← Go to **ACL ARR 2024 June** homepage (/group?id=aclweb.org/ACL/ARR/2024/June)

Mitigating Hallucination in Fictional Character Role-Play

PDF (/pdf?id=CrF1t2Qqyi)

Nafis Sadeq (/profile?id=~Nafis_Sadeq1), Zhouhang Xie (/profile?id=~Zhouhang_Xie1), Byungkyu Kang (/profile?email=jay_kang%40intuit.com), Prarit Lamba (/profile?email=prarit_lamba%40intuit.com), Xiang Gao (/profile?id=~Xiang_Gao4), Julian McAuley (/profile?id=~Julian_McAuley1)

16 Jun 2024 (modified: 09 Aug 2024) ACL ARR 2024 June Submission Everyone, Ethics Reviewers Revisions (/revisions?id=CrF1t2Qqyi) BibTeX

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.

Paper Type: Long
Research Area: Dialogue and Interactive Systems
Research Area Keywords: factuality,retrieval-augmented generation
Contribution Types: NLP engineering experiment, Publicly available software and/or pre-trained models, Data resources
Languages Studied: English
Reviewing Volunteers: Nafis Sadeq (/profile?id=~Nafis_Sadeq1)
Reviewing Volunteers For Emergency Reviewing: The volunteers listed above are only willing to serve as regular reviewers.
Reviewing No Volunteers Reason: N/A - An author was provided in the previous question.
Reassignment Request Action Editor: This is not a resubmission
Reassignment Request Reviewers: This is not a resubmission
Software: zip (/attachment?id=CrF1t2Qqyi&name=software)
Data: zip (/attachment?id=CrF1t2Qqyi&name=data)
Preprint: yes
Preprint Status: We plan to release a non-anonymous preprint in the next two months (i.e., during the reviewing process).
Consent To Share Data: yes
Consent To Share Submission Details: On behalf of all authors, we agree to the terms above to share our submission details.
Author Submission Checklist: I confirm that the paper is anonymous and that all links to data/code repositories in the paper are anonymous., I confirm that the paper has proper length (Short papers: 4 content pages maximum, Long papers: 8 content pages maximum, Ethical considerations and Limitations do not count toward this limit), I confirm that the paper is properly formatted (Templates for *ACL conferences can be found here: https://github.com/acl-org/acl-style-files (https://github.com/acl-org/acl-style-files).)
A1 Limitations Section: This paper has a limitations section.

A2 Potential Risks: No
A3 Abstract And Introduction Summarize Claims: Yes
B Use Or Create Scientific Artifacts: Yes
B1 Cite Creators Of Artifacts: Yes
B2 Discuss The License For Artifacts: N/A
B3 Artifact Use Consistent With Intended Use: N/A
B4 Data Contains Personally Identifying Info Or Offensive Content: N/A
B5 Documentation Of Artifacts: Yes
B6 Statistics For Data: Yes
C Computational Experiments: No
C1 Model Size And Budget: N/A
C2 Experimental Setup And Hyperparameters: Yes
C3 Descriptive Statistics: Yes
C4 Parameters For Packages: Yes
D Human Subjects Including Annotators: No
D1 Instructions Given To Participants: N/A
D2 Recruitment And Payment: N/A
D3 Data Consent: N/A
D4 Ethics Review Board Approval: N/A
D5 Characteristics Of Annotators: N/A
E Ai Assistants In Research Or Writing: No
E1 Information About Use Of Ai Assistants: N/A
Association For Computational Linguistics - Blind Submission License Agreement: On behalf of all authors, I agree
Submission Number: 4477

Discussion (/forum?id=CrF1t2Qqyi#discussion)

Filter by reply type | Filter by author | Search keywords...

