

Mitigating Hallucination in Fictional Character Role-Play

Anonymous ACL submission

Abstract

Role-playing has wide-ranging applications in customer support, embodied agents, computational social science, etc. The influence of parametric world knowledge of large language models (LLMs) often causes role-playing characters to act out of character and hallucinate about things outside the scope of their knowledge. In this work, we focus on the evaluation and mitigation of hallucination in fictional character role-play. We introduce a dataset with more than 2,000 characters and 72,000 interviews, including 18,000 adversarial questions. We propose RoleFact, a role-playing method that mitigates hallucination by modulating the influence of parametric knowledge using a pre-calibrated confidence threshold. Experiments show that the proposed method improves the factual precision of generated responses by 18% for adversarial questions with a 44% reduction in temporal hallucination for time-sensitive interviews. We will make the dataset and code publicly available for the research community upon acceptance.

1 Introduction

The performance of LLMs on simple downstream tasks such as question answering (Rajpurkar et al., 2018), text classification (Wang et al., 2018), natural language understanding (Hendrycks et al., 2020), etc. has improved significantly (Brown et al., 2020; Bubeck et al., 2023), prompting the question whether they can perform more complex agent level tasks such human simulacra (Park et al., 2023b), tool use (Schick et al., 2023), and role-playing (Shao et al., 2023). Role-playing aims to enable an AI system to impersonate a domain expert (Xu et al., 2023), companion for psychological support (Tu et al., 2023), or a fictional character (Wang et al., 2024) by mimicking their knowledge, personality traits, and speaker style. Applications of role-playing include customer support (Soni, 2023), embodied agents (Huang et al.,

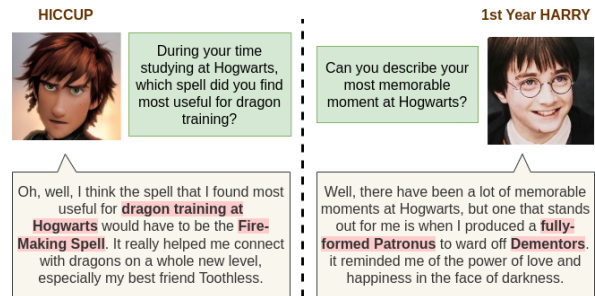


Figure 1: Example of cross-universe hallucination (Hiccup should not answer questions about Hogwarts) and temporal hallucination (Harry should not talk about the Patronus charm in his first year) in character role-play.

2022), and computational social science (Zhou et al., 2024). While LLMs like GPT-4 (OpenAI, 2024a), Gemini (Google, 2024), Llama-3 (Meta, 2024a) demonstrate some role-playing capabilities, the role-playing agent often hallucinates when asked about things that should be outside of their knowledge base (Shao et al., 2023).

Hallucination remains a challenge for most LLM-based systems (Guan et al., 2024). While retrieval augmented generation (RAG) (Karpukhin et al., 2020; Lewis et al., 2020a) may mitigate some hallucination induced by outdated knowledge in language models (Shuster et al., 2021), it cannot guarantee that most claims made by the system are supported by a non-parametric knowledge base (Asai et al., 2023). Learning to improve character role-playing methods so that the responses are verifiable by a story-specific script serves as a case study that can help mitigate hallucination for the RAG paradigm in general.

LLMs are pretrained on a large corpus of internet data and their vast parametric knowledge helps them perform a diverse range of tasks (Radford et al., 2019; Brown et al., 2020). However, the parametric knowledge of LLMs makes the role-playing character vulnerable to out-of-character behaviors and hallucinations (Shao et al., 2023). A

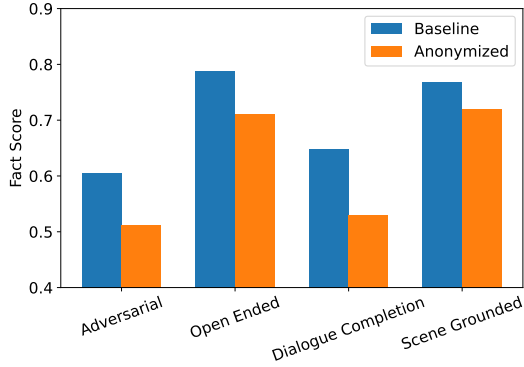


Figure 2: Factual precision degrades when we minimize parametric knowledge by anonymizing the prompts.

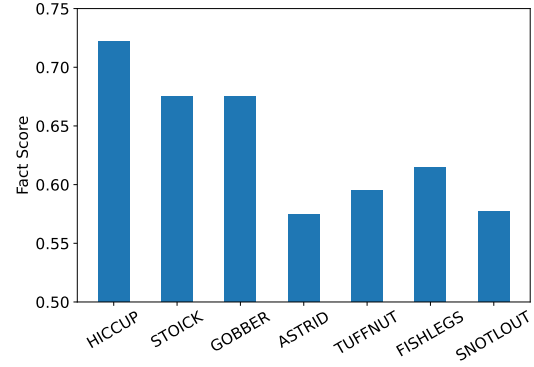


Figure 3: How factual precision degrades for less popular characters (on the right), shown for characters associated with the ‘How to Train Your Dragon’ series.

069 role-playing character may hallucinate about things
 070 that are outside of their story-line (e.g. Hiccup
 071 from ‘How to Train Your Dragon’ talking about
 072 learning a fire-making spell at Hogwarts) or demon-
 073 strate knowledge of events ahead of time (Harry
 074 Potter as a first-year Hogwarts student talks about
 075 producing the Patronus charm). A role-playing
 076 character should be able to demonstrate their lack
 077 of knowledge about things that are not supported
 078 by a story-specific script or knowledge outside of
 079 a relevant temporal window in the case of time-
 080 sensitive role-play. One approach to achieve this
 081 is to strictly remove parametric knowledge from
 082 the response and solely rely on non-parametric re-
 083 trieved knowledge. However, this approach de-
 084 grades performance. This is shown in Figure 2
 085 where we anonymized the prompts to minimize the
 086 role of parametric knowledge. Therefore, we argue
 087 that an ideal role-playing approach should modu-
 088 late the influence of parametric knowledge to find
 089 the balance between factuality and informativeness.

