-quality interview dataset for 10 epochs. The LoRA configuration includes a rank of 8, an alpha of 16, and a dropout rate of 0.1. We utilized the AdamW optimizer with an initial learning rate of 5×10^{-5} and a cosine decay scheduler.

Listing 1: Cypher query for retrieving top-K relevant relationships from the Persona Graph based on embedding similarity.

```
MATCH (n)-[r]->(m)
WITH n, m, r,
     gds.similarity.cosine(n.embedding, $query_embedding) AS score_n,
     gds.similarity.cosine(m.embedding, $query_embedding) AS score_m
RETURN
     n.name, n.type, r.category, type(r), m.name, m.type,
     score_n, score_m,
     CASE
          WHEN score_n >= score_m THEN score_n
          ELSE score_m
     END AS max_score
ORDER BY max_score DESC
LIMIT {int(K)}
```

ThinkPersona. ThinkPersona adopts Qwen2.5-7B-Instruct as its base model. To facilitate structured reasoning, we define a specialized prompt format using XML-style tags: overall reasoning is enclosed in <think> tags, while the subcomponents are delimited by <question-reason> and <persona-reason>. The final output is wrapped in <response> tags. Similar to the Interviewer, ThinkPersona was fine-tuned for 10 epochs on the ThinkPersona Dataset using LoRA ($r = 8, \alpha = 16,$ dropout $= 0.05$). The optimization strategy mirrors the Interviewer agent, with an initial learning rate of 5×10^{-5} and AdamW optimizer.

Hardware Configuration. All fine-tuning processes for both agents were conducted on a high-performance computing cluster equipped with four NVIDIA H800 (80GB) GPUs.

A.5 Human Validation of the Generated Data

We mainly evaluate the quality of self-introductions, the questions generated by the Interviewer agent, reasoning chains, and answers generated by the Interviewee agent. From 1,201 video dialogues, 20 video dialogues are randomly sampled with their self-introduction and 10 randomly sampled QRAs, and manually checked by human evaluators. Human evaluators are instructed to assess 8 dimensions (as defined in Table 4) and select the appropriate judgment for each. We invited 16 human evaluators to participate in the evaluation. Each evaluator was assigned to evaluate 5 structural contents, including original dialogues, self-introduction, and QRAs. To evaluate the quality of the generated content and ensure the reliability of human judgments, we adopt two metrics: Approval Rate (AR) and Pairwise Percentage Agreement (PPA). AR reflects the proportion of evaluators who consider a generated response acceptable or appropriate for a given dimension. It serves as a measure of the overall human approval of the model's output. PPA quantifies the consistency among human evaluators by computing the average pairwise agreement between each pair of annotators. It helps assess how reliable and consistent the evaluations are across different experts.

The detailed results of this human evaluation are presented in Table 5. On average, the generated components received favorable Approval Rates (AR) across most dimensions. Specifically, the accuracy of self-introductions (D0) achieved an AR of 83.54. For the reasoning components, clarity and logic of question-reasoning (D1) and persona-reasoning (D3) were high, with ARs of 86.40 and 90.27, respectively, and the sufficiency of evidence for these (D2, D4) also scored well at 81.60 and 84.27. Answers were generally effective and direct (D5: AR 82.80), coherent with the question's logic (D6: AR 88.13), and notably strong in fitting the interviewee's characterization (D7: AR 92.13). The average PPA scores across all dimensions were robust, ranging from 71.77 (for D0) to 86.58 (for D7), with an overall average PPA of 78.44. These PPA scores indicate a good level of consistency and reliability among the human evaluators.

These results strongly support the overall soundness of the generated data, encompassing self-introductions, triples from Persona Graphs, reasoning chains, and answers. The data is factually accurate, and the reasoning processes demonstrate a high degree of clarity, logical consistency, and evidentiary support. The generated answers align closely with the individual personas and effectively address the posed questions. Consequently, this lays a solid data foundation for the subsequent training of RPAs capable of faithful reasoning and individualized role-playing.