Sort: Newest First | - | = | ≡

Everyone | Submission4477... | Submission4477 Area... | Submission4477 Authors | Submission4477... | Program Chairs | Submission4477... | Submission4477... | Ethics Chairs | Submission4477... | Submission4477... | Submission4477... | ✖

11 / 11 replies shown

Add: **Author-Editor Confidential Comment** **Withdrawal**

- = ≡

Meta Review of Submission4477 by Area Chair iq2J

Meta Review ✎ Area Chair iq2J 09 Aug 2024, 06:33 (modified: 22 Aug 2024, 15:39)

Senior Area Chairs, Area Chairs, Authors, Reviewers Submitted, Program Chairs, Commitment Readers

Revisions (/revisions?id=X1LY0UNJvj)

Metareview:
This paper addresses the issue of hallucination in fictional character role-play using large language models (LLMs). Hallucination occurs when a character makes statements that are outside their knowledge base or time frame, leading to inconsistencies and inaccuracies. The paper proposes RoleFact, a method to mitigate hallucination by balancing the influence of LLM’s parametric knowledge with non-parametric retrieved knowledge. The paper also introduces the SGR dataset, containing over 2,000 characters and 72,000 interviews. This dataset is specifically designed for studying character

hallucinations and includes script-specific knowledge, facilitating a more nuanced evaluation. Experiments show RoleFact improves factual precision by 18% for adversarial questions, reduces temporal hallucination by 44% for time-sensitive interviews, and improves factual precision by 23% for less popular characters.

Summary Of Reasons To Publish:

- The task is of importance in terms of mitigating hallucination in fictional character role-play.
- The proposed dataset SGR is significantly valuable and able to contribute to the community.
- The proposed RoleFact method experimentally demonstrated its efficacy.

Summary Of Suggested Revisions:

- The quality analysis of the proposed dataset is needed.
- Ensuring the reliability of the FactScore metric in the context of role-playing evaluations in the proposed SGR dataset is needed.
- There's lack of details such as prompt anonymization, inference cost, and temporal hallucination should be included.
- Proposing a method beyond a simple prompt-based method would enhance the completeness and novelty of this study. In addition, the selection method of confidence threshold could be improved.
- This paper is absent of connection with core related papers [1, 2].

[1] Lu et al., "Large Language Models are Superpositions of All Characters: Attaining Arbitrary Role-play via Self-Alignment", ACL 2024

[2] Ahn et al., "TimeChara: Evaluating Point-in-Time Character Hallucination of Role-Playing Large Language Models", ACL 2024 Findings

Overall Assessment: 3 = There are major points that may be revised

Suggested Venues: ARR; NAACL; COLING

Best Paper Ae: No

Ethical Concerns:

There are no concerns with this submission

Needs Ethics Review: No

Author Identity Guess: 1 = I do not have even an educated guess about author identity.

Add: **Author-Editor Confidential Comment**

Official Review of Submission4477 by Reviewer 3yj6

Official Review Reviewer 3yj6 19 Jul 2024, 17:33 (modified: 22 Aug 2024, 15:39)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer 3yj6, Commitment Readers

Revisions (/revisions?id=MRasCsZcOt)

Paper Summary:

This paper focuses on mitigating hallucination in fictional character role-play using large language models (LLMs). The authors introduce a dataset called SGR with over 2,000 characters and 72,000 interviews, including 18,000 adversarial questions, to enable systematic study of character hallucinations. They propose RoleFact, a role-playing method that modulates the influence of parametric knowledge using a pre-calibrated confidence threshold to improve factual precision of generated responses. Experiments show RoleFact improves factual precision by 18% for adversarial questions, reduces temporal hallucination by 44% for time-sensitive interviews, and improves factual precision by 23% for less popular characters.

Summary Of Strengths:

- The paper addresses an important problem of mitigating hallucination in fictional character role-play, which has applications in customer support, embodied agents, computational social science, etc.
- The authors introduce a novel dataset SGR specifically designed to study various types of hallucinations like cross-universe, temporal, and for less popular characters. SGR enables automated evaluation of hallucination.