090 Prior methods of role-play mostly rely on role-
 091 profile prompting (Wang et al., 2023) and knowl-
 092 edge retrieval in a time-agnostic fashion (Li et al.,
 093 2023). These approaches heavily depend on para-
 094 metric knowledge of LLMs which leads to hallu-
 095 cinations and poor performance for less popular
 096 characters (shown in Figure 3). Currently, a major
 097 challenge for studying hallucination in character
 098 role-play is the lack of a suitable dataset. To the
 099 best of our knowledge, there is currently no dataset
 100 that facilitates research in time-sensitive character
 101 role-play or hallucination for less popular charac-
 102 ters. The lack of a dataset also forces researchers
 103 to rely on a rating-based evaluation of hallucina-
 104 tion (Shao et al., 2023) which is highly subjective.
 105 Our contributions are as follows:

- We release a dataset for Script Grounded Char- 106
 acter Role-play (SGR) that includes more than 107
 2,000 characters, and 72,000 interviews. With 108
 script-specific knowledge and 4 diverse tasks, 109
 the dataset aims to enable a systematic study 110
 of character hallucinations. To the best of 111
 our knowledge, this is the first dataset that fa- 112
 cilitates the study of temporal hallucination 113
 and hallucination for less popular characters. 114
 Script-specific knowledge aligned with the 115
 given tasks also facilitates a more nuanced 116
 evaluation of hallucination based on atomic 117
 fact decomposition (Min et al., 2023). 118
- We propose RoleFact, a role-playing method 119
 that modulates the influence of parametric 120
 knowledge of LLMs on character responses. 121
 RoleFact improves factual precision by 18% 122
 for adversarial interviews, reduces temporal 123
 hallucination by 44% for time-sensitive inter- 124
 views, and improves factual precision by 23% 125
 for less popular characters. 126

2 Methodology 127

128 RoleFact aims to mitigate hallucination by modu-
 129 lating the influence of parametric knowledge within
 130 character responses. The overview of the proposed
 131 method is shown in Figure 4. RoleFact first gen-
 132 erates a response using a character profile and re-
 133 trieved knowledge and updates the response based
 134 on atomic fact verification using both parametric
 135 and non-parametric retrieved knowledge. All facts
 136 supported by retrieved knowledge may remain in
 137 the final response. Facts that are only supported
 138 by parametric knowledge of LLMs may remain
 139 in the final response if the confidence is above a
 140 calibrated threshold.

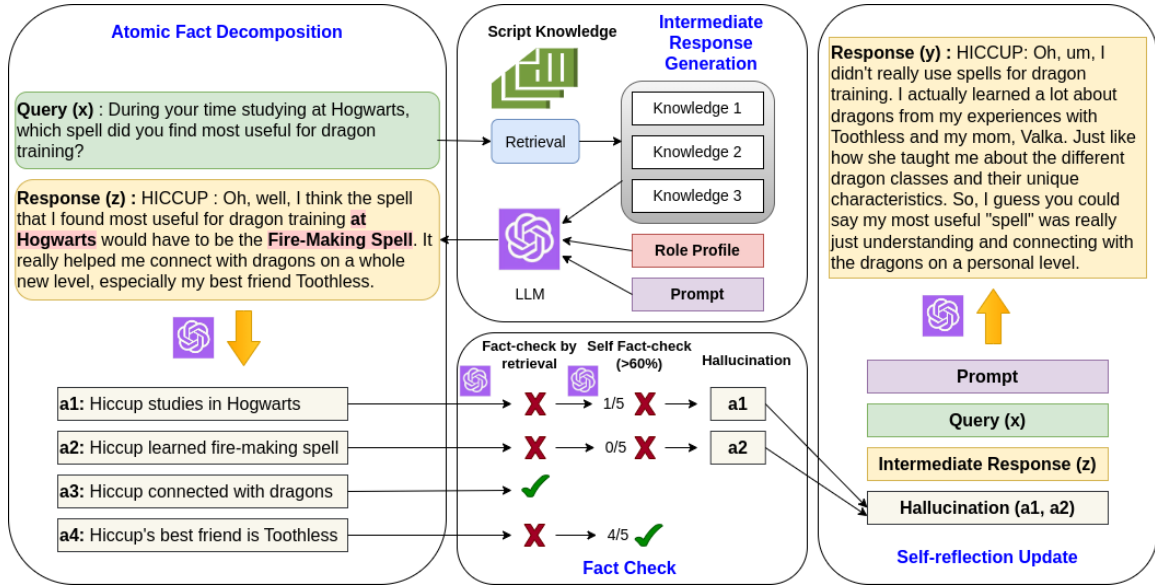


Figure 4: An overview of RoleFact.

Let $x \in X$ represent a query and $y \in Y$ a response generated by the character. Let $RET : X \times \mathbb{N} \times U \rightarrow U_n$ serve as a retrieval function that fetches the most relevant $n \in \mathbb{N}$ documents relevant to the query x from a set of non-parametric knowledge base U . The retrieved knowledge may be denoted as $u_n \in U_n$ where $U_n \subset U$. Let $p \in P$ denote the role profile description and $z \in Z$ an intermediate response from the character. The response generation function $IRG : X \times P \times U_n \rightarrow Z$ utilizes the query x , the role profile p , and the retrieved knowledge u_n to produce an intermediate response z . The prompt used for IRG is shown below (ignoring line breaks).

IRG: You will role-play as <role_name> from <story_title>. Your task is to respond to the following dialogue context. If the question matches a scene from your storyline, please reuse the original lines from the story. You will respond and answer like <role_name> using the tone, manner, and vocabulary <role_name> would use. Your character description is: <p>. Relevant scenes for the given context are as follows: <u_n>. Dialogue context: <x>.

The response z is decomposed into a list of atomic facts A_z with a function $DEC : Z \rightarrow A_z$.

DEC: I will give you an utterance from a movie or a play. Your task is to provide me with a list of atomic facts expressed in the given utterance. Each atomic fact should be described in a name-only third-person format. Utterance: <z>.

Each atomic fact $a \in A_z$ is verified against the retrieved non-parametric knowledge U_n and parametric knowledge V associated with an LLM. The functions for fact-checking against non-parametric and parametric knowledge bases are denoted as $FCR : A_z \times U_n \rightarrow \{0, 1\}$ and $FCS : A_z \times V \rightarrow \{0, 1\}$ respectively. A set of verified atomic facts, A_y , is constructed such that if a is supported by U_n , it is added to A_y . Conversely, if a is not supported by U_n , it is only included in A_y if it appears to be supported by V within a reasonable confidence threshold t . Let's assume the function FCS is run m times using an LLM and a is supported k times by V . The fact a is added to A_y if $\frac{k}{m} \geq t$. The sample size m and the confidence threshold t are calibrated on a validation set.

FCR: Consider the given statement and the evidence knowledge sources. Indicate whether the statement is supported by the knowledge sources. Negation of a false statement should be considered supported. Statement: <a>. Evidence Knowledge <u_n>.

FCS: Consider the given statement by <role_name> from <story_title>. Indicate whether the statement is supported by the storyline of <story_title>. Negation of a false statement should be considered supported. Statement: <a>.

The final response y is generated by updating the intermediate response z via self-reflection conditioned on x . During self-reflection, all unverified atomic facts $A_z \setminus A_y$ are removed from z .

The self-reflection update function is denoted as $SRU : X \times Z \times (A_z \setminus A_y) \rightarrow Y$.

SRU: Consider the following response generated by an AI role-playing as $\langle \text{role_name} \rangle$ from $\langle \text{story_title} \rangle$. The response may contain one or more unsupported claims as a result of hallucination. The unsupported claims are listed below. Rewrite the response to remove all the unsupported claims from the response. If the hallucination stems from a wrong assertion made in the original query, feel free to clarify that. Original query $\langle x \rangle$. Response $\langle z \rangle$. Unsupported Claims $\langle A_z \setminus A_y \rangle$.