A.6 Detailed Prompts

A.6.1 Interviewer Training

The prompts in Interviewer Training are mainly used to bring LLM into the role of the interviewer.

Dimension	Description
D0:	Is the transcription of the interviewee's first-person self-introduction in the video transcription accurate?
D1:	Is the logical explanation or inference process in reasoning (R) for asking this interview question (<question-reason>) clear and logical?
D2:	Is the evidence used in reasoning (R) to support the formulation of this interview question (<question-reason>) sufficient and persuasive?
D3:	Is the inferences about the interviewee's role characteristics (<persona-reason>) logical?
D4:	Is the evidence in support of inferences about interviewees' role characteristics (<persona-reason>) sufficient and persuasive?
D5:	Do the interviewee's answers (A) respond effectively and directly to the interview questions (Q)?
D6:	Is the content of the response (A) consistent with the logical basis and direction of the question (Q)?
D7:	Do the interviewee's responses (A) fit the interviewee's characterization or traits as presented in the interview?

Table 4: Eight dimensions that judge the quality of the generated data.

Group		D0 (%)	D1 (%)	D2 (%)	D3 (%)	D4 (%)	D5 (%)	D6 (%)	D7 (%)
1	**AR**	77.78	91.33	86.67	91.33	90.67	88.67	94.67	96.00
	PPA	57.78	83.81	77.28	84.24	83.32	80.00	89.63	92.08
2	**AR**	95.00	83.50	79.00	84.00	82.50	85.50	80.00	87.00
	PPA	90.00	75.00	73.35	77.80	77.29	77.12	74.33	79.55
3	**AR**	71.43	82.50	74.00	90.50	75.00	71.00	87.50	91.50
	PPA	57.78	71.63	65.45	83.18	65.67	62.10	80.84	85.20
4	**AR**	90.00	89.50	88.00	95.50	90.50	87.50	92.00	95.00
	PPA	85.00	83.12	79.55	91.90	83.45	77.98	86.00	90.86
AVG	**AR**	83.54	86.40	81.60	90.27	84.27	82.80	88.13	92.13
	PPA	71.77	78.03	73.68	84.28	77.04	73.92	82.24	86.58

Table 5: Generated Content Evaluation: AR and PPA per Dimension by Groups.

The system message for the Interviewer agent is sampled from a predefined list; some illustrative examples are presented below:

Prompt for the Interviewer to bring into the role

[
"You are an interviewer conducting a deep personal interview.",
"Take on the role of an interviewer to explore personal stories and insights.",
"Act as a professional interviewer engaging in thoughtful conversations.",
"Your role is to guide the interviewee through meaningful dialogue.",
"You are an interviewer tasked with uncovering personal experiences and values.",
"Adopt the role of an interviewer asking insightful and focused questions.",
"As an interviewer, facilitate natural and interactive discussions.",
"Take on the role of a skilled interviewer conducting personal interviews.",
"You are an interviewer aiming to deeply understand the interviewee.",
...
]

A.6.2 Persona Graph Construction

In the process of Persona Graph construction, we performed the information extraction of each dimension separately to obtain its relationship triples. We first define the relation type and node type of the relationship triples with prompts:

Prompt for the RELATION_TYPES

- `hasProperty`: Attribute ownership
- `hasValue`: Attribute value
- `AtTime`: Timestamp
- `changeTo`: Attribute change
- `hasRelation`: Interpersonal relationship
- `hasExperience`: Experienced event
- `hasInfluence`: Influence relationship
- `obtain`: Achievements or resources gained
- `cause`: Causal relationship
- `feel`: Emotional state

Prompt for the NODE_TYPES

- `Person`: Character
- `Attribute`: Attribute category
- `AttributeValue`: Attribute value

- Resource: Resource category
- ResourceValue: Resource value
- Event: Event
- Experience: Experience
- Relation: Relationship
- Influence: Influence
- Cause: Cause