- The proposed RoleFact method shows significant improvements in factual precision, reduction in temporal hallucination, and better performance for less popular characters compared to baselines.
- Detailed experiments, ablation studies and human evaluations are conducted to comprehensively analyze the performance of RoleFact.

Summary Of Weaknesses:

- The method is almost entirely prompt-based, with the prompt template remaining fixed throughout the experiment. One metric (SFPR) used in Figure 2 might be sensitive to the prompt. For example, we can encourage the LLM to provide longer outputs, which will potentially lead to larger SFPR values.
- The method for setting the confidence threshold is less appealing.

Comments Suggestions And Typos:

1. Does the selected confidence threshold (0.6) generalize well for all models? Is 0.6 used for all models in Table 2?
2. What retrieval method is used for Table 2? I assume it is BM25 because Figure 7 shows BM25 is the best. What is the reason that BM25 performs better than dense retrieval? Ideally, sparse retrieval is keyword matching, while dense retrieval might be able to understand semantics. I think the better performance of BM25 for fairness scores might reveal some special patterns in the data.
3. How are the decoding parameters (e.g., greedy or not, top_k, top_p, temperature) used for different models?

Confidence: 3 = Pretty sure, but there's a chance I missed something. Although I have a good feel for this area in general, I did not carefully check the paper's details, e.g., the math or experimental design.

Soundness: 4 = Strong: This study provides sufficient support for all of its claims/arguments. Some extra experiments could be nice, but not essential.

Overall Assessment: 3 = Good: This paper makes a reasonable contribution, and might be of interest for some (broad or narrow) sub-communities, possibly with minor revisions.

Best Paper: No

Needs Ethics Review: No

Reproducibility: 5 = They could easily reproduce the results.

Datasets: 5 = Enabling: The newly released datasets should affect other people's choice of research or development projects to undertake.

Software: 5 = Enabling: The newly released software should affect other people's choice of research or development projects to undertake.

Knowledge Of Or Educated Guess At Author Identity: No

Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources

Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Add: **Author-Editor Confidential Comment**

- = ≡

Clarification regarding the confidence threshold, retrieval performance and decoding parameters

Official Comment

Authors (jay_kang@intuit.com (/profile?id=jay_kang@intuit.com), prarit_lamba@intuit.com (/profile?id=prarit_lamba@intuit.com), Xiang Gao (/profile?id=~Xiang_Gao4), Julian McAuley (/profile?id=~Julian_McAuley1), +2 more (/group/info?id=aclweb.org/ACL/ARR/2024/June/Submission4477/Authors))

26 Jul 2024, 16:16 (modified: 22 Aug 2024, 15:39)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer 3yj6, Commitment Readers

Revisions (/revisions?id=KVUOErC6P2)

Comment:

Thank you for your valuable feedback! We are providing the requested information below and are happy to add the provided details to the paper for better clarity.

Question 1: Does the selected confidence threshold (0.6) generalize well for all models? Is 0.6 used for all models in Table 2?

For the whole dataset, 0.6 is the best confidence threshold for all models. RoleFact performance reported for all LLMs in Table 2 is generated using threshold 0.6. For a specific subset of the dataset such as the open-ended interviews, we have found some variance in the optimal confidence threshold across models (between 0.6 and 0.8). As we mentioned in line 229, open-ended interviews constitute 25% of the SGR dataset. Therefore, the subset-level variance in the optimal confidence threshold has a limited impact on the dataset-level optimal confidence threshold. If RoleFact is used with another dataset with a much larger share of open-ended interviews (e.g. RoleLLM, Wang et al. 2024), we expect the optimal confidence to vary across LLMs.

Question 2: What retrieval method is used for Table 2? What is the reason that BM25 performs better than dense retrieval?

