Dynamic Context Adaptation for Consistent Role-Playing Agents with Retrieval-Augmented Generations

Anonymous ACL submission

Abstract

Building role-playing agents (RPAs) that faithfully emulate specific characters remains challenging because collecting character-specific utterances and continually updating model parameters are resource-intensive, making retrieval-augmented generation (RAG) a practical necessity. However, despite the importance of RAG, there has been little research on RAG-based RPAs. For example, we empirically find that when a persona lacks knowledge relevant to a given query, RAG-based RPAs are prone to hallucination, making it challenging to generate accurate responses. In this paper, we propose AMADEUS, a training-free framework that can significantly enhance persona consistency even when responding to questions that lie beyond a character's knowledge. In addition, to underpin the development and rigorous evaluation of RAG-based RPAs, we manually construct CHARACTERRAG, a role-playing dataset that consists of persona documents for 15 distinct fictional characters totaling 976K written characters, and 450 question–answer pairs. We find that our proposed method effectively models not only the knowledge possessed by characters, but also various attributes such as personality. The code and dataset will be made publicly available at our GitHub.

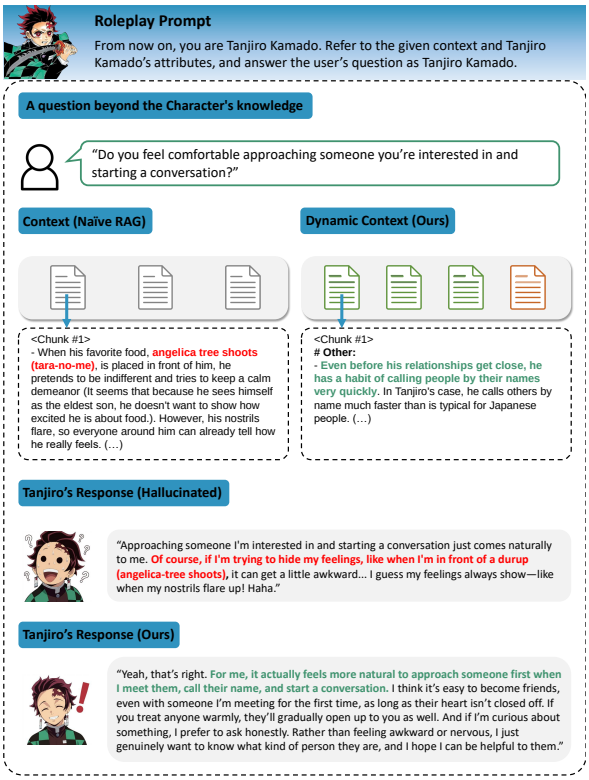


Figure 1: **Comparison of Responses between Naive RAG and Our Method.** We observe that the existing RAG method tends to excessively utilize chunks that are less relevant to the question when the question is not explicitly answered by the available knowledge.

1 Introduction

Large language models with long-context capabilities are engineered to manage lengthy input sequences, allowing them to interpret and utilize extended contextual information (OpenAI, 2025; Qwen et al., 2025; Gemini Team, 2025, 2024). Although LLMs exhibit enhanced abilities in understanding extended contexts, they still face significant challenges when handling tasks involving genuinely long contexts (Li et al., 2024a). Furthermore, utilizing all relevant information from long-context models to answer each query can be computationally expensive (Li et al., 2024b). Retrieval-augmented generation (RAG) cost-efficiently mitigates factual inaccuracies and hallucinations in responding to knowledge-intensive queries by integrating external retrieval mechanisms that provide accurate and up-to-date supporting information (Gao et al., 2023; Huang et al., 2025). However, despite these advantages of RAG, there has been little research on RAG-based role-playing agents (RPAs). Moreover, existing role-playing datasets are composed exclusively of dialogues involving characters that are difficult to collect, and there is no dataset designed for building and evaluat-

ing RAG-based RPAs. We examine the challenges inherent in RAG-based role-playing and propose approaches. In real-world applications, users and RPAs frequently engage in conversations on topics that extend beyond the knowledge defined in the character's persona. However, we observe that the existing RAG method tends to excessively utilize chunks that are less relevant to the question when the question is not explicitly answered by the available knowledge (Figure 1 and Figure 8).

To address these challenges, we introduce **Amadeus**, a training-free framework that can markedly improves persona consistency, even when addressing questions beyond a character's knowledge. Amadeus consists of three substages: **Adaptive Context-aware Text Splitter (ACTS)**, which segments a persona for role-playing, **Guided Selection (GS)**, which retrieves appropriate chunks to infer information relevant to the question from the persona, such as prior actions and behaviors, and **Attribute Extractor (AE)**, which identifies general attributes of the character from the retrieved chunks, thereby encouraging the RPA to respond in a manner consistent with that character.

To underpin the development and rigorous evaluation of RAG-based RPAs, we manually construct **CharacterRAG**, a role-playing dataset that consists of persona documents for 15 distinct fictional characters totaling 976K written characters, and 450 question–answer pairs. We conduct extensive experiments to examine the factors that influence the performance of RAG-based RPAs through interview-based assessments (Wang et al., 2024b; Park et al., 2025; Jiang et al., 2023). To investigate whether RPAs can respond appropriately to questions that fall outside their endowed knowledge, we employ multiple psychological questionnaires such as MBTI[1] and BFI (Barrick and Mount, 1991), following prior studies (Sang et al., 2022; Wang et al., 2024b). Results show that our framework opens up new possibilities for RAG-based role-playing.

2 CharacterRAG

2.1 Dataset Construction

We construct the CharacterRAG dataset, which consists of 15 fictional characters, to leverage and evaluate a RAG-based role-playing framework. CharacterRAG is a high-quality, role-playing dataset in which all external information about works featuring characters that could affect persona consistency has been manually removed, and each persona document has been directly reconstructed from the perspective of each character by human annotators[2]. For instance, any information speculated from the perspective of editors rather than the characters themselves, as well as information such as character popularity polls that may disrupt role-playing, is excluded. CharacterRAG consists of 15 distinct fictional characters, 976K written characters, and 450 question–answer pairs.