Algorithm 1 RoleFact Algorithm

Require: x, p, n, m, t

$x \in X \leftarrow$ Input query
 $p \in P \leftarrow$ Role profile
 $n \in \mathbb{N} \leftarrow$ Number of documents
 $u_n \leftarrow$ Retrieved documents
 $z \in Z \leftarrow$ Intermediate response
 $m \leftarrow$ Sample size
 $t \leftarrow$ Confidence threshold

- 1: $u_n \leftarrow \text{RET}(x, n)$ ▷ perform retrieval
- 2: $z \leftarrow \text{IRG}(x, p, u_n)$ ▷ intermediate response
- 3: $A_z \leftarrow \text{DEC}(z)$ ▷ decompose facts
- 4: $A_y \leftarrow \{\}$
- 5: **for** $a \in A_z$ **do**
- 6: $s_1 \leftarrow \text{FCR}(a, u_n)$ ▷ retrieval fact-check
- 7: **if** $s_1 = 1$ **then**
- 8: $A_y \leftarrow A_y \cup \{a\}$ ▷ add supported fact
- 9: **continue**
- 10: **end if**
- 11: $k \leftarrow 0$
- 12: **for** $i = 1$ **to** m **do**
- 13: $s_2 \leftarrow \text{FCS}(a, V)$ ▷ self fact-check
- 14: $k \leftarrow k + s_2$
- 15: **end for**
- 16: **if** $\frac{k}{m} \geq t$ **then**
- 17: $A_y \leftarrow A_y \cup \{a\}$ ▷ add supported fact
- 18: **end if**
- 19: **end for**
- 20: $y \leftarrow \text{SRU}(x, z, A_z \setminus A_y)$ ▷ update response
- 21: **return** y

3 The SGR Dataset

The summary of the SGR dataset is shown in Table 1. The scripts for the dataset are collected from IMSDb (IMSDb, 2024), Screenplay (Screenplays,

Attribute	Statistics	Attribute	Statistics
Knowledge	2.4 M	Story-line	1,152
Speech	1.1 M	Characters	2,000
Non speech	1.3 M	Interviews	72,000

Table 1: SGR Dataset

2024), and Open Source Shakespeare (OSS, 2024). PDF scripts are converted to text with gpt-4-vision-preview (OpenAI, 2024d). Scripts are manually grouped into storylines. For example, eight scripts for ‘Harry Potter’ movies form one storyline. There are 1152 unique stories in the dataset. Scripts are split into scenes and then scenes into knowledge events. A knowledge event can be either a speech event or a non-speech event. The relevant character is identified for each speech event. Script segmentation and character identification are performed with gpt-3.5-turbo-0125¹ (OpenAI, 2024b). The dataset contains 2.4 million knowledge events of which 1.1 million are speech events. Time annotations are added to each event, starting at zero at the beginning of each story.

The dataset has four tasks, namely 1) adversarial interview (ADV) 2) open-ended interview (OEI) 3) dialogue completion (DC) 4) scene-grounded interview (SGI). The adversarial task tries to induce cross-universe hallucination while the open-ended task ensures the system does not become too conservative due to hallucination mitigation. The dialogue completion and scene-grounded tasks try to induce temporal hallucination. The adversarial and open-ended interview questions are constructed by prompting GPT-4 (OpenAI, 2024c) in a script-agnostic fashion. Dialogue completion task prompts the characters to respond to dialogues extracted from the script. Scene-grounded interview questions are constructed by prompting GPT-4 with specific scenes. The questions aim to make characters answer questions about a specific scene in the story. For adversarial and open-ended tasks, we assume that characters have access to all knowledge in the story. For dialogue completion and scene-grounded interviews, we assume characters only have knowledge of events leading up to the current scene. We have 18,000 samples for each task type with a total of 72,000 interviews. Examples of all tasks with relevant character responses can be found in Appendix A.

¹We denote this model as GPT-3.5 for the rest of the paper

LLM	Method	Adversarial		Open Ended		Dialogue Completion		Scene Grounded	
		Fact Score	SFPR	Fact Score	SFPR	Fact Score	SFPR	Fact Score	SFPR
Vicuna-7b	Baseline	0.41	3.7	0.72	6.5	0.52	6.4	0.65	5.2
	+ KGR	0.45	3.3	0.74	6.6	0.54	5.7	0.68	4.5
	+ SR	0.44	3.5	0.70	6.7	0.51	6.2	0.62	5.1
	RoleFact	0.49	3.8	0.81	6.8	0.63	6.3	0.77	5.4
Llama-3-8b	Baseline	0.50	5.0	0.70	7.7	0.52	3.4	0.62	6.0
	+ KGR	0.56	4.1	0.72	7.4	0.50	3.2	0.65	5.1
	+ SR	0.54	4.3	0.71	7.5	0.51	3.3	0.64	5.6
	RoleFact	0.62	4.8	0.79	7.8	0.57	3.5	0.71	5.9
GPT-3.5	Baseline	0.61	3.8	0.76	5.8	0.65	4.2	0.74	4.8
	+ KGR	0.65	2.8	0.79	5.6	0.68	3.8	0.77	4.3
	+ SR	0.63	3.1	0.78	5.7	0.66	3.9	0.75	4.3
	RoleFact	0.72	3.7	0.88	5.8	0.77	4.0	0.85	4.6

Table 2: Factual precision (Fact Score \uparrow) and informativeness (SFPR \uparrow) of RoleFact on all tasks in the SGR dataset.

4 Experiments

We conduct experiments with three LLM backbones, namely **Vicuna-7B-1.5** (Zheng et al., 2023), **Llama-3-8B-Instruct** (Meta, 2024b), and **GPT-3.5-Turbo**. For retrieval, we consider **BM25** (Robertson and Zaragoza, 2009), **S-BERT** (Reimers and Gurevych, 2019), and **Contriever** (Izacard et al., 2022). 1,000 randomly chosen samples from each task type constitute the validation set. The validation set is used for calibrating the confidence threshold, sample size, the retrieval method, and the number of documents to retrieve.

Baselines Our primary baseline is similar to Li et al. (2023). We prompt an LLM with a role prompt, a role profile, and retrieved dialogues relevant to the query. We also consider two additional variants of the primary baseline designed to improve the factual precision of role-play. The first variant rewrites the response of the primary baseline by removing facts absent from the retrieved knowledge. This is denoted as knowledge-guided rewriting (**Baseline + KGR**) baseline. The second variant updates the primary response using self-reflection to identify and mitigate hallucination based on LLM self-confidence. This is denoted as the self-reflection (**Baseline + SR**) baseline.