All results in Table 2 use BM25. One special pattern of our dataset is that out of 2.4 million knowledge events, 1.1 million are utterances explicitly annotated with character names (mentioned in section 3). Off-the-shelf dense retrieval such as S-BERT seems sensitive to named entities. If the query and target document share a named entity, S-BERT generates a high similarity score even if the rest of the content does not have a semantic match. Consider the two utterances from Tony Stark:

Query: "TONY: Not a great plan. When they come, and they will, they'll come for you."

Target: "TONY: This thing on? Hey, Ms. Potts. Pep. If you find this recording, don't post it on social media. It's going to be a real tearjerker."

The S-BERT similarity score between the query and target is 0.4 (it would have been 1 for an exact match). The BM25 similarity score between the query and target is 5.62 (would have been 102 for an exact match). Since BM25 is less sensitive to named entities, it has a better chance of finding fine-grained similarity in the overall content.

Question 3: How are the decoding parameters used for different models?

Decoding for character response generation is performed with temperature = 0.7 and top_p = 0.95 for all models.

Add: **Author-Editor Confidential Comment**

Reply

Official Comment Reviewer 3yj6 30 Jul 2024, 09:20 (modified: 22 Aug 2024, 15:39)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer 3yj6, Commitment Readers

Revisions (/revisions?id=j9dK3r1okE)

Comment:

Thanks for the reply. Those clarification are helpful. So I decide to increase the soundness from 3.5 to 4. However, I keep my overall assessment to 3 as I am not a big fan to the method you proposed.

Add: **Author-Editor Confidential Comment**

Official Review of Submission4477 by Reviewer 1AXu

Official Review Reviewer 1AXu 18 Jul 2024, 01:13 (modified: 22 Aug 2024, 15:39)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer 1AXu, Commitment Readers

Revisions (/revisions?id=uClAGG3kdG)

Paper Summary:

This paper addresses the issue of hallucination in fictional character role-play using large language models (LLMs). Hallucination occurs when a character makes statements that are outside their knowledge base or time frame, leading to inconsistencies and inaccuracies. The paper proposes RoleFact, a method to mitigate hallucination by balancing the influence of LLM's parametric knowledge with non-parametric retrieved knowledge. The paper also introduces the SGR dataset, containing over 2,000 characters and 72,000 interviews. This dataset is specifically designed for studying character hallucinations and includes script-specific knowledge, facilitating a more nuanced evaluation.

Summary Of Strengths:
S1: The paper clearly identifies and addresses the specific challenge of hallucination in fictional character role-play, providing a targeted solution for this issue.

S2: The introduction of the SGR dataset is a significant contribution. It fills a gap in the field by providing a large, diverse, and script-grounded dataset specifically designed for studying character hallucinations. This dataset enables systematic evaluation and analysis, facilitating further research in this area.

S3: The paper employs a comprehensive evaluation strategy, measuring factual precision, informativeness, and temporal hallucination rate. This multi-faceted evaluation provides a thorough understanding of the method's performance and effectiveness. Additionally, human evaluations are conducted to assess the quality of speaker style imitation.

Summary Of Weaknesses:
W1: RoleFact's performance is heavily dependent on the quality of the retrieved knowledge. Poor retrieval can lead to inaccurate or incomplete responses, potentially compromising the method's effectiveness.

W2: The paper acknowledges the need to tune the confidence threshold for optimal performance. However, finding the ideal threshold can be challenging and may require significant effort. This process could be further automated or streamlined to make the method more accessible.