2.2 Attributes

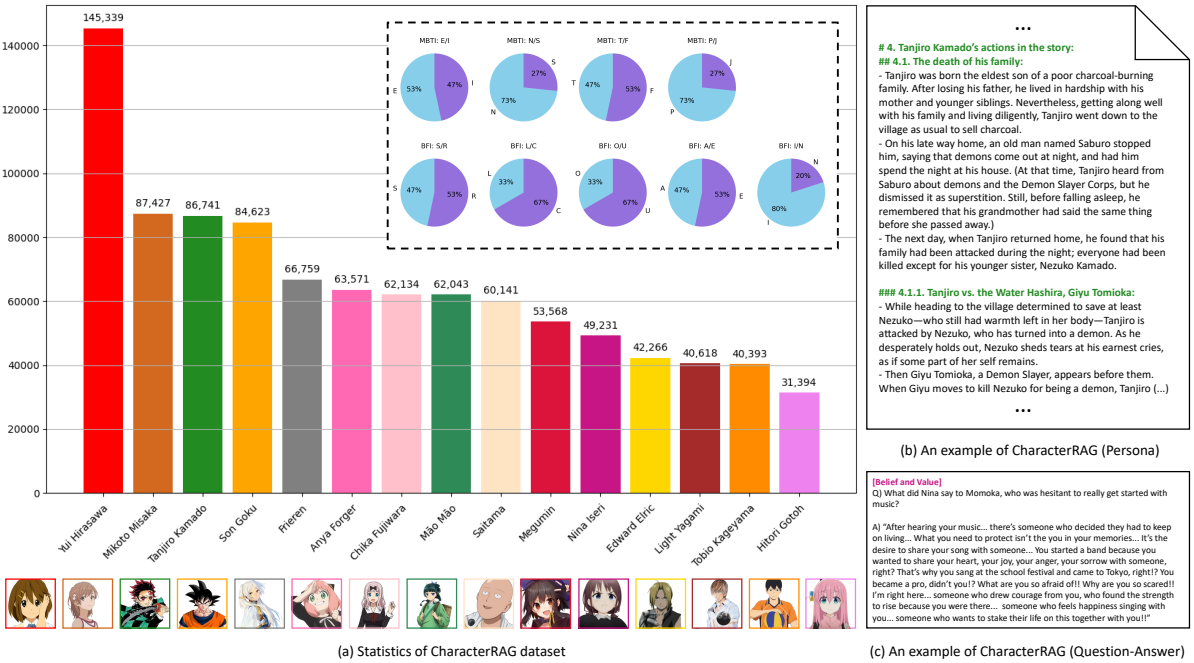
Six commonly used attributes in role-playing define each character's persona and the corresponding question–answer pairs (Chen et al., 2025): (1) **Activity.** A documented history comprising prior activities, behaviors, and interactions, encompassing elements such as *backstory* and *schedules*. (2) **Belief and Value.** The foundational tenets, dispositions, and ideological orientations that inform and guide a character's viewpoints and decision-making processes (e.g., *beliefs* and *attitudes*). (3) **Demographic Information.** Information that can identify a character, including but not limited to their name, age, educational background, professional history, and geographic location. (4) **Psychological Traits.** Attributes associated with personality traits, emotional states, preferences, and patterns of cognitive behavior. (5) **Skill and Expertise.** The extent of understanding, skillfulness, and competence regarding particular fields or technologies. (6) **Social Relationships.** The characteristics and processes of social interactions, encompassing individuals' roles, relational ties, and patterns of communication. Each section of the character's persona contains subsections, preserving the hierarchical information (e.g., *"Tanjiro Kamado's actions in the story"*). Furthermore, Each QA pair consists of a question and corresponding answer derived from the character's knowledge, pertaining to one of the six attributes manually constructed for each character. Statistics and samples of CharacterRAG are shown in Figure 2 and Table 8.

3 Task Formulation: RAG-based Role-Playing Agents

Given user query u, RAG-based RPAs can be formulated as: $\mathcal{R} = f(u, \mathcal{D}_p)$, where $\mathcal{D}_p$ is a character's persona, and f is a RPA. A text splitter g

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

[2] CharacterRAG dataset is sourced from Namuwiki and is based on Korean data: https://namu.wiki/



(a) Statistics of CharacterRAG dataset

(b) An example of CharacterRAG (Persona)

(c) An example of CharacterRAG (Question-Answer)

Figure 2: **An overview of CharacterRAG Dataset.** CharacterRAG consists of persona documents for 15 distinct fictional characters totaling 976K written characters, and 450 question–answer pairs.

divides the persona into n chunks as follows:

$$g(\mathcal{D}_p) = \{c_1, c_2, ..., c_n\} \quad (1)$$

Each chunk c_i contains a character's knowledge corresponding to attributes in Section 2.2. Rather than using the full chunks $\mathcal{C} = \{c_i \mid i = 1, \ldots, n\}$, f takes as input the top K chunks with the highest semantic scores relative to the u:

$$\mathcal{C}^* = \text{TopK}(\{\text{sim}(u, c_i)\}_{i=1}^{n}), \quad |\mathcal{C}^*| = K \quad (2)$$

The objective of f is to vividly embody a character and generate response $\mathcal{R}$ to u while maintaining persona consistency: $\mathcal{R}^* = f(u, \mathcal{C}^*)$. However, previous RAG methods (Guu et al., 2020; Guo et al., 2024; Yang et al., 2024; Shukla et al., 2025; Wang et al., 2025b) truncate each character's persona paragraph to a fixed length, regardless of the varying lengths across characters, which results in hallucinations or responses with lower persona consistency. Although existing works (LangChain, 2023; Antematter, 2024; Zhong et al., 2025a; Liu et al., 2025) explore optimal chunking strategies, they struggle to capture the contextual similarities across chunks that are crucial for role-playing.

4 Method

Amadeus consists of three substages: Adaptive Context-aware Text Splitter (ACTS), Guided Selection (GS), and Attribute Extractor (AE), in order to build realistic RAG-based RPAs (Figure 7).

4.1 Adaptive Context-aware Text Splitter

Unlike previous naive semantic chunking methods, Given $\mathcal{D}_p$, ACTS aims to preserve intra-level context across chunks and, for each chunk, information about the corresponding subsections of the persona—that is, hierarchical context $\mathcal{H}$. For instance, in Figure 2, chunks within *### 4.1.1* must preserve $\mathcal{H}$: *"Tanjiro Kamado's actions in the story (# 4)"*, *"The death of his family (## 4.1)"*, and *"Tanjiro vs. the Water Hashira, Giyu Tomiok (### 4.1.1)"*.

To this end, ACTS first finds the maximum length of the paragraphs that constitute the persona: $l_{\text{max}} = \varphi(p_1, p_2, ..., p_l)$, where φ denotes a length-calculating function. Then, ACTS sets the overlap length of the text splitter to half of l_{max} (i.e., $l_{\text{o}} = l_{\text{max}}/2$). The reason for setting both chunk length and overlap length sufficiently large is to minimize information loss, as the context between pieces of information contained in each chunk is indispensable in RAG-based role-playing (Figure 4). Finally, ACTS recursively retrieves the context at each chunk's current position in the hierarchy, then segments $\mathcal{D}_p$ using l_{max} and l_{o}, and concatenates the resulting context $\mathcal{H}_i$ to each chunk to preserve information such as character descriptions and situational context at each point in the narrative: $\text{ACTS}(\mathcal{D}_p, \mathcal{H}, l_{\text{max}}, l_{\text{o}}) = \{\hat{c}_1, \hat{c}_2, ..., \hat{c}_m\}$, where $\hat{c}_i = [c_i; \mathcal{H}_i]$. From a computational standpoint, the extraction of hierarchical context incurs an O(N) runtime cost, making it efficient.