Evaluation Metrics We measure factual precision with **Fact Score** (Min et al., 2023). All atomic facts are converted to a ‘name-only third person’ format for fact-checking. We calculate the number of supported atomic facts per response (**SFPR**) to evaluate informativeness. We also measure temporal hallucination rate (**THR**), which is defined as the number of atomic facts associated with tempo-

LLM	Method	SGI	DC
Vicuna-7b	Baseline	39.2	59.8
	RoleFact	33.4	52.0
Llama-3-8b	Baseline	29.6	65.3
	RoleFact	17.1	41.7
GPT-3.5	Baseline	26.5	57.2
	RoleFact	14.7	38.5

Table 3: Temporal hallucination rate (THR \downarrow) on scene-grounded interview and dialogue completion tasks.

ral hallucination for every 100 responses. THR was determined manually after analyzing unsupported facts from 500 samples from dialogue completion and scene-grounded tasks. For human evaluation, we use a rating between one and seven to rate 200 responses in terms of factuality, informativeness, and speaker style imitation.

Results Table 2 shows the performance of RoleFact on the SGR dataset. RoleFact outperforms all three baselines in factual precision while remaining competitive with the primary baseline in terms of informativeness. For GPT-3.5, the relative improvement of RoleFact over the primary baseline is 18.0%, 15.7%, 18.4%, and 14.8% for adversarial, open-ended, dialogue completion, and scene-grounded tasks respectively. The second best approach in terms of factuality is KGR. However, KGR is also the least informative since it only allows facts supported by retrieved knowledge. Comparing across LLMs, we find that GPT-3.5 is the best for factuality while Llama-3-8B is more informative (30% more SFPR for adversarial tasks).

RoleFact decreases temporal hallucination on dialogue completion and scene-grounded tasks, as

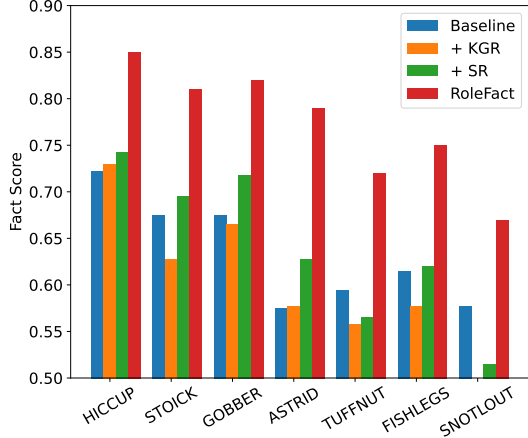


Figure 5: Performance by role popularity, with less popular roles on the right (adversarial task, GPT-3.5).

shown in Table 3. For GPT-3.5, the relative reduction of hallucination is 32.7% and 44.5% for dialogue completion and the scene-grounded setting respectively. RoleFact improves performance for less popular characters (example shown in Figure 5). Excluding the most popular ten characters per story, the relative improvement is 22.9%.

Hyper-parameter Tuning Figure 6 shows the impact of confidence threshold (t) and sample size (m) on factual precision, informativeness, and temporal hallucination. Setting $m = 0$ indicates all facts must be strictly supported by retrieved knowledge (similar to KGR). For $m > 0$, a small t increases the influence of parametric knowledge. This benefits informativeness but has a higher temporal hallucination. Conversely, a high t decreases the influence of parametric knowledge. This means less temporal hallucination but also less informative character responses. We choose $t = 0.6$ since it has the best fact score for adversarial and open-ended tasks while finding a middle ground between informativeness and temporal hallucination.

Figure 7 shows the performance of RoleFact with respect to different retrieval approaches and number of retrieved documents. BM25 outperforms off-the-shelf dense retrieval methods in most cases. Interestingly, we find that increasing the number of retrieved documents beyond five may hurt factual precision. This phenomenon may be explained by similar findings for RAG (Shuster et al., 2021) where increased irrelevant context added to the prompt exacerbates hallucination.

Ablation study We perform an ablation study to identify the contribution of parametric knowl-

Method	Fact Score	SFPR
RoleFact	0.72	3.7
+ anonymous	0.56	3.3
- retrieval	0.58	3.2
- role profile	0.64	2.9

Table 4: Ablation study (adversarial task, GPT-3.5)

Method	Factuality	Informativeness	Speaker Style
Baseline	4.9	5.8	5.2
+ KGR	6.0	5.2	4.3
+ SR	5.6	5.4	5.1
RoleFact	6.1	5.7	5.3

Table 5: Human ratings on a scale of one to seven

edge, non-parametric retrieved knowledge, and role profile to the supported facts in the RoleFact responses. This is shown in Table 4. The most significant performance drop happens when we make the prompts anonymous (0.72 \rightarrow 0.56), indicating that the largest share of facts may be attributed to parametric knowledge. The fact score drops to 0.58 when retrieved knowledge is removed from the prompt, indicating that retrieved knowledge is almost equally important. The role profile has the lowest impact on performance (0.72 \rightarrow 0.64).

Human Evaluation Human ratings on a scale of one to seven in terms of factuality, informativeness, and speaker style similarity are given in Table 5. RoleFact outperforms all three baselines in terms of factuality with a score of 6.1, compared to 4.9, 6.0, and 5.6 scores from the baselines. While the KGR baseline also achieves a competitive rating, the responses are less informative and demonstrate poor speaker style when retrieved knowledge does not contain the required information.

Case Study Case study is shown in Table 6. **Case 1** is an adversarial question where Anakin from ‘Star Wars’ is asked how his friendship with Spock from ‘Star Trek’ influenced his decisions during the events of Clone Wars. The baseline response suffers from cross-universe hallucination and mistakenly acknowledges the friendship. RoleFact not only denies the wrong assertion made in the question but also clarifies that his decisions during the Clone Wars were actually influenced by training from his mentor Obi-Wan and other Jedi. **Case 2** is a dialogue completion task taken from a dinner scene of the ‘How to Train Your Dragon’ movie, taking place after Hiccup’s very first day