Comments Suggestions And Typos:
N/A

Confidence: 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.
Soundness: 4 = Strong: This study provides sufficient support for all of its claims/arguments. Some extra experiments could be nice, but not essential.
Overall Assessment: 3.5
Best Paper: No
Needs Ethics Review: No
Reproducibility: 4 = They could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.
Datasets: 4 = Useful: I would recommend the new datasets to other researchers or developers for their ongoing work.
Software: 4 = Useful: I would recommend the new software to other researchers or developers for their ongoing work.
Knowledge Of Or Educated Guess At Author Identity: No
Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources
Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources
Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Add: **Author-Editor Confidential Comment**

-
=
≡

Official Comment by Authors

Official Comment

✎ Authors (👁 jay_kang@intuit.com (/profile?id=jay_kang@intuit.com), prarit_lamba@intuit.com (/profile?id=prarit_lamba@intuit.com), Xiang Gao (/profile?id=~Xiang_Gao4), Julian McAuley (/profile?id=~Julian_McAuley1), +2 more (/group/info?id=aclweb.org/ACL/ARR/2024/June/Submission4477/Authors))

📅 28 Jul 2024, 23:04 (modified: 22 Aug 2024, 15:39)

👁 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer 1AXu, Commitment Readers

🗐 Revisions (/revisions?id=2QLOVzgxHb)

Comment:
We sincerely appreciate Reviewer 1AXu for their valuable feedback. The reviewer pointed out a couple of limitations. We want to address them as follows:

Dependence on retrieval quality

As mentioned in the limitation section, the performance of RoleFact may vary based on the retrieval quality. However, we want to clarify that this limitation is not unique to our approach. All character role-play methods with retrieval components suffer from this limitation. As we have explained in the response to Reviewer 3yj6, off-the-shelf dense retrieval methods have some limitations in retrieving semantically similar utterances due to the repeated presence of named entities. The quality of dense retrieval methods can be significantly improved with task-specific fine-tuning. We leave this for future work.

Tuning confidence threshold

Even though we did not include the code for automated hyper-parameter search for the confidence threshold with this submission, we are very happy to include this with the publicly released codebase. We hope this will help generalize RoleFact across multiple datasets.

Add: **Author-Editor Confidential Comment**

Official Review of Submission4477 by Reviewer R364

Official Review Reviewer R364 17 Jul 2024, 07:10 (modified: 22 Aug 2024, 15:39)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer R364, Commitment Readers

Revisions (/revisions?id=JJjQKkP2Bv)

Paper Summary:

- This paper introduced a new character role-play dataset with more than 2,000 characters and 72,000 interviews, which is intended to evaluate diverse character hallucinations.
- The authors proposed a role-playing method, RoleFact, to mitigate the character hallucinations.

Summary Of Strengths:

- The authors tackled the critical problem of diverse character hallucinations, including cross-universe hallucination and temporal hallucination, and they collected a large dataset for evaluation.
- The authors proposed the RoleFact method, which empirically mitigates the character hallucinations by self-refining its knowledge and response.
- The authors provided various analyses, including minimizing parametric knowledge by anonymizing the prompts, scoring for less popular characters vs. more popular characters, ablation studies, etc.

Summary Of Weaknesses:

1. Missing key references

- While the SGR dataset is intended to evaluate diverse character hallucinations, it is not the first to tackle such hallucinations.
- For instance, [1] already introduced cross-universe hallucination, and [2] introduced temporal hallucination, where role-playing agents simulate time-sensitive characters.

2. Dataset

- While the authors utilized an automated dataset construction pipeline, they didn't provide any dataset quality check (i.e., human evaluation). For example, LLMs (e.g., GPT-4) are considered to have difficulty generating adversarial, open-ended, and scene-grounded interview questions.

3. Method

- Although the RoleFact method mitigated the character hallucinations, it will increase inference cost due to multiple text generation steps. It will be better to compare the inference cost of the proposed method with the others for a fair comparison.

- It would be better to add empirical results of the 'Ditto' method [1] and the 'Narrative-Experts' method [2] as baselines since they also deal with the relevant problems.

4. Evaluation Metrics and Empirical Results

- Did the authors utilize LLMs or humans to measure FactScore? If the former, they should demonstrate the reliability of evaluation metrics by comparing LLM judges with human judges. If the latter, please elaborate on the detailed process.