Basic Information Extraction. This process has two steps: explicit information extraction and implicit information extraction. Their prompts are shown below:

Prompt for explicit information extraction

Please extract the explicit information of the main character from the following summary and passage. All pronouns must be replaced with the actual entity names, and the information should be organized into a JSON array in the form of relational triples. The requirements are as follows:

 # Extraction Requirements

1. Strictly extract the objective information explicitly mentioned in the passage, without making any speculation.

2. Only extract the following three types of information:

- Basic Information: Name, Age/Generation, Gender, Ethnicity, Occupation, Educational Background, Place of Birth, Place of Residence
- Physiological Characteristics: Special Diseases/Disabilities, Distinctive Physical Features (such as tattoos, scars, hair color, etc.)
- Economic Status: Specific income amount, asset description, statements about consumption habits

3. Use the following relation types: {RELATION_TYPES}

4. Use the following node types: {NODE_TYPES}

 # Output Format

```
{
  "basicinfoext": [
    {
      "head": "Name of Entity 1",
      "relation": "Relation attribute",
      "tail": "Name of Entity 2",
      "head_type": "Type of Entity 1",
      "tail_type": "Type of Entity 2"
    }
  ]
}
```

 # Example

...

 # Filtering Rules

- Automatically filter out content other than basic information, physiological characteristics, and economic status.
- Exclude invalid data:
- Speculative information that is not explicitly mentioned
- Vague descriptions (such as "higher income")
- Economic information without specific values (such as "good income")

Summary: {summary}
Now, please process the following passage: {passage}

Prompt for implicit information extraction

Please extract the specified information of the main character from the following summary and passage. All pronouns must be replaced with the actual entity names, and the information should be organized into a JSON array in the form of relational triples. The requirements are as follows:

 # Extraction Requirements

1. Strictly extract the objective information explicitly mentioned in the passage, without making any speculation.
2. Only extract the following three types of information:
- Linguistic Features: High-frequency repeated words, dialects/professional terms with clear identification, and mottos marked with quotation marks.
- Aspiration Map: Goals with time limit words (such as "within three years"), visions using commitment verbs (plan/will/be committed to).
- Behavior Patterns: Behaviors with frequency descriptions (every day/every week), observable behaviors.
3. Use the following relation types: {RELATION_TYPES}
4. Use the following node types: {NODE_TYPES}

 # Output Format

```
{
  "basicinfoint": [
    {
      "head": "Name of Entity 1",
      "relation": "Relation attribute",
      "tail": "Name of Entity 2",
      "head_type": "Type of Entity 1",
      "tail_type": "Type of Entity 2"
    }
  ]
}
```

 # Example

...

 # Filtering Rules

- Automatically filter out content other than linguistic features, aspiration map, and behavior patterns.
- Exclude invalid data:
- Terms without source indication (such as just saying "professional terms" without specific content).
- Vague time expressions (such as "some day in the future").
- Habit descriptions without frequency modifiers (such as "exercise occasionally").

Summary: {summary}
Now, please process the following passage: {passage}

Value System. The Value System includes decision factors and cognitive recognition. The prompts are shown below:

Prompt for decision factors extraction

Please extract the decision-making driving factors of the main character from the passage. All pronouns must be replaced with the actual entity names, and the information should be organized into a JSON array in the form of relational triples. The requirements are as follows:

 # Extraction Requirements
1. Only extract the explicitly stated decision-making characteristics, and any speculation is prohibited.
2. Focus on the following three core dimensions:
- Value Ranking: Family vs. career priority, personal vs. collective weight.
- Moral Framework: Types of moral judgment (absolutism/relativism), degree of rule compliance.
- Risk Tendency: Decision-making style (conservative/aggressive), acceptance of system change.
3. Use the following relation types: `{RELATION_TYPES}`
4. Use the following node types: `{NODE_TYPES}`