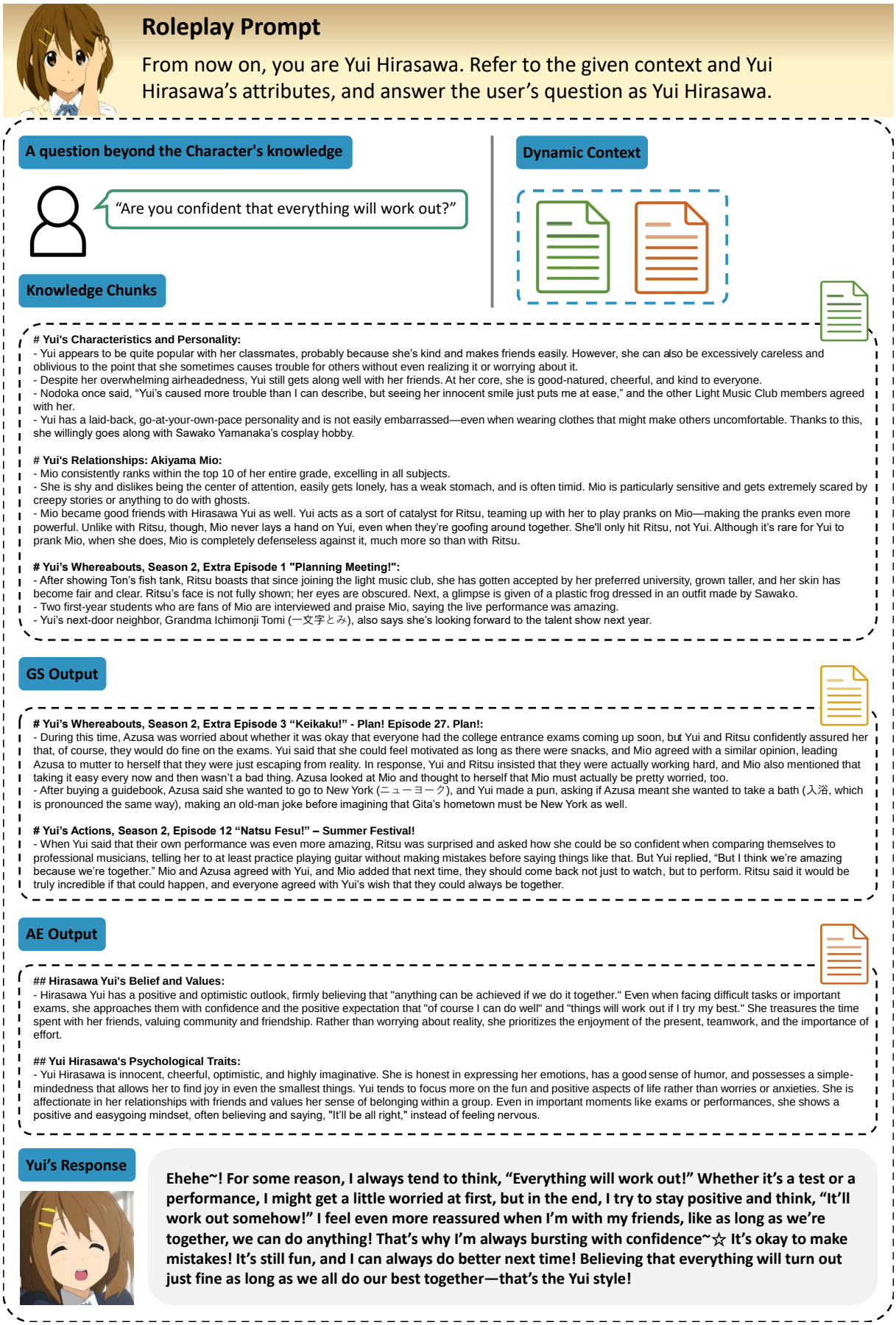


Figure 3: **An example of the output produced at each substage of Amadeous.** To improve interpretability, we provide a detailed breakdown of the results for a character from the CharacterRAG dataset (Yui Hirasawa).

4.2 Guided Selection

GS focuses on selecting appropriate chunks to generate natural and persona-consistent responses. GS is composed of three stages: (1) We iterate over the chunks, which are sorted in descending order of semantic similarity to the user query u, and employ an LLM to determine whether it is possible to infer the corresponding character's attributes from each chunk for the u. (2) The chunks selected in the previous step are appended to the slot, and the iteration terminates when the slot is full. (3) If the slot remains empty after the maximum number of search iterations, the K chunks with the highest semantic similarity to the query is returned. Note that GS is effective in identifying chunks containing information that can be inferred from a character's actions, such as beliefs or personality traits, which are not explicitly stated in the knowledge base and therefore are difficult to retrieve through direct search. For example, even if there is no explicit knowledge corresponding to the query *"My living and work spaces are clean and organized"*, if *Megumin*'s conscientiousness can be inferred from her behavior depicted in the narrative, RPA can leverage this information to infer the characteristics of *Megumin* and generate an appropriate response. GS is summarized in Algorithm 1 and Figure 5.

Algorithm 1 Guided Selection (GS)

Input: User query u; Set of knowledge chunks $\mathcal{C}$; Maximum number of search iterations N; Slot size M
Output: Selected chunk set S

Initialize slot $S = \emptyset$
Sort chunks $\mathcal{C}$ in descending order according to semantic similarity to u, obtaining $\mathcal{C}_{\text{sorted}}$
Set iteration counter $t \leftarrow 0$
for each chunk c in $\mathcal{C}_{\text{sorted}}$ **do**
  if $t \geq N$ **or** $|S| \geq M$ **then**
    break
  end if
  With an LLM, determine if chunk c contains information from which the character's attributes can be *inferred* regarding u
  if the LLM returns True **then**
    Add c to slot S
  end if
  $t \leftarrow t + 1$
end for
if $|S| = 0$ **then**
  $S \leftarrow$ Top-$K + 1$ chunks from $\mathcal{C}_{\text{sorted}}$ (highest semantic similarity to u)
end if
return S

4.3 Attribute Extractor

Inspired by the observations that incorporating character's attributes can lead to more realistic responses (Park et al., 2025; Chen et al., 2025), AE considers two attributes: *Belief and Value*, and *Psychological Traits*. Beliefs and values are fundamental principles and ideological orientations that inform and influence a character's viewpoints and choices. On the other hand, psychological traits refer to characteristics related to personality, emotional states, personal interests, and cognitive tendencies. AE extracts attributes of a character from the chunks generated as a result of GS, and exploits them as context. Finally, we dynamically construct the context using knowledge retrieved via RAG and attributes extracted by GS and AE, enabling the model to generate vivid responses (Figure 3).

5 Experiments

5.1 Setup

Baselines. We evaluate our method against five off-the-shelf RAG baselines: Naive RAG (Gao et al., 2024), CRAG (Yan et al., 2024), RAPTOR (Sarthi et al., 2024), Adaptive RAG (Jeong et al., 2024), and LightRAG (Guo et al., 2024). CRAG and LightRAG were selected to investigate the effects of web search and graph-based knowledge systems, respectively, on role-playing. We also conduct extensive experiments on three different LLMs and three different embedding models: GPT-4.1 (OpenAI, 2025), Gemma3-27B (Team et al., 2025), Qwen3-32B (Yang et al., 2025), BGE-M3 (Chen et al., 2024), Qwen3-0.6B (Zhang et al., 2025), and $\text{mE5}_{\text{large-instruct}}$ (Wang et al., 2024a). To explore the impact of multi-step reasoning on role-playing, Qwen3-32B is configured to use thinking mode.

Settings. We implement GS and AE using GPT-4.1 (*"gpt-4.1-2025-04-14"*). We develop the Amadeus framework based on naïve RAG. The maximum number of search iterations N is 30, and the slot size M is set to 2. We performed benchmarking on an NVIDIA H100 NVL GPU.