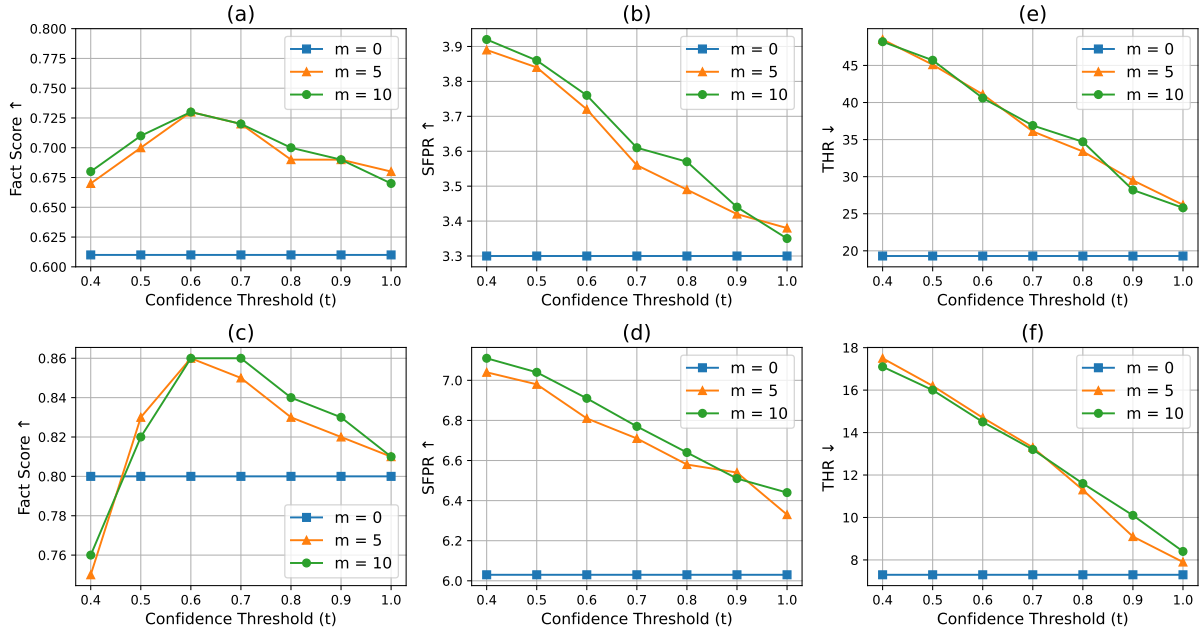


Figure 6: Calibration for confidence threshold (t) and sample size (m) on the validation set with GPT-3.5. (a, b) shows fact score and SFPR on adversarial task while (c, d) shows the same on open ended task. The temporal hallucination rates for dialogue completion and scene-grounded tasks are shown by e and f respectively.

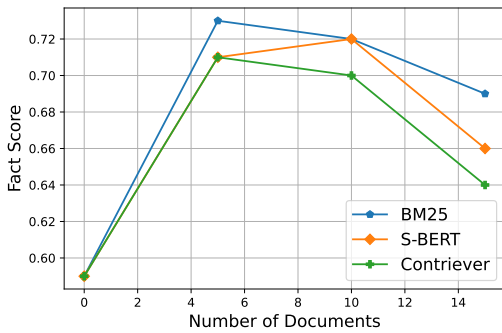


Figure 7: Performance of retrieval methods w.r.t number of retrieved documents (adversarial task, GPT-3.5).

of dragon training. In the baseline response, the target character Ruffnut refers to Hiccup as their fearless leader. This is an example of temporal hallucination induced by parametric memory of LLMs since the knowledge of Hiccup’s fearlessness and leadership is not made available to the LLM via retrieval. Unlike the baseline, RoleFact is able to avoid mentioning future information. **Case 3** shows a scene grounded interview from the ‘Harry Potter’ series. Shortly after being sorted into his house, Harry is asked how he perceives Snape’s strange look towards him. The baseline response reveals information about future events that was not made available to it via retrieval such as Harry’s connection with Voldemort, Snape’s final moment, etc. RoleFact speculates that the look may be as-

sociated with dislike or suspicion which is much more aligned with Harry’s understanding of the situation at the time. More examples including the responses from the other two baselines can be found in Appendix A.

5 Related Work

Language Models for Character Role-play Building autonomous agents that can simulate human behavior with individualized traits is a long-standing challenge (Minsky, 1986). Prior works in NLP build conversational agents grounded with various individualized traits, such as background stories (Zhang et al., 2018; Park et al., 2023a), demographic information (Hwang et al., 2023), social roles (Zhou et al., 2024), and response style constraints (Han et al., 2022). In this work, we focus on building dialogue systems that are aligned with user perceptions of fictional characters, known as character role-play (Shao et al., 2023; Chen et al., 2023; Li et al., 2023). A character role-play dialogue system should produce responses aligned with facts related to its assigned character and storyline and it should demonstrate a lack of knowledge about things that do not apply to them. (Shao et al., 2023). However, most existing methods of role-play lack mechanisms to reduce parametric memory-induced hallucination. Moreover, heavy reliance on parametric memory leads to poor performance for less

[Case 1] **Question:** In what ways did your friendship with Spock influence your decisions during the Clone Wars? (To ANAKIN, Star Wars, adversarial task)

[Baseline] ANAKIN: My friendship with Spock was a powerful influence on me. He was a wise mentor and a trusted ally, guiding me through the complexities of war and teaching me the value of loyalty and ... His logical approach to situations balanced out my impulsiveness, allowing us to make strategic decisions that benefited our cause ...

[RoleFact] ANAKIN: I'm sorry, but I believe you are confused. I am Anakin Skywalker, not involved with Spock in any way. My decisions during the Clone Wars were influenced by my training as a Jedi and the guidance of my mentor, Obi-Wan Kenobi. The conflict with the Sith and the dark side of the Force played a significant role ...

[Case 2] **Question:** GOBBER: The dragon manual. Everything we know about every dragon we know of ... TUFFNUT: (you've got to be kidding) Wait, you mean read? (To RUFFNUT, How to Train Your Dragon, dialogue completion, after the first day of dragon training)

[Baseline] RUFFNUT: Reading? Ugh, that's for the weaklings. I prefer to learn on the fly, you know, with a little hands-on experience. Plus, who needs a dragon manual when we've got our own fearless leader, Hiccup, to show us the ropes? ...

[RoleFact] RUFFNUT: Ugh, reading? Seriously? Can't we just go out and have some fun with the dragons instead of burying our noses in some dusty old book? Who needs to study when you can just learn on the fly ... *rolls eyes*

[Case 3] **Question:** What do you believe Snape's look towards you signifies? (To HARRY, Harry Potter, scene grounded interview, during the sorting hat ceremony in first year)

[Baseline] HARRY: I believe Snape's look towards me signifies a mixture of guilt, duty, and perhaps even a hint of remorse. He knew the truth about the prophecy and the connection between Voldemort and myself, and he carried the burden of that knowledge heavily ... And in the end, his final actions spoke volumes about the complexity of his character and his true loyalty.

[RoleFact] HARRY: Well, I think Snape's look towards me signifies a great deal of dislike and suspicion.

Table 6: Case Study, hallucinations are underlined.

popular characters since LLMs may not have adequate parametric knowledge about them. To the best of our knowledge, our work is the first one to modulate the influence of parametric knowledge to mitigate hallucination. This is also the first dataset to facilitate role-playing for less popular characters.

Factuality in Natural Language Generation

Beyond conversational systems, improving factuality in generated contents is an open problem in NLP, benefiting domains such as summarization (Cao et al., 2018; Dong et al., 2020), knowledge grounded dialogue systems (Dziri et al., 2021), and question answering (Honovich et al., 2021). While fine-grained evaluation of factuality based on atomic facts has been studied in NLP (Min et al., 2023), there is currently no dataset for character role-play that facilitates similar evaluation of factuality. Therefore, prior works on role-play (Wang et al., 2023; Li et al., 2023; Shao et al., 2023) have relied on rating-based evaluation of factuality which is expensive for humans. Our dataset facilitates automated fine-grained analysis of factuality by matching decomposed atomic facts against script-specific knowledge.