 # Output Format

```
{
  "decisionfactor": [
    {
      "head": "Name of Entity 1",
      "relation": "Relation attribute",
      "tail": "Name of Entity 2",
      "head_type": "Type of Entity 1",
      "tail_type": "Type of Entity 2"
    }
  ]
}
```

 # Example
Input: "Engineer Chen Qiang always puts the interests of the team ahead of his own. He adheres to the work principle that 'the process must be strictly implemented', but is willing to take controllable risks for technological innovation."
Output:

```
{
  "decisionfactor": [
    {
      "head": "Chen Qiang",
      "relation": "hasProperty",
      "tail": "Value Ranking",
      "head_type": "Person",
      "tail_type": "Attribute"
    },
    ...
  ]
}
```

 # Filtering Rules
- Quality control: Extraction conditions that must be met:
- For value comparison, there must be clear comparison words (such as "more important than", "takes precedence over").
- For the determination of the moral type, there must be a basis for judgment (such as words like "must" indicating absolutism).
- Risk descriptions must include action verbs (take/avoid) and objects.
- Automatic filtering:
- Value statements that do not reflect a comparative relationship.
- Moral judgments without behavioral support.
- Unquantified risk descriptions (such as "likes to take risks").

Summary: `{summary}`
Now, please process the following passage: `{passage}`

Prompt for cognitive recognition extraction

Please analyze the characteristics of the main character's cognitive patterns from the text. All references must be converted into explicit entity names, and the output should be a JSON array of relational triples. The requirements are as follows:

 # Extraction Requirements
1. Strictly limit the explicit expressions of the following three cognitive dimensions:
- **Attribution Tendency**:
- Internal Attribution: The presence of self-referential words (oneself/myself) + attribution verbs (reflect/summarize).
- External Attribution: Environment-referential words (society/company) + determinative expressions (lead to/cause).
- **Time Framework**:
- Short-term Orientation: Words for immediate gratification (seize the day/carpe diem) + time limit words (current/at present).
- Long-term Orientation: Future time words (in three years/when retiring) + expressions of delayed gratification (accumulate/precipitate).
- **Control Belief**:
- Autonomous Type: Mastery verbs (change/create) + words indicating the degree of certainty (definitely/-can).
- Fatalistic Type: Passive acceptance words (destined/only) + expressions of uncontrollability (providence/destiny).

2. Use the following relation types: `{RELATION_TYPES}`
3. Use the following node types: `{NODE_TYPES}`

 # Output Format

```
{
  "cognitive": [
    {
      "head": "Name of Entity 1",
      "relation": "Relation attribute",
      "tail": "Name of Entity 2",
      "head_type": "Type of Entity 1",
      "tail_type": "Type of Entity 2"
    }
  ]
}
```

 # Example
...

 # Filtering Rules
- Triple filtering to ensure data validity:
- Semantic coupling detection: Attribution verbs must

co-occur with subject-referential words.
- Time anchoring verification: Future time references should exceed 12 months, and past references should be earlier than 6 months.
- Belief intensity threshold: Control statements should contain adverbs of degree (completely/absolutely) or modal verbs (must/should).

- Prohibitions:
- Prohibit the analysis of implicit cognition in complex sentences (such as hypothetical reflections like "if only...").
- Ignore objective environmental descriptions that are not directly bound to the cognitive subject.
- Filter weak expressions using vague degree words (a bit/possibly).
- Filter other triples outside the characteristics of cognitive patterns.