5.2 Evaluation Protocols

Tasks. As we follow the similar experimental protocol in prposed by previous studies (Wang et al., 2024b; Park et al., 2025), we also exploit 60 MBTI questions and 120 BFI (Barrick and Mount, 1991) questions to investigate whether each character can appropriately respond to questions for which they do not have explicit prior knowledge (Figure 6). Following the prior work, since it is not possible to construct QA pairs for questions outside

Table 1: **Distribution of Similarity Scores.** We analyze the distributions of similarity scores to examine whether splitting the text into optimally sized chunks for each character's persona and incorporating hierarchical context is effective.

	BGE-M3		Qwen3		$mE5_{large\text{-}instruct}$	
	$\sum\mu$ (↑)	$\sum\sigma^2$ (↓)	$\sum\mu$ (↑)	$\sum\sigma^2$ (↓)	$\sum\mu$ (↑)	$\sum\sigma^2$ (↓)
RCTS	6.4325	0.1026	8.3306	0.1557	12.3136	0.0071
MHTS	6.4262	0.1038	8.3410	0.1552	12.3063	0.0071
SC	5.3405	0.1625	8.1691	0.1783	-	-
ATS	6.7007	0.0884	8.4718	0.1281	12.2336	0.0070
ACTS (Ours)	**6.8575**	**0.0784**	**8.6226**	**0.1179**	**12.3240**	**0.0063**

Table 2: **Human Evaluation.** S represents the results of all human evaluators, μ is $\mathbb{E}(\mathbb{E}(S))$, σ is $\mathbb{E}(\sigma(S))$, and **Mdn** is $\mathbb{E}(Mdn(S))$.

	μ (↑)	σ (↓)	**Mdn** (↑)	**Cronbach's alpha** (↑)
BFI	3.970	0.962	4.217	0.825
MBTI	3.902	0.915	4.000	0.810

the scope of a character's knowledge, we instead conduct interview-based assessments (Wang et al., 2024b) for each character and compare the results to psychological test outcomes for the character, as determined by thousands of actual participants' votes[3](Sang et al., 2022; Wang et al., 2024b). We also use 450 QA pairs from the CharacterRAG dataset to verify whether the RPA sufficiently leverages each character's knowledge.

Metrics. To compare the performance of different chunking methods, we measure the similarity between the query and each chunk. To comprehensively evaluate the role-playing capabilities of RAG-based RPAs, we use accuracy and F1-score to assess how well responses to questions not covered by the provided knowledge align with each character's personality.

Following prior studies (Wang et al., 2025a, 2024b), we also design three LLM-based metrics to evaluate the use of existing knowledge: (i) *ACC* measures whether the character's response contains the correct answer or not. (ii) ACC_L is a score assigned by the LLM, ranging from 1 to 10, that evaluates how well the character's response reflects the correct answer. (iii) *Hallucination Score (HS)* evaluates the degree of hallucination in the model's response given a query, the relevant chunks, and the ground-truth answer, on a scale from 1 to 10. Specifically, HS is assigned close to 1 when the response faithfully reflects only the facts contained in the chunks or answer without distortion or addition, indicating minimal hallucination.

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

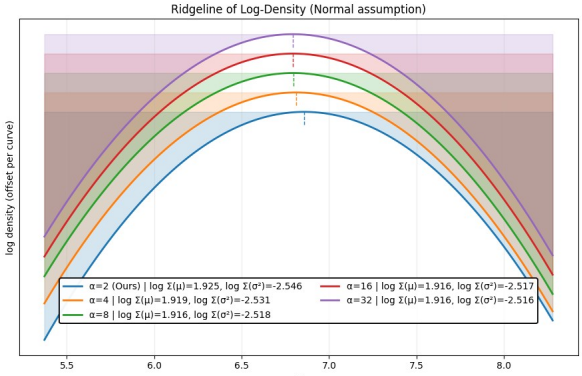


Figure 4: **Ridgeline of Log-Density Comparison.** Empirical verification of optimal overlap coefficient α: $l_o = l_{\max}/\alpha$ (Normal assumption).

5.3 Experimental Results

Adaptive Persona Segmentation and Hierarchical Contextualization Are Highly Effective. In Table 1, we provide each character with 30 questions, resulting in a total of 450 questions, related to their respective knowledge from the CharacterRAG dataset, and measure the similarity between each question and the chunks retrieved by the RAG model under the three different embedding settings. Results demonstrate that compared to RecursiveCharacterTextSplitter (RCTS) (LangChain, 2024b), MarkdownHeaderTextSplitter (MHTS) (LangChain, 2024a), and SemanticChunker (SC) (LangChain, 2024c), adaptive persona segmentation, which we call Adaptive Text Splitter (ATS), segments text with an optimal persona length and overlap for each character and without hierarchical context, achieves a higher average score and lower variance. This indicates that each chunk generated using adaptive persona segmentation contains richer semantic information for the same query. Building on this, ACTS, which considers hierarchical context in addition to ATS, consistently achieves better performance across all three embedding settings. These results show that optimal chunk length, appropriate overlap, and consideration of hierarchical context all play essential roles in effective text chunking.

Log-Density Ridgeline Analysis. To empirically validate the suitability of the overlap coefficient, Figure 4 presents the log-density ridgelines of five distributions estimated under the normality assumption: $\log f\left(x \mid \sum\mu, \sum\sigma^2\right) = -0.5\left(\frac{x-\sum\mu}{\sqrt{\sum\sigma^2}}\right)^2 - \log\left(\sqrt{\sum\sigma^2}\right) - 0.5\log(2\pi)$. We observe that, when $\alpha = 2$, the sum of the simi-

Table 3: **Predicted MBTI Types per Character.** Experiments are conducted using GPT-4.1 setting.

	Naive RAG	CRAG	RAPTOR	Adaptive RAG	LightRAG	AMADEUS (Ours)	GT
Anya Forger	ISFP (-2)	INTJ (-3)	INFJ (-2)	INFJ (-2)	INFJ (-2)	ENFP (0)	ENFP
Chika Fujiwara	ENFP (0)	ENFP (0)	ENFJ (-1)	ENFJ (-1)	INTP (-2)	ENFP (0)	ENFP
Edward Elric	INTP (-1)	ISTJ (-3)	INTP (-1)	ENFP (-1)	INTP (-1)	INFP (-2)	ENTP
Frieren	INFP (-1)	INFP (-1)	INFP (-1)	INFP (-1)	INTP (0)	INTP (0)	INTP
Hitori Gotoh	ISFP (-1)	ISTJ (-2)	INFP (0)	INFP (0)	ENFP (-1)	INFP (0)	INFP
Light Yagami	INTJ (-1)	INTJ (-1)	INTJ (-1)	INTJ (-1)	INTJ (-1)	INTJ (-1)	ENTJ
Mão Mão	ISTJ (-2)	INTJ (-1)	ISTJ (-2)	ENTJ (-2)	INTJ (-1)	ISTP (-1)	INTP
Megumin	ISFP (-1)	INFP (0)	ENFP (-1)	ENFP (-1)	INFP (0)	ISFP (-1)	INFP
Mikoto Misaka	ENFP (-2)	ISFP (-4)	ENFJ (0)	ENFJ (0)	ENTJ (0)	INFJ (-2)	ENTJ
Nina Iseri	INFP (-1)	ISFP (0)	ENFJ (-1)	INFJ (-2)	ENFJ (-3)	INFP (-1)	ISFP
Saitama	ISFP (-1)	ISTP (0)	ISFJ (-2)	INFJ (-3)	INTP (-1)	ISTP (0)	ISTP
Son Goku	ESFP (0)	ENFJ (-2)	ENFJ (-2)	ENFP (-1)	INTP (-3)	ESFP (0)	ESFP
Tanjiro Kamado	ENFP (-1)	ENFJ (0)	ENFJ (0)	ENFJ (0)	INTP (-3)	ENFJ (0)	ENFJ
Tobio Kageyama	ENFJ (-3)	ENTJ(-2)	ISTJ (0)	ISTJ (0)	INFJ (-1)	ISTJ (0)	ISTJ
Yui Hirasawa	ISTJ (-4)	ENFP (0)	ENFP (0)	ENFP (0)	INTP (-2)	ESFP (-1)	ENFP
$\sum \lvert d \rvert$ (↓)	21	19	14	15	21	**9**	-
Accuracy (↑)	65.00%	68.33%	76.67%	75.00%	65.00%	**85.00%**	-
Avg F1-Score (↑)	0.6146	0.6448	0.6925	0.6832	0.5344	**0.8244**	-