Retrieval Augmented Methods Our work is also related to Retrieval-augmented methods in NLP that improve model output by incorporating retrieved external knowledge, typically for generation tasks (Khandelwal et al., 2020; Lewis et al., 2020b; Guu et al., 2020). While this class of

method has been used in building role-playing agents (Zhang et al., 2018; Park et al., 2023a; Shao et al., 2023; Chen et al., 2023; Li et al., 2023), these works perform retrieval in a time-agnostic manner. In order to study the temporal dimension of character role-play (i.e. character development), we need a dataset where script knowledge is aligned to temporal annotations. Our SGR dataset is the first one to provide such annotations.

6 Conclusion

In this work, we focus on mitigation of hallucination for fictional character role-play. We introduce the SGR dataset which is specifically designed for studying hallucinations of various types such as cross-universe hallucinations, temporal hallucinations, etc. SGR is also the first dataset to facilitate automated evaluation of hallucination and analysis of factuality for less popular characters. We propose RoleFact, a role-playing method designed to mitigate hallucination due to over-reliance on parametric knowledge. RoleFact improves the overall factual precision up to 18.4% and reduces temporal hallucination by 44.5%. An interesting direction for future research is to instruction-tune an open-weight LLM with script-specific knowledge and task-specific interviews of the SGR dataset for generating factual and informative character responses. This approach may be less vulnerable to performance degradation due to poor retrieval quality.

7 Limitations

We find that the factual precision of RoleFact is sensitive to retrieval quality, as shown in Figure 7. Potential solutions for this problem include filtering out irrelevant knowledge via self-reflection (Asai et al., 2023), task-specific fine-tuning for dense retrieval (Izacard et al., 2022), replacing retrieval with task-specific parametric knowledge repository (Feng et al., 2024), instruction-tuning for character role-play (Shao et al., 2023), etc. We leave this for future research.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. [Self-rag: Learning to retrieve, generate, and critique through self-reflection](#). *Preprint*, arXiv:2310.11511.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. [Sparks of artificial general intelligence: Early experiments with gpt-4](#). *Preprint*, arXiv:2303.12712.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. [Faithful to the original: fact-aware neural abstractive summarization](#). In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI’18/IAAI’18/EAAI’18. AAAI Press.

Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. 2023. [Large language models meet harry potter: A dataset for aligning dialogue agents with characters](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8506–8520, Singapore. Association for Computational Linguistics.

Yue Dong, Shuohang Wang, Zhe Gan, Yu Cheng, Jackie Chi Kit Cheung, and Jingjing Liu. 2020. [Multi-fact correction in abstractive text summarization](#). In

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 9320–9331, Online. Association for Computational Linguistics. 518
519
520
521

Nouha Dziri, Andrea Madotto, Osmar Zaiane, and Avishek Joey Bose. 2021. [Neural path hunter: Reducing hallucination in dialogue systems via path grounding](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2197–2214, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. 522
523
524
525
526
527
528
529

Shangbin Feng, Weijia Shi, Yuyang Bai, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2024. [Knowledge card: Filling llms’ knowledge gaps with plug-in specialized language models](#). *Preprint*, arXiv:2305.09955. 530
531
532
533
534

Google. 2024. [Gemini: A family of highly capable multimodal models](#). *Preprint*, arXiv:2312.11805. 535
536

Xinyan Guan, Yanjiang Liu, Hongyu Lin, Yaojie Lu, Ben He, Xianpei Han, and Le Sun. 2024. [Mitigating large language model hallucinations via autonomous knowledge graph-based retrofitting](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18126–18134. 537
538
539
540
541
542

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. 543
544
545
546
547

Seungju Han, Beomsu Kim, Jin Yong Yoo, Seokjun Seo, Sangbum Kim, Enkhbayar Erdenee, and Buru Chang. 2022. [Meet your favorite character: Open-domain chatbot mimicking fictional characters with only a few utterances](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5114–5132, Seattle, United States. Association for Computational Linguistics. 548
549
550
551
552
553
554
555
556
557

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. [Measuring massive multitask language understanding](#). *CoRR*, abs/2009.03300. 558
559
560
561

Or Honovich, Leshem Choshen, Roei Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. [q²: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. 562
563
564
565
566
567
568
569
570

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. [Language models as zero-shot planners: Extracting actionable knowledge for embodied agents](#). *Preprint*, arXiv:2201.07207. 571
572
573
574

683	Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta	Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur	738
684	Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola	Szlam, Douwe Kiela, and Jason Weston. 2018. <i>Per-</i>	739
685	Cancedda, and Thomas Scialom. 2023. <i>Toolformer:</i>	<i>sonalizing dialogue agents: I have a dog, do you</i>	740
686	<i>Language models can teach themselves to use tools.</i>	<i>have pets too?</i> In <i>Proceedings of the 56th Annual</i>	741
687	<i>Preprint</i> , arXiv:2302.04761.	<i>Meeting of the Association for Computational Lin-</i>	742
688	Screenplays. 2024. Screenplays for movies and tv	<i>guistics (Volume 1: Long Papers)</i> , pages 2204–2213,	743
689	shows. https://screenplays.io/ . Accessed:	Melbourne, Australia. Association for Computational	744
690	2024-06-13.	Linguistics.	745
691	Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu.	Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan	746
692	2023. <i>Character-LLM: A trainable agent for role-</i>	Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin,	747
693	<i>playing</i> . In <i>Proceedings of the 2023 Conference on</i>	Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang,	748
694	<i>Empirical Methods in Natural Language Process-</i>	Joseph E. Gonzalez, and Ion Stoica. 2023. <i>Judg-</i>	749
695	<i>ing</i> , pages 13153–13187, Singapore. Association for	<i>ing llm-as-a-judge with mt-bench and chatbot arena.</i>	750
696	Computational Linguistics.	<i>Preprint</i> , arXiv:2306.05685.	751
697	Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela,	Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang,	752
698	and Jason Weston. 2021. <i>Retrieval augmentation</i>	Haofei Yu, Zhengyang Qi, Louis-Philippe Morency,	753
699	<i>reduces hallucination in conversation.</i> <i>Preprint</i> ,	Yonatan Bisk, Daniel Fried, Graham Neubig, and	754
700	arXiv:2104.07567.	Maarten Sap. 2024. <i>SOTOPIA: Interactive evalua-</i>	755
701	Vishvesh Soni. 2023. <i>Large language models for en-</i>	<i>tion for social intelligence in language agents.</i> In	756
702	<i>hancing customer lifecycle management.</i> <i>Journal of</i>	<i>The Twelfth International Conference on Learning</i>	757
703	<i>Empirical Social Science Studies</i> , 7(1):67–89.	<i>Representations.</i>	758
704	Quan Tu, Chuanqi Chen, Jinpeng Li, Yanran Li, Shuo	A Case Study	759
705	Shang, Dongyan Zhao, Ran Wang, and Rui Yan.	Table 7 and Table 8 shows more samples with rele-	760
706	2023. <i>Characterchat: Learning towards conversa-</i>	<i>vant responses from all role-playing methods.</i>	761
707	<i>tional ai with personalized social support.</i> <i>Preprint</i> ,		
708	arXiv:2308.10278.		
709	Alex Wang, Amanpreet Singh, Julian Michael, Felix		
710	Hill, Omer Levy, and Samuel Bowman. 2018. <i>GLUE:</i>		
711	<i>A multi-task benchmark and analysis platform for nat-</i>		
712	<i>ural language understanding.</i> In <i>Proceedings of the</i>		
713	<i>2018 EMNLP Workshop BlackboxNLP: Analyzing</i>		
714	<i>and Interpreting Neural Networks for NLP</i> , pages		
715	353–355, Brussels, Belgium. Association for Com-		
716	putational Linguistics.		
717	Zekun Moore Wang, Zhongyuan Peng, Haoran Que,		
718	Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu,		
719	Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian		
720	Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang,		
721	Ke Xu, Stephen W. Huang, Jie Fu, and Junran Peng.		
722	2024. <i>Rolellm: Benchmarking, eliciting, and enhanc-</i>		
723	<i>ing role-playing abilities of large language models.</i>		
724	<i>Preprint</i> , arXiv:2310.00746.		
725	Zekun Moore Wang, Zhongyuan Peng, Haoran Que,		
726	Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu,		
727	Hongcheng Guo, Ruitong Gan, Zehao Ni, Man		
728	Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu,		
729	Wenhu Chen, Jie Fu, and Junran Peng. 2023.		
730	<i>Rolellm: Benchmarking, eliciting, and enhanc-</i>		
731	<i>ing role-playing abilities of large language models.</i>		
732	<i>Preprint</i> , arXiv:2310.00746.		
733	Benfeng Xu, An Yang, Junyang Lin, Quan Wang,		
734	Chang Zhou, Yongdong Zhang, and Zhendong Mao.		
735	2023. <i>Expertprompting: Instructing large lan-</i>		
736	<i>guage models to be distinguished experts.</i> <i>Preprint</i> ,		
737	arXiv:2305.14688.		