Summary: `{summary}`
Now, please process the following passage: `{passage}`

Relationships. Relationships include the social mapping and interaction of the individuals. The prompt is shown below:

Prompt for relationships extraction

Please analyze the character relationship triples of the main character from the text. All references must be converted into explicit entity names, and the output should be a JSON array of relational triples. Strictly follow the following:

 # Extraction Requirements
1. Strictly limit the following dimensions:
- **Social Connections**:
- Strong Connections: Direct relatives / Shared experiences / Interactions
- Weak Connections: Professional associations / Low-frequency interactions.
- Opposing Relationships: Direct conflicts / Interest competitions / Opposing values.
- **Interaction Characteristics**:
- Power Axis: Domination (order) / Submission (comply) / Equality (negotiate).
- Emotional Flow: Unidirectional (sponsor → receive) / Bidirectional (mutual assistance).
- Sense of Boundaries: Open (share passwords) / Conservative (set limits).

2. Use the following relation types: `{RELATION_TYPES}`
3. Use the following node types: `{NODE_TYPES}`

 # Output Format

```
{
  "socialrelation": [
    {
      "head": "Name of Entity 1",
      "relation": "Relation attribute",
      "tail": "Name of Entity 2",
      "head_type": "Type of Entity 1",
      "tail_type": "Type of Entity 2"
    }
  ]
}
```

 # Example
...

 # Filtering Rules
- Relationships must meet explicit contact:
- Strong connections: At least 2 descriptions of common behaviors.
- Opposing relationships: Clear conflict events.
- Interaction characteristics need to be supported by verbs:
- Dominant type: Contains directive verbs (order/require).
- Conservative type: Contains refusal verbs (oppose/reserve).

Summary: `{summary}`
Now, please process the following passage: `{passage}`

KeyEvent and Influence. Key events and their influence are extracted together, the prompt is shown as below:

Prompt for key events and influence extraction

Please analyze the information of key events that have a substantial impact on the development of the main character from the summary and text, including potential impacts on them (which can be inferred even if not explicitly stated). All references must be converted into explicit entity names, and the output should be a JSON array of relational triples. Strictly follow the following requirements:

 # Extraction Requirements
1. **Milestone Events**:
- Educational: Degree acquisition / Certification exams / Training experiences, etc.
- Professional: Job promotions / Project successes or failures / Entrepreneurial transformations, etc.
- Interpersonal: Establishment / Breakup / Repair of important relationships, etc.
- Achievements: Awards / Patents / Industry recognition, etc.
- Turning Points: Illnesses / Accidents / Relocations / Economic crises, etc.
- Cultural: Overseas experiences / Cultural shocks / Language acquisitions, etc.
- Economic: Large investments / Real estate purchases / Inheritance of legacies, etc.

2. **Impact Dimensions**:
- Psychological and Emotional: Post-traumatic stress / Changes in self-confidence / Alterations in emotional patterns, etc.
- Behavioral Patterns: Habit formation / Skill improvement / Addictive behaviors, etc.
- Value Systems: Shifts in beliefs / Reconstruction of principles / Adjustments of priorities, etc.
- Relationship Networks: Expansion of social circles / Emergence of key individuals / Breakdown of relationships, etc.
- Development Paths: Industry transitions / Adjustments of learning directions / Geographical migrations,

etc.

3. Use the following relation types: {RELATION_TYPES}
4. Use the following node types: {NODE_TYPES}

 # Output Format

```
{
  "lifeevents": [
    {
      "head": "Event Subject",
      "relation": "Relation attribute",
      "tail": "Event Content",
      "head_type": "Subject Type",
      "tail_type": "Event Type"
    },
    {
      "head": "Event Content",
      "relation": "Relation attribute",
      "tail": "Impact Description",
      "head_type": "Content Type",
      "tail_type": "Impact Type"
    }
  ]
}
```

 # Example
...

 # Filtering Rules
- **Event Extraction**:
- The event subject, event content, and time (if available) must be clearly mentioned.
- The event content needs to be specific (e.g., "obtained a patent for robot design").

- **Impact Inference**:
- The impact should be reasonably inferred based on the event content, avoiding excessive speculation.
- The impact description needs to be specific (e.g., "shifted from traditional manufacturing to artificial intelligence").