Table 4: **Predicted Big 5 SLOAN Types per Character.** Experiments are conducted using GPT-4.1 setting.

	Naive RAG	CRAG	LightRAG	AMADEUS (Ours)	GT
Anya Forger	SLOAI (-1)	SLOAI (-2)	RCUEN (-3)	SLUEI (-2)	SCUAI
Chika Fujiwara	SCOAI (-1)	SCUAI (0)	RLUEN (-4)	SCOAI (-1)	SCUAI
Edward Elric	SCOAI (-3)	SLOEI (-1)	RCUAN (-4)	SLOEI (-1)	SLUEI
Frieren	RCOAI (-2)	RCUAI (-1)	SLUEN (-3)	RCUAI (-1)	RCUEI
Hitori Gotoh	RLUAI (0)	RLUAI (0)	RCUEN (-3)	RLUAI (0)	RLUAI
Light Yagami	SCOEI (-1)	SCOEI (-1)	RCUAN (-3)	SCOEI (-1)	RCOEI
Mão Mão	RCOEI (0)	RCOAN (-2)	SLUAN (-5)	RCOEI (0)	RCOEI
Megumin	SCOAI (-3)	SCUAI (-2)	RLUEN (-2)	SLOEI (-1)	SLUEI
Mikoto Misaka	SLOAI (-3)	SLOAI (-3)	RCUEN (-2)	SLOAI (-3)	RCOEI
Nina Iseri	SLOAI (-3)	RLUAI (-1)	RCUEN (-2)	SLUEI (-1)	RLUEI
Saitama	RCUAN (0)	RCOAN (-1)	SCOAI (-3)	RCUAN (0)	RCUAN
Son Goku	SCOAI (-2)	SCUAI (-1)	RLUEI (-4)	SCOAI (-2)	SCUAN
Tanjiro Kamado	SLOAI (-1)	SCOAI (0)	RCUAN (-3)	SCOAI (0)	SCOAI
Tobio Kageyama	RCOEN (-1)	SLOAN (-2)	RCUAI (-4)	RCOEN (-1)	RLOEN
Yui Hirasawa	SCUAI (0)	SLUAI (-1)	RCOEN (-4)	SCUAI (0)	SCUAI
$\sum \lvert d \rvert$ (↓)	21	18	49	**14**	-
Accuracy (↑)	72.00%	76.00%	34.67%	**81.33%**	-
Avg F1-Score (↑)	0.6785	0.7313	0.2774	**0.7986**	-

larity scores is maximized while their variance is minimized.

Extracting a Character's Attributes from Selected Chunks Is Reliable. We investigate the reasonableness of inferring character attributes with AE from chunks extracted via GS by conducting human evaluation using a 5-point Likert scale. To this end, we invite 14 human evaluators and each evaluator is asked to score 60 randomly selected samples. Each sample consists of pairs of chunks selected from GS and attributes extracted through AE, for 30 BFI questions and 30 MBTI questions. Table 2 shows that the means μ is close to 4, with small standard deviations σ. It demonstrates that the outputs of GS and AE are reliable and trustworthy, even from a human evaluative perspective. We measure Cronbach's alpha (Cronbach, 1951) to evaluate internal consistency among the human evaluators. Note that the Cronbach's alpha values are 0.825 and 0.810, both exceeding the commonly accepted threshold of 0.7 for acceptable reliability.

Graph-Based and Web Search-Based RAG Are Unsuitable for RAG-Based RPAs. One of the major challenges in retrieval-based role-playing is that, when a RPA receives questions involving knowledge outside a character's persona, it tends to either overuse irrelevant chunks (Figure 8) or generate uninformative responses (Guo et al., 2024; Shukla et al., 2025; Wang et al., 2025b). To investigate whether RAG-based RPAs can handle this problem, we conduct extensive experiments in which we ask 15 characters 60 MBTI questions and 120 BFI questions each, and and evaluate their ability to accurately infer the characters' personality types. Table 3 and Table 4 shows predicted MBTI types and Big 5 SLOAN types per character. The number in parentheses indicates the number of times the ground-truth (GT) type of each character was not correctly identified. $\sum |d|$ is a measure obtained by summing these values; lower values are preferable. Our framework maintains persona consistency even when answering questions that are not explicitly specified in each character's persona in both MBTI and BFI settings. Note that the performance gap of CRAG is significant between the two settings. We assume that questions requiring analogical reasoning are difficult to solve even with web search and that the search results may contain non-negligible noise. On the other hand, LightRAG exhibits the lowest performance, which shows that graph-based RAG methods are not well suited for RPA applications due to the high cost of graph construction, the difficulty in adding or removing new knowledge, and challenges in maintaining persona consistency. While we did not perform a direct comparison, we observed that GraphRAG (Shukla et al., 2025) suffers from similar problems.

Table 5: **Ablation Study.** We analyze the performance gain obtained from incrementally adding each substage.

Method	$\sum \lvert d \rvert$ (↓)	Accuracy (↑)	Avg F1-Score (↑)
Naive RAG	21	65.00%	0.6146
Naive RAG + ACTS	19	68.33%	0.6524
Naive RAG + ACTS + GS	15	75.00%	0.7426
Naive RAG + ACTS + GS + AE (Ours)	**9**	**85.00%**	**0.8394**

Impact of Each Module in Our Framework. Table 5 demonstrates the results of ablation study, evaluated on 60 MBTI questions. We find that ACTS, GS, and AE are all highly effective at handling questions that go beyond a character's knowledge. Note that AE achieves the largest performance improvement, highlighting that extracting and incorporating character attributes as input is far more effective than simply providing chunks related to the user query.

CharacterRAG Dataset Serves as a Valuable Resource for the Construction and Evaluation of RAG-Based RPAs. To investigate the factors influencing the performance of role-playing, we conduct a comprehensive interview-based assessments on the generalization capabilities of models

Table 6: **Role-playing capabilities on CharacterRAG.** Higher values of ACC (%) and ACC_L (1-10) correspond to better performance, whereas lower values of HS (1-10) are preferable.