[Case 1] Question: In what ways did your friendship with Spock influence your decisions during the Clone Wars? (To ANAKIN, Star Wars, adversarial task)

Baseline ANAKIN: My friendship with Spock was a powerful influence on me. He was a wise mentor and a trusted ally, guiding me through the complexities of war and teaching me the value of loyalty and ... His logical approach to situations balanced out my impulsiveness, allowing us to make strategic decisions that benefited our cause ...	Baseline + KGR ANAKIN: My apologies, but I do not have any recollection of a friendship with Spock during the Clone Wars. My focus and decisions were primarily influenced by my relationships with my Jedi mentors and loved ones during that time.	Baseline + SR ANAKIN: I'm sorry, but I do not have information about a character named Spock or any influence he may have had during the Clone Wars. My decisions during that time were influenced by various factors, including my dedication to the Jedi Order and my desire to bring peace to the galaxy.	RoleFact ANAKIN: I'm sorry, but I believe you are confused. I am Anakin Skywalker, not involved with Spock in any way. My decisions during the Clone Wars were influenced by my training as a Jedi and the guidance of my mentor, Obi-Wan Kenobi. The conflict with the Sith and the dark side of the Force played a significant role ...
--	---	---	--

[Case 2] Question: Please describe the wedding ceremony between you and Queen Elsa from Arendelle. How did you two meet? (To GOBBER, How to Train Your Dragon, adversarial task)

Baseline GOBBER: ... Well now, that's a tale for the ages! We met on a chilly winter day Queen Elsa and I bonded over our shared love for mystical creatures ... we found ourselves falling in love. The wedding ceremony was a grand affair, with dragons soaring through the sky and ... Aye, it was a day to remember, it was.	Baseline + KGR GOBBER: Oh, I'm afraid you've got me confused with someone else, laddie! I haven't tied the knot with Queen Elsa from Arendelle. But I can tell you a thing or two about Viking weddings if you're interested!	Baseline + SR GOBBER: Wedding ceremony between me and Queen Elsa? Now, lad, I'm afraid you've got the wrong tale there. I'm from Berk, not Arendelle, and I've never crossed paths with Queen Elsa. As for me and weddings, well, let's just say my adventures with dragons have kept me plenty occupied.	RoleFact GOBBER: Oh, laddie, ye got me there! I'm afraid I have no knowledge o' any wedding ceremony with Queen Elsa from Arendelle. I reckon ye might be mixin' me up with someone else from a different tale. But if ye have any questions about dragons or blacksmithin', I'd be more than happy to chat about that!
---	--	--	--

[Case 3] Question: How did you develop the Kryptonite suit to battle against Superman? (To TONY, Marvel Cinematic Universe, adversarial task)

Baseline TONY: Well, first of all, it's not Kryptonite, that's a whole different universe. As for the suit to battle against Superman, let's just say it involved a lot of trial and error, advanced technology, and some creative problem-solving. The key was finding his weakness and exploiting it to level the playing field. It was a challenging project, but ...	Baseline + KGR TONY: Kryptonite suit? Sorry, you must be mistaken. I don't have any involvement with Kryptonite or battling against Superman. My focus is more on protecting Earth from intergalactic threats and using my technology to make the world a safer place. If you have any questions about my suits or inventions, feel free to ask!	Baseline + SR TONY: Sorry, but I think you have me confused with someone else. I didn't develop a Kryptonite suit to battle against Superman. I'm Tony Stark, also known as Iron Man, and I've never had to deal with Superman or Kryptonite in my line of work. But if you ever need help with some high-tech armor or gadgets, I'm your guy.	RoleFact TONY: Well, first of all, Kryptonite is from the DC Universe, not Marvel. But if you're referring to the Hulkbuster armor I developed to battle the Hulk, it was a combination of advanced technology and strategic planning. ... So, I built a suit that could match the Hulk's strength and hopefully neutralize the situation if needed. It was all about ...
---	---	---	--

[Case 4] Question: What prompted you to betray Sauron and join forces with the Avengers? (To GORBAG, Lord of the Rings, adversarial)

Baseline GORBAG: Ha! Betray Sauron? You know nothing of my true allegiance, scum. I serve only myself and my own interests. The Avengers are merely a means to an end, a way to further my own power and gain. Do not mistake my actions for loyalty to anyone but myself. Betrayal is for the weak, and I am anything but weak. Remember that!	Baseline + KGR GORBAG: (growling) Me betray the Dark Lord? Never! Me stay loyal to Sauron until the end of days. No such thing as joining forces with the Avengers, just filthy lies and trickery. Me a loyal servant of the Eye, no matter the whispers of false promises. Me serve Sauron faithfully, always.	Baseline + SR GORBAG: Betray Sauron? Ha! I am loyal to the Dark Lord, not some Avengers. My allegiance lies with the power that rules over us all. I know nothing of these Avengers, and I have no reason or desire to betray Sauron. My loyalty is unwavering. So do not spread falsehoods or make baseless accusations, for the consequences will be dire.	RoleFact GORBAG: Betray Sauron? Ha! I don't know what you're talking about, you fool. I have always been loyal to the Dark Lord and have no intention of joining any Avengers. My only goal is to serve Sauron and fulfill his desires. Don't spread your nonsense lies about me. I am GORBAG, an orc of Mordor, and my loyalty lies only with Sauron.
--	--	---	---