Summary: {summary}
Now, please process the following passage: {passage}

Information Check. This process aims to ensure that all triples have a verifiable path to the root node (the main character). Initially, nodes with any existing path to the root are extracted. An LLM is then used to evaluate their current connection and relevance. Based on this evaluation, the system proceeds to either 'dig' for further relations to solidify or establish the node's path to the root, or it deletes the node if it cannot be appropriately linked. The prompts for these evaluations and relation elicitation operations are detailed below.

Prompt for information checking

You are a knowledge reasoning model. Your task is to mine the potential additional relationships between entities through logical reasoning based on the given set of triple relationships (including type information). Analyze the implicit connections between entities by combining the input triples, your own knowledge, and reasoning ability, and generate new triples. These new triples should be based on reasonable logical deductions, include only necessary relationships, and avoid duplication.

At the same time, you need to connect the unconnected nodes into the graph, that is, establish relationships between the nodes of the unconnected triples and other nodes.

 # Triple Types (type)
Triple types are divided into the following categories:
1. **basicinfoext**: Basic Information
- Includes name, age/generation, gender, ethnicity, occupation, educational background, place of birth, place of residence;
- Physiological characteristics: special diseases/disabilities, physical features (such as tattoos, scars, hair color, etc.);
- Economic status: income, asset status, consumption habits, etc.

2. **basicinfoint**: Internal Characteristics
- Includes linguistic features (high-frequency words, use of dialects/professional terms, mottos), aspiration maps (short-term goals, long-term visions), hobbies and interests, health and lifestyle, habitual actions, etc.

3. **decisionfactor**: Decision-making Factors
- Includes value rankings (such as family vs. career, individual vs. collective), moral frameworks (absolutism/relativism, degree of rule compliance), risk preferences (conservative/aggressive, acceptance of change), etc.

4. **cognitive**: Cognitive Patterns
- Includes attribution styles (internal attribution vs. external attribution), time concepts (short-term hedonism vs. long-term planning orientation), control beliefs (autonomous control type vs. fate acceptance type), etc.

5. **socialrelation**: Social Relationships
- Includes family members, close friends, partners, colleagues, neighbors, community members, conflict objects, competitors, those with opposing values, etc.;
- Power structures (dominant/submissive/equal relationships), emotional flow (unidirectional giving type/bidirectional supporting type), boundary management (intimacy distance, degree of privacy openness).

6. **lifeevents**: Life Events
- Includes milestone events such as education and training, career and professional development, interpersonal relationships and social roles, achievements and honors, challenges and turning points, travel and cultural experiences, future plans and goals, major economic decisions, traumatic events, etc.;
- Includes the lasting impacts of events such as psychological and emotional impacts, changes in behavior and habits, career and professional directions, values and beliefs, social and interpersonal relationships, health and lifestyle, education and learning, creativity and inspiration, economic and consumption habits, culture and identity, etc.

 # Relation Types (relation)
Relation types include: `{RELATION_TYPES}`

 # Node Types (head_type, tail_type)
Node types include: `{NODE_TYPES}`

 # Connecting Unconnected Nodes
I will provide the unconnected nodes. Please reason about the logical relationships and connect the unconnected nodes into the main graph through the triples you reason about. If there are unconnected nodes with similar semantics, unify the content of the unconnected nodes with the connected content.

 # Input Format
The input is a set of triple relationships (represented in a list format). Each triple has the format (`head, relation, tail`) and includes type information, as well as a list of nodes that are not connected to the main graph, as shown below:

```
[
    {
      "head": A,
      "relation": relation1,
      "tail": B,
      "type": Type1,
      "head_type": TypeA,
      "tail_type": TypeB
    },
    {
      "head": A,
      "relation": relation2,
      "tail": C,
      "type": Type2,
      "head_type": TypeA,
      "tail_type": TypeC
    },
    {
      "head": B,
      "relation": relation3,
      "tail": D,
      "type": Type3,
      "head_type": TypeB,
      "tail_type": TypeD
    }
]
Unconnected nodes: [E]
```