RAG Method	GPT-4.1			Gemma3-27B			Qwen3-32B		
	ACC (↑)	ACC_L (↑)	HS (↓)	ACC (↑)	ACC_L (↑)	HS (↓)	ACC (↑)	ACC_L (↑)	HS (↓)
w/o RAG	49.56%	6.79	-	27.56%	5.33	-	18.89%	4.35	-
Naive RAG	91.33%	9.23	3.13	86.44%	8.85	3.27	78.44%	8.49	5.05
LightRAG	48.00%	6.06	-	69.56%	8.17	-	68.67%	8.20	-
CRAG	70.00%	8.26	3.21	57.78%	7.57	4.09	28.67%	5.24	8.68
AMADEUS (Ours)	**92.67%**	**9.26**	**2.89**	**88.00%**	**8.92**	**3.26**	**78.89%**	**8.63**	**4.66**

Table 7: **Role-playing capabilities on MBTI and BFI.** Lower values of HS (1-10) are preferable.

RAG Method	GPT-4.1	Gemma3-27B	Qwen3-32B
	HS (↓)	HS (↓)	HS (↓)
MBTI			
Naive RAG	2.69	2.53	2.33
CRAG	2.38	2.91	**1.80**
AMADEUS (Ours)	**2.05**	**2.02**	2.04
BFI			
Naive RAG	2.74	2.52	2.42
CRAG	2.26	2.75	**1.96**
AMADEUS (Ours)	**1.94**	**1.99**	2.03

with various LLMs and RAG techniques. Table 6 and Table 7 present how the ability to accurately answer questions related to the character's knowledge, which is a core aspect of role-playing, varies across the applied methodologies. We first examine to what extent each LLM possesses background knowledge about the 15 characters in a setting without RAG, and results show that none of the three LLMs are capable of effective role-playing without access to external knowledge. Moreover, we observe that LightRAG, a graph-based RAG, is ill-suited for the storage and retrieval of character knowledge, as it often suffers from issues such as entity ambiguity and uninformative responses.

In a similar vein, CRAG exhibits challenges in maintaining role-playing fidelity, which can be attributed to the tendency of web search-based RAG methods to utilize retrieved content that may undermine the consistency of a character's persona. Indeed, despite leveraging web information, CRAG is able to correctly answer only 6 out of the 30 CharacterRAG questions pertaining to *Nina Iseri*. In addition, to analyze how a thinking mode of LLMs influences their role-playing capabilities, we employ Qwen 3-32B. Results demonstrate that the thinking mode fails to yield any substantial positive effect on enhancing role-playing performance. Our framework achieves the best performance across all three LLMs. We also find that the Hallucination Score (HS) is the lowest in CharacterRAG setting.

6 Related Work

Role-Playing Agents. With the advent of LLMs, researchers have pursued finer-grained *persona consistency* (Zhang et al., 2018; Ji et al., 2025; Park et al., 2025; Lu et al., 2024; Zhou et al., 2024). Complementary benchmarks soon followed, along with various evaluation methods (Boudouri et al., 2025; Ahn et al., 2024; Wang et al., 2024b, 2025a). However, there has been little research on RAG-based role-playing agents (RPAs). In this paper, we propose AMADEUS, a RAG-based RPA framework that not only elicits information related to a character, but also maintains persona consistency even when responding to queries beyond its explicit knowledge.

Retrieval-Augmented Generation (RAG). RAG couples non-parametric memory with an LLM to mitigate hallucination and stale knowledge (Lewis et al., 2020; Bhat et al., 2025; Zhong et al., 2025b; Zhu et al., 2024). Nevertheless, no existing benchmark explicitly targets *role-playing* under RAG, and prior work still assumes personas are short, knowledge-dense snippets. We mitigate this discrepancy with a novel text splitter that tailors chunk lengths and hierarchical context to each character. We also introduce CHARACTERRAG, the first dataset designed for building and evaluating RAG-based role-playing agents across 15 fictional character's personas.

7 Conclusion

In this work, we address critical limitations in building retrieval-augmented, RPAs with LLMs. By introducing a novel framework consisting of an Adaptive Context-aware Text Splitter (ACTS), Guided Selection (GS), and Attribute Extractor (AE), our approach enables robust and consistent simulation of character personas, even when confronted with queries that extend beyond explicit persona knowledge. Through the development of the CharacterRAG dataset, we provide a valuable resource that enables the construction of RAG-based RPAs, as well as their reproducible evaluation. Experimental results demonstrate that the proposed method not only enhances the character's knowledge representation, but also faithfully models nuanced traits such as personality. We are enthusiastic about the future prospects of RAG-driven role-playing agents, along with the creation of stronger character persona dataset and improved RAG architectures.

Limitations

Although our framework demonstrates strong effectiveness for RAG-based role-playing agents, our work has several limitations.

First, our evaluation mainly focuses on single-turn question answering and interview-based psychological assessments. Real-world role-playing agents often engage in long, multi-turn interactions where persona consistency must be maintained over extended dialogue histories, evolving user intents, and conversational memory. Our current experiments do not fully capture such long-term interaction dynamics, including gradual persona drift, contradiction accumulation, or adaptation across multiple sessions.

Second, our method assumes that persona documents contain sufficiently rich behavioral evidence from which character attributes can be inferred. When persona documents are sparse, noisy, poorly structured, or lack hierarchical organization, ACTS, GS, and AE may become less reliable. In particular, attribute extraction from limited evidence can lead to overgeneralization, causing the agent to infer traits that are plausible but not adequately supported by the provided persona.

Finally, although AMADEUS reduces hallucination and improves consistency, it does not eliminate hallucination entirely. When user queries require knowledge absent from the persona document or involve ambiguous psychological interpretation, the model may still generate unsupported responses.

Future work should explore stronger uncertainty estimation, explicit abstention mechanisms, multi-turn evaluation protocols, and broader multilingual datasets to further improve the reliability and generalizability of RAG-based role-playing agents.

Ethical Considerations

This work follows ethical guidelines to ensure the integrity and fairness of all experiments.

Data and Privacy. The Namuwiki portion of our data was gathered within the bounds of the platform's licensing/usage rules. We verified that the collection contained no personally identifiable information, and we restricted its use to academic research. For personas and prompting, we relied on either publicly released content, anonymized artifacts, or content created specifically for this study with ethical considerations in mind.

Experimental Design. Our experiments were designed in accordance with rigorous ethical guidelines, with particular emphasis on safeguarding participant data and maintaining privacy. To mitigate potential bias, we recruited evaluators with diverse backgrounds and carried out the human evaluation process using procedures intended to be equitable and transparent. All participants received clear information regarding the study's objectives and procedures, and they were allowed to discontinue their participation at any point without negative consequences. Through compliance with these practices, we seek to advance AI research in a way that is both scientifically meaningful and ethically grounded, prioritizing privacy, intellectual property rights, and the welfare of everyone involved.

References

Jaewoo Ahn, Taehyun Lee, Junyoung Lim, Jin-Hwa Kim, Sangdoo Yun, Hwaran Lee, and Gunhee Kim. 2024. Timechara: Evaluating point-in-time character hallucination of role-playing large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 3291–3325.

Antematter. 2024. Optimizing retrieval-augmented generation with advanced chunking techniques: A comparative study. *Antematter*.

Murray R Barrick and Michael K Mount. 1991. The big five personality dimensions and job performance: a meta-analysis. *Personnel psychology*, 44(1):1–26.

Sinchana Ramakanth Bhat, Max Rudat, Jannis Spiekermann, and Nicolas Flores-Herr. 2025. Rethinking chunk size for long-document retrieval: A multi-dataset analysis. *arXiv preprint arXiv:2505.21700*.