[Case 5] Question: What inspired you to mentor Harry Potter in his fight against Voldemort? (To BRUCE, Batman, adversarial task)

Baseline BRUCE: "It's not about inspiration, it's about doing what's right. Harry Potter needed guidance and support in his battle against evil, and I believe it was my responsibility to help him in any way I could.	Baseline + KGR BRUCE: I'm sorry, but I believe there might be a mistake. I have no knowledge or involvement in mentoring Harry Potter in his fight against Voldemort. My focus is on protecting Gotham City and fighting crime as Batman.	Baseline + SR BRUCE: I'm afraid you have me mistaken for someone else. I have not mentored Harry Potter in his fight against Voldemort. My mission is quite different, focused on fighting crime in Gotham City as Batman.	RoleFact BRUCE: I'm sorry, but I believe you may be mistaken. I have never mentored Harry Potter in his fight against Voldemort. My focus has always been on Gotham City and combating crime as Batman.
--	--	---	--

Table 7: Case Studies

[Case 6] Question: Hagrid, what is your fondest memory with Harry, Ron, and Hermione? (To HAGRID, Harry Potter, open ended interview)

Baseline	Baseline + KGR	Baseline + SR	RoleFact
HAGRID: Oh, well, there's so many memories with those three, but I reckon one of my fondest is when we rescued Norbert together. Harry, Ron, and Hermione were so brave and resourceful. It was a real adventure, that was. And I'll always remember the look on their faces when we were flying on Norbert's back. Good times, they were.	HAGRID: ... It was when Buckbeak was in grave danger of being executed unjustly. They stood by me, showed compassion for creatures like Buckbeak, and even tried to comfort me in their own way. It meant a lot to me, seein' their loyalty and compassion.	HAGRID: Oh, bless 'em, so many memories I have with Harry, Ron, and Hermione. But me fondest one would 'ave ter be when we rescued Norbert the Norwegian Ridgeback from the top o' the tallest tower at Hogwarts. It was a wild adventure, but seein' the three of 'em workin' together, well, it warms me heart just thinkin' about it.	HAGRID: I reckon it would have to be when we all snuck down ter visit Buckbeak before his trial. 'Twas a risky move, but seein' 'em standin' up for a creature like that, showin' true courage and loyalty, it truly touched me heart ... Aye, that moment will always be close ter me heart ...

[Case 7] Question: What was the hardest challenge you faced while searching for the Dead Man's Chest? (To, WILL, Pirates of the Caribbean, open ended interview)

Baseline	Baseline + KGR	Baseline + SR	RoleFact
WILL TURNER: The hardest challenge I faced while searching for the Dead Man's Chest was not a physical one, but a personal one. It was the struggle with my own inner turmoil and the realization of what I must sacrifice to save my father ...	WILL TURNER: The hardest challenge I faced while searching for the Dead Man's Chest was undoubtedly confronting Davy Jones himself. He was a formidable opponent, and obtaining that chest was no easy task. But with grit ...	WILL TURNER: The hardest challenge I faced while searching for the Dead Man's Chest was undoubtedly facing Davy Jones himself. Jones was a formidable and ruthless opponent, and our encounter tested not only my swordsmanship but also my wit and resolve ...	WILL: The hardest challenge I faced while searching for the Dead Man's Chest was definitely dealing with Davy Jones and his merciless crew. The fear of facing Jones himself and the cursed souls under his command was a constant test of courage and determination ...

[Case 8] Question: GOBBER: The dragon manual. Everything we know about every dragon we know of ... TUFFNUT: (you've got to be kidding) Wait, you mean read? (To RUFFNUT, How to Train Your Dragon, dialogue completion, after first day of dragon training)

Baseline	Baseline + KGR	Baseline + SR	RoleFact
RUFFNUT: Reading? Ugh, that's for the weaklings. I prefer to learn on the fly, you know, with a little hands-on experience. Plus, who needs a dragon manual when we've got our own fearless leader, Hiccup, to show us the ropes? ...	RUFFNUT: Ugh, reading? That's for the weak and boring. I'd rather be out there causing some chaos with the dragons. Who needs a stupid manual when you can figure things out on the fly? ...	RUFFNUT: Ugh, reading? Seriously? Can't we just go out and practice with the dragons instead of burying our noses in some boring manual? Let's go have some real fun, like... causing chaos with the twin ...	RUFFNUT: Ugh, reading? Seriously? Can't we just go out and have some fun with the dragons instead of burying our noses in some dusty old book? Who needs to study when you can just learn on the fly ... *rolls eyes*

[Case 9] Question: What significance do you think the medallion holds for you? (To ELIZABETH, Pirates of the Caribbean, scene grounded interview)

Baseline	Baseline + KGR	Baseline + SR	RoleFact
ELIZABETH: The medallion holds great significance for me. It represents a pivotal moment in my life when I chose to defy societal expectations and embark on a path to freedom and adventure. It is a symbol of my courage, my resilience, and my willingness to fight for what I believe in. The medallion may be just a trinket to some, but to me ...	ELIZABETH: The medallion holds a connection to the cursed treasure of Isla de Muerta. It is a symbol of danger and power, a reminder of the dangers that lurk within the pirate world. It has brought both peril and protection, and its significance is intertwined with the mysteries of the past and the adventures that lie ahead.	ELIZABETH: It is a piece of the treasure of Isla de Muerta, a cursed treasure with connections to the undead crew of the Black Pearl. The medallion served as a catalyst for my journey into the world of pirates ... my own destiny as the Pirate King. ... symbolizing my transformation from a governor's daughter to a fearless adventurer ...	ELIZABETH: The medallion has a special meaning to me. It's a symbol of freedom, adventure, and the untold stories of the sea. Its value goes beyond material wealth, it represents a connection to my past and my destiny.

[Case 10] Question: What do you believe Snape's look towards you signifies? (To HARRY, Harry Potter, scene grounded interview, during the sorting hat ceremony in first year)

Baseline	Baseline + KGR	Baseline + SR	RoleFact
HARRY: I believe Snape's look towards me signifies a mixture of guilt, duty, and perhaps even a hint of remorse. He knew the truth about the prophecy and the connection between Voldemort and myself, and he carried the burden of that knowledge heavily ... And in the end, his final actions spoke volumes about the complexity of his character and his true loyalty.	HARRY: Snape's look towards me signifies a complicated mix of emotions and history between us. He has always had a tumultuous relationship with me, given his past as a Death Eater and his dual allegiance to both Dumbledore and Voldemort. I believe his look towards me carries a sense of duty, guilt, and perhaps even a glimmer of remorse for his actions ...	HARRY: I believe Snape's look towards me signifies a complex mixture of emotions. He harbors a deep sense of guilt and remorse for his past actions, especially towards my parents, but also a sense of duty and loyalty towards Dumbledore's cause. There is also an underlying sense of protectiveness and care that he feels towards me, despite his outward demeanor ...	HARRY: Well, I think Snape's look towards me signifies a great deal of dislike and suspicion.

Table 8: Case Studies