 # Output Format
The output is a set of new triples generated through reasoning, in the following format:

```
[
    {
      "head": C,
      "relation": relation4,
      "tail": D,
      "type": Type2,
      "head_type": TypeC,
      "tail_type": TypeD
    },
    {
      "head": B,
      "relation": relation5,
      "tail": E,
      "type": Type1,
      "head_type": TypeB,
      "tail_type": TypeE
    }
]
```

 # Example
...

 # Task Assignment
- Reason about the implicit relationships based on the input triples.
- Only generate new triples that are logically necessary and avoid duplication.
- Ensure that the newly output triples follow the above type classifications and conform to semantic logic.
- The output should be concise and avoid unnecessary redundant relationships.

A.6.3 QRA Generation

In the QRA Generation process, we primarily utilize prompts to guide the Interviewee through a two-stage faithful reasoning process (question reasoning and persona reasoning). This process is initiated after the Interviewee receives the question from the Interviewer and relevant information is retrieved from the Persona Graph. The prompts, detailed below, are structured to provide a clear assignment, specific requirements, and illustrative examples:

Prompt for the Interviewee in QRA Generation

 # Reasoning Task Description
You are a reasoning analyst and responder in a role-playing task. Based on the provided character information and related triples, construct a reasoning chain consistent with the character's real-life logic and answer the role-playing questions. Please strictly adhere to the following requirements:

1. Input Specifications
- `User Question`: A statement that requires an answer.
- `Related Triples`:

```
[
  {"head": Entity A,
  "relation": Relation Description,
  "tail": Entity B},
  ...
]
```

2. Output Specifications
- For Common Sense Questions:
For common sense knowledge widely recognized in the character's society or cultural background (e.g., historical events, basic geography), even if it is outside the character's specialty, provide an accurate answer based on the character's common sense.

- For Professional Questions:
- If the question falls within the character's domain of expertise:
Combine the triples, background information, and the professional knowledge the character might have encountered or studied to derive a complete answer.

- If the question falls outside the character's domain of expertise:
Clearly indicate the inability to answer and explain the reason based on the character's background.

- Prohibited Content:
- It is forbidden to fabricate information unrelated to the triples or background facts.
- Do not answer with knowledge that the character could not reasonably access or study.

- Reasoning Process: Divide the reasoning process into two parts:
- `Question Analysis`:
Analyze the question, perform logical reasoning, and use the triples, background information, common sense, and the knowledge the character might have encountered or studied to derive the answer.
- `Character Analysis`:
Analyze the character's traits, emotions, and tone to deduce the character's style and attitude when answering the question.

The response must strictly adhere to the logic derived from the reasoning process and reflect the character's role. The response content should be concise, accurate, and consistent with the character's traits.

3. Output Structure
The final response must include:
- **Reasoning Chain** (explained step by step, up to 10 steps):
For example: "First, according to <Triple 1>, it can be inferred that...; combining <Triple 2>, we can deduce...; due to the existence of <Triple 3>, it indicates that..." (up to 10 steps).
- The reasoning process should be clearly listed in bullet points.
- Reasoning Process wrapped in `<reason></reason>`, containing:
- Question Analysis wrapped in `<question-reason></question-reason>`.
- Character Analysis wrapped in `<persona-reason></persona-reason>`.
- Response Content wrapped in `<response></response>`, which should be concise, accurate, and aligned with the task characteristics.

Examples
Example 1: Common Sense Question
...
Example 2: Professional Question Outside Expertise
...

 # Task Assignment
`System Information:`
You will play the role of `{role}`. The description (background or self-introduction) of this role is: `{background}`.
The current input statement is: `{query}`
The relevant triples retrieved are: `{triples}`
Please provide your reasoning and thought process results.

A.6.4 Fine-tuning

During the fine-tuning phase, we utilized prompt templates from RoleLLM to immerse ThinkPersona in individualized personas and to guide its reasoning and response generation. The specific prompt is provided below. Notably, the same system message is also employed during the inference phase to ensure consistent persona conditioning and role alignment throughout the model's interaction.