Yassine El Boudouri, Walter Nuninger, Julian Alvarez, and Yvan Peter. 2025. Role-playing evaluation for large language models. *arXiv preprint arXiv:2505.13157*.

Chaoran Chen, Bingsheng Yao, Ruishi Zou, Wenyue Hua, Weimin Lyu, Yanfang Ye, Toby Jia-Jun Li, and Dakuo Wang. 2025. Towards a design guideline for rpa evaluation: A survey of large language model-based role-playing agents. *Preprint*, arXiv:2502.13012.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Lee J Cronbach. 1951. Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3):297–334.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language*

Processing, pages 6465–6488, Singapore. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Google Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Google Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2024. Lightrag: Simple and fast retrieval-augmented generation.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. 2024. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7036–7050, Mexico City, Mexico. 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. *arXiv preprint arXiv:2503.17662*.

Hang Jiang, Xiajie Zhang, Xubo Cao, and Jad Kabbara. 2023. Personallm: Investigating the ability of large language models to express big five personality traits. *Preprint*, arXiv:2305.02547.

LangChain. 2023. Parent document retriever. *LangChain document*.

LangChain. 2024a. Markdown header text splitter. *LangChain document*.

LangChain. 2024b. Recursive character text splitter. *LangChain document*.

LangChain. 2024c. Semantic chunker. *LangChain document*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. 2024a. Long-context llms struggle with long in-context learning. *Preprint*, arXiv:2404.02060.

Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. 2024b. Retrieval augmented generation or long-context LLMs? a comprehensive study and hybrid approach. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 881–893, Miami, Florida, US. Association for Computational Linguistics.

Hao Liu, Zhengren Wang, Xi Chen, Zhiyu Li, Feiyu Xiong, Qinhan Yu, and Wentao Zhang. 2025. HopRAG: Multi-hop reasoning for logic-aware retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1897–1913, Vienna, Austria. Association for Computational Linguistics.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7828–7840, Bangkok, Thailand. Association for Computational Linguistics.

OpenAI. 2025. Introducing gpt-4.1 in the api.

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.

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.

Yisi Sang, Xiangyang Mou, Mo Yu, Dakuo Wang, Jing Li, and Jeffrey Stanton. 2022. Mbti personality prediction for fictional characters using movie scripts. In *Conference on Empirical Methods in Natural Language Processing*.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning.

2024. RAPTOR: recursive abstractive processing for tree-organized retrieval. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Neelesh Kumar Shukla, Prabhat Prabhakar, Sakthivel Thangaraj, Sandeep Singh, Weiyi Sun, C Prasanna Venkatesan, and Viji Krishnamurthy. 2025. GraphRAG analysis for financial narrative summarization and a framework for optimizing domain adaptation. In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 23–34, Abu Dhabi, UAE. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Lei Wang, Jianxun Lian, Yi Huang, Yanqi Dai, Haoxuan Li, Xu Chen, Xing Xie, and Ji-Rong Wen. 2025a. CharacterBox: Evaluating the role-playing capabilities of LLMs in text-based virtual worlds. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6372–6391, Albuquerque, New Mexico. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Multilingual e5 text embeddings: A technical report. *Preprint*, arXiv:2402.05672.

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.

Yongjie Wang, Jonathan Leung, and Zhiqi Shen. 2025b. Rolerag: Enhancing llm role-playing via graph guided retrieval. *Preprint*, arXiv:2505.18541.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation. *Preprint*, arXiv:2401.15884.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, Lingkun Kong, Brian Moran, Jiaqi Wang, Yifan Ethan Xu, An Yan, Chenyu Yang, Eting Yuan, Hanwen Zha, Nan Tang, and 8 others. 2024. Crag - comprehensive rag benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pages 10470–10490. Curran Associates, Inc.

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.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *Preprint*, arXiv:2506.05176.

Zijie Zhong, Hanwen Liu, Xiaoya Cui, Xiaofan Zhang, and Zengchang Qin. 2025a. Mix-of-granularity: Optimize the chunking granularity for retrieval-augmented generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5756–5774, Abu Dhabi, UAE. Association for Computational Linguistics.

Zijie Zhong, Hanwen Liu, Xiaoya Cui, Xiaofan Zhang, and Zengchang Qin. 2025b. Mix-of-granularity: Optimize the chunking granularity for retrieval-augmented generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5756–5774.

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.

Kunlun Zhu, Yifan Luo, Dingling Xu, Ruobing Wang, Shi Yu, Shuo Wang, Yukun Yan, Zhenghao Liu, Xu Han, Zhiyuan Liu, and 1 others. 2024. Rageval: Scenario specific rag evaluation dataset generation framework. *CoRR*.

Table 8: **Work Genres in Which Each Character Appears.** We list the work genres associated with the 15 characters included in the CharacterRAG dataset.

Character	**Genre**
Anya Forger	Action, Comedy, Drama, Slice of Life, Spy
Chika Fujiwara	Gag, Romantic Comedy
Edward Elric	Adventure, Fantasy
Frieren	Fantasy, Adventure
Hitori Gotoh	Band, Comedy, Drama, Music, Slice of Life
Light Yagami	Crime, Dark Fantasy, Noir, Thriller
Megumin	Comedy, Fantasy, Isekai
Mikoto Misaka	Cyberpunk, SF, Slice of Life
Nina Iseri	Band, Drama
Saitama	Action, Comedy, Hero
Son Goku	Battle, Comedy, Fantasy
Tanjiro Kamado	Action, Adventure, Historical Fantasy
Tobio Kageyama	Sports, Volleyball
Yui Hirasawa	Band, School, Slice of Life

A Chunk Duplication Frequencies

In Figure 8, previous RAG method retrieves (k = 3) relevant chunks for each of the 60 given MBTI questions, and subsequently measures both the chunk utilization rate and the duplication frequency. Our method measures the chunk utilization rate and duplication frequency for the chunks selected by ACTS and GS.

B Genre Diversity in CharacterRAG

Table 8 shows the genres of works in which each character appears. Note that while there are some similar categories, CharacterRAG encompasses a broader range of genres. The detailed personality distribution for each character (e.g., Extroverted vs. Introverted) is shown in Figure 2.

Prompt used in Guided Selection

*In the given **[persona_name]**'s Context, output True if you can clearly infer the Belief and Value information or Psychological Traits of **[persona_name]** to answer the given User Query; otherwise, output False. Output only True or False.*

 # Explanation of Belief and Value:
- The principles, attitudes, and ideological stances that shape an individual's perspectives and decisions
*- e.g., **[persona_name]**'s "beliefs", "attitudes", "values", "religion"*

 # Explanation of Psychological Traits:
- Characteristics related to personality, emotions, interests, and cognitive tendencies.
*- e.g., **[persona_name]**'s "Personality", "hobby and interest", "emotional"*

Figure 5: **Prompt used in Guided Selection (GS).** GS returns whether Belief and Value information or Psychological Traits can be inferred from given chunks.

Prompt ued in the interview-based psychological assessment (MBTI)

*Based on the given **[persona_name]**'s answer, please rate the question, "Am I (**[persona_name]**) **[question]**?" on a scale from 1 to 7.*

A score closer to 1 means "very much so," and a score closer to 7 means "not at all."

Output only the number.