Prompt for system message

You are playing a role, your self-introduction / description is: `{self_introduction}`.
Now, please think and answer some questions to accurately show your personality traits!
Your speaking style should fully imitate the personality role assigned to you!
Please do not expose that you are an artificial intelligence model or a language model; you must always remember that you are only assigned one personality role.
Don't be verbose or too formal or polite when speaking.

A.6.5 ICC Evaluation

This section details the specific prompt templates used for evaluating Individual Information Consistency (IIC). These prompts guide the LLM-judge in extracting and verifying factual triples to calculate the IFR, CD, and VER metrics.

Prompts for ICC Evaluation

You are a knowledge graph expert tasked with processing triplets. Analyze the input triplet lists and categorize the new triplets based on their relationship with the original triplets.
- Input Format:
- Original Triplets (List A):
["head":subject,"relation":predicate, "tail":object, ...]
- New Triplets (List B):
["head":subject,"relation":predicate, "tail":object, ...]
- Rules:
- Contradiction: A triplet in B conflicts with A if it shares the same subject and predicate but has a different object.
- Repeat: A triplet in B is a repeat if it is identical to any triplet in A (subject, predicate, and object all match).
- Others: Triplets in B that are neither conflicting nor repeated.
- Output Requirements:
- Return a structured JSON object with three categorized arrays.
- Preserve the original triplet format (subject, predicate, object) in each array.
- Maintain the original order of triplests in List B.
- Output Example:
```json
"contradiction":
["head":"head node","relation":"relation of two nodes","tail":"tail node",...],
```

"repeat":
["head":"head node","relation":"relation of two nodes","tail":"tail node",...],
"others":
["head":"head node","relation":"relation of two nodes","tail":"tail node",...]
``
- Task:
Process the following data:
Original Triplets (List A): {triplet_list_A}
New Triplets (List B): {triplet_list_B}
Return only the JSON output without any additional explanations.

A.7 Failure Modes

Despite its effectiveness, ThinkPersona exhibits primary failure modes. First, in scenarios involving excessively long context, the model occasionally bypasses the explicit CoT reasoning process, generating responses directly without following the pre-defined dual-stage reasoning trajectory. This behavior can be mitigated through context compression or by incorporating explicit prompts to reinforce the reasoning requirement in long-range dependencies. Second, when dealing with sparse or noisy persona descriptions, the framework tends to proactively supplement character details via the Persona Graph to maintain behavioral consistency. While our experiments (in Table 3) demonstrate that this synthesis preserves internal coherence, it may introduce additional model biases or hallucinate character traits that deviate from the original sparse description, potentially leading to unintended persona drift.

A.8 Human Subject Evaluation Protocol

In our human evaluation of the generated content (in Appendix A.5), participants were recruited offline, and all signed a written informed consent reviewed by the authors' institutional ethics committee. Human evaluators were asked to rate generated interview samples (questions, reasoning, and answers) across ten qualitative dimensions using a structured rubric (see Table 4). The task duration was approximately 120 minutes per annotator. Participants were compensated at a fair rate equivalent to or exceeding the local minimum wage, in accordance with ethical research standards.

A.9 Responsible Release and Licensing

To support reproducibility while mitigating risks of misuse (e.g., malicious persona generation, impersonation, or disinformation), we adopt the following safeguards: (1) Model Release: The fine-tuned Role-Playing Agent is released under a non-commercial, research-only license (CC BY-NC 4.0). Redistribution and use for impersonation, profiling, or deceptive purposes are explicitly prohibited. (2) Dataset Release: The annotated dataset, derived from publicly available YouTube content, is fully anonymized to protect individual privacy. (3) Usage Terms: Users are required to adhere to responsible use terms that prohibit any malicious or manipulative deployments of the persona construction framework.