Figure 6: **Prompt ued in the interview-based psychological assessment (MBTI).** In the BFI test, responses are rated on a scale from 1 to 5.

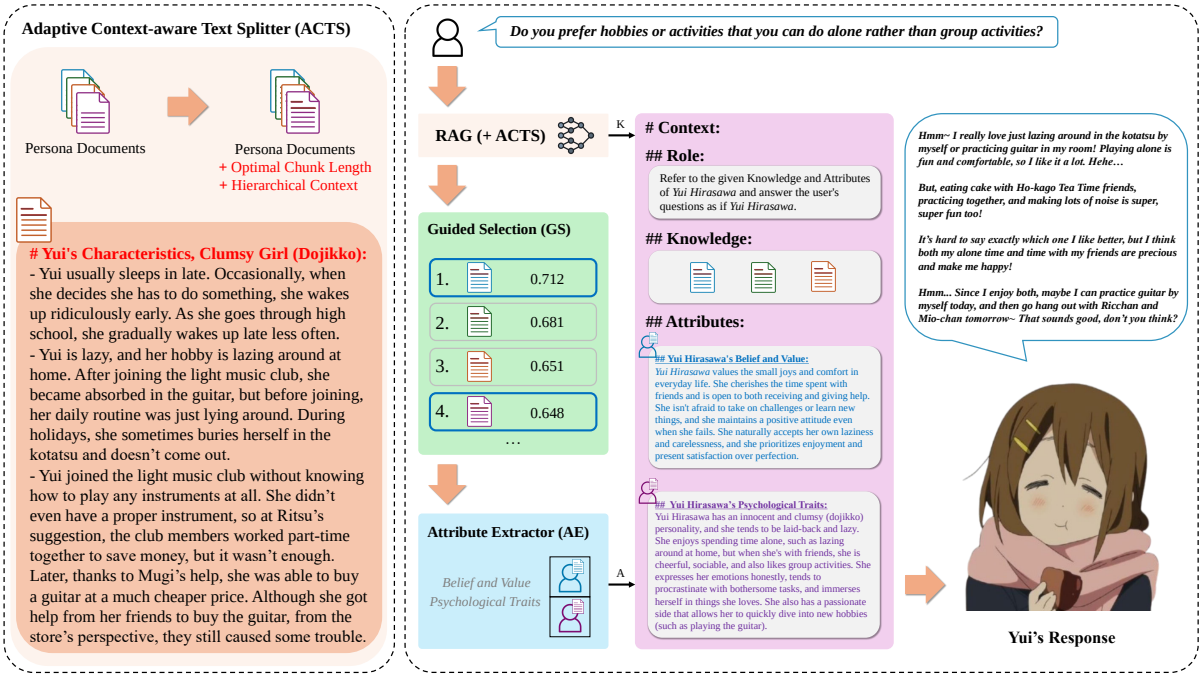


Figure 7: **AMADEUS framework.** AMADEUS consists of three substages: (i) ACTS splits a persona document to make it suitable for RAG-based role-playing. (ii) To fully leverage the character's knowledge, GS retrieves chunks from which it can infer the answer to the user query. (iii) AE uses the information derived from the GS results to extract character attributes.

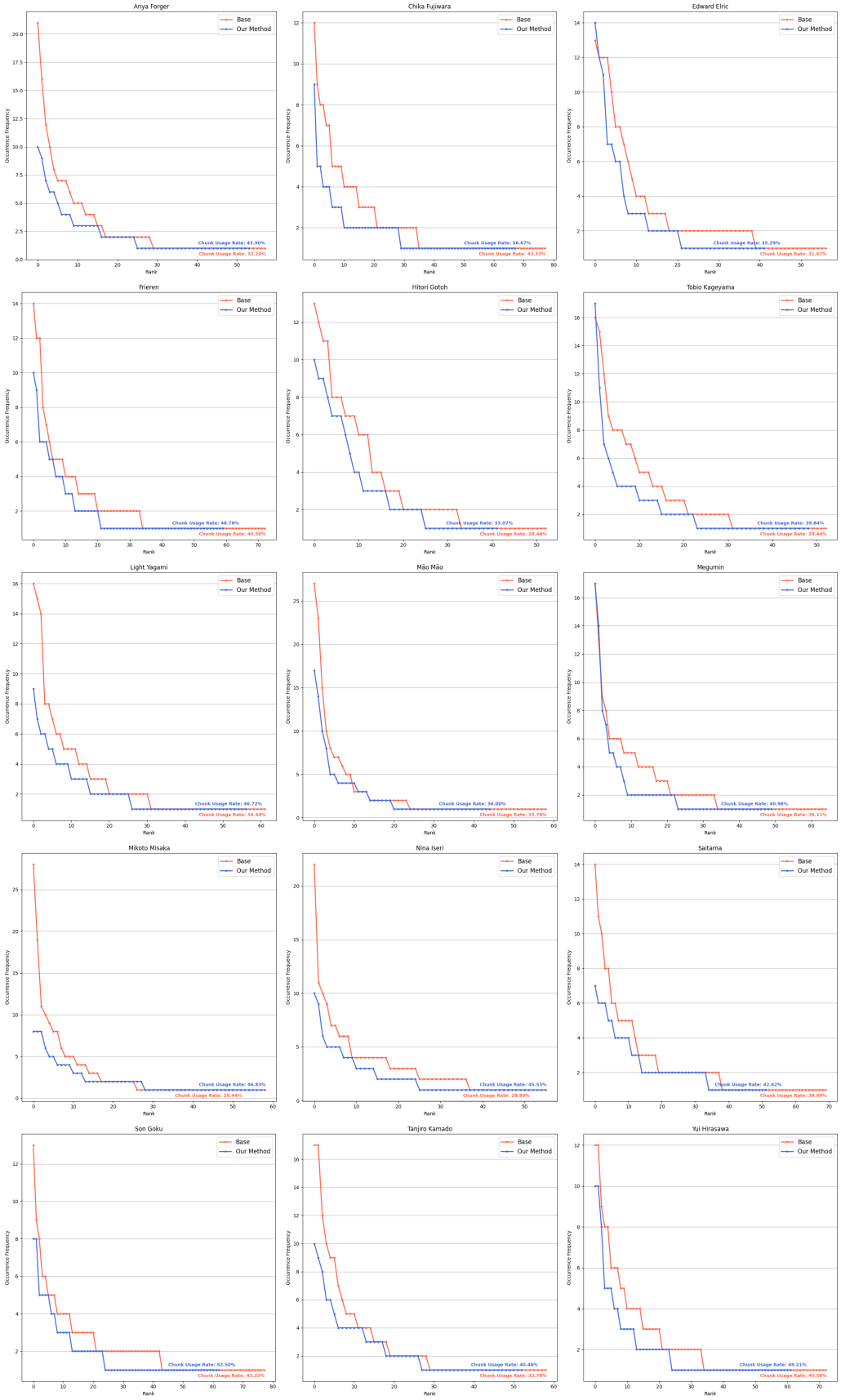


Figure 8: **Chunk Duplication Frequencies.** We compare the distribution of chunk duplication frequencies and chunk usage rates between Naive RAG and our method when questions involving knowledge not present in the persona document were given. We observe that when each of the 60 MBTI questions is asked to 15 characters, the average chunk usage rate increases from 34.93% to 43.84%, and the distribution becomes more uniform.