[1] Lu et al., "Large Language Models are Superpositions of All Characters: Attaining Arbitrary Role-play via Self-Alignment", ACL 2024

[2] Ahn et al., "TimeChara: Evaluating Point-in-Time Character Hallucination of Role-Playing Large Language Models", ACL 2024 Findings

Comments Suggestions And Typos:

- No details on prompt anonymization. Please explain how the authors constructed the anonymized prompt template.
- Please provide the THR of KGR and SR in Table 3.
- I'm curious about the multi-turn conversation scenarios beyond the single-turn interview question. Do role-playing agents keep hallucinating during multi-turn interactions?
- I'm also curious about the training-based method beyond the inference-based (i.e., RoleFact) method. Can smaller trainable agents (e.g., Character-LLM) be implemented to avoid such character hallucinations?

Confidence: 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.
Soundness: 3 = Acceptable: This study provides sufficient support for its major claims/arguments. Some minor points may need extra support or details.
Overall Assessment: 3 = Good: This paper makes a reasonable contribution, and might be of interest for some (broad or narrow) sub-communities, possibly with minor revisions.
Best Paper: No
Limitations And Societal Impact:
Yes

Ethical Concerns:
Copyright issues in the proposed dataset.

Needs Ethics Review: Yes
Reproducibility: 4 = They could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.
Datasets: 4 = Useful: I would recommend the new datasets to other researchers or developers for their ongoing work.
Software: 4 = Useful: I would recommend the new software to other researchers or developers for their ongoing work.
Knowledge Of Or Educated Guess At Author Identity: Yes
Knowledge Of Paper: Before the review process
Knowledge Of Paper Source: Preprint on arxiv
Impact Of Knowledge Of Paper: Not at all

Add: **Author-Editor Confidential Comment**

Rebuttal (comparison with relevant work, evaluation metric, etc.)

Official Comment

Authors (jay_kang@intuit.com (/profile?id=jay_kang@intuit.com), prarit_lamba@intuit.com (/profile?id=prarit_lamba@intuit.com), Xiang Gao (/profile?id=~Xiang_Gao4), Julian McAuley (/profile?id=~Julian_McAuley1), +2 more (/group/info?id=aclweb.org/ACL/ARR/2024/June/Submission4477/Authors))

28 Jul 2024, 13:02 (modified: 22 Aug 2024, 15:39)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer R364, Commitment Readers

Revisions (/revisions?id=4W9vgMQgzO)

Comment:
We genuinely appreciate the fact that Reviewer R364 strongly resonates with our paper’s motivation and contribution of the released dataset. We respond to the reviewer's concerns below:

Comparison with DITTO and TimeChara

We thank the reviewer for pointing out the very relevant works and we will cite them in our paper. The reviewer pointed out the DITTO paper (Lu et al., ACL 2024) as a prior example of cross-universe hallucination. However, we did not claim to be the first work that tackles cross-universe hallucinations. In fact, we already cited prior work such as Character-LLM ((Shao et al., EMNLP 2023) that tackles such hallucination. We have three important contributions compared to DITTO (and Character-LLM):

1. Unlike DITTO, our approach for mitigating cross-universe hallucinations is training-free.
2. Our dataset has a more comprehensive character-specific knowledge base. DITTO and Character-LLM use Wikipedia pages associated with characters to construct character-specific golden knowledge. If a character is asked a question that does not align with their Wikipedia-backed knowledge, characters are expected to reject that as irrelevant. On the other hand, the character-specific knowledge in our SGR dataset includes every scene, event, and conversation associated with a particular fictional character (mentioned in lines 196 - 206). This comprehensive knowledge is constructed from story-specific scripts. This makes it less likely that characters will falsely reject relevant knowledge as irrelevant.
3. DITTO and Character-LLM use ratings from LLM-judge for evaluating hallucinations which are highly subjective and unstable. The comprehensive story-specific knowledge in our dataset facilitates a more fine-grained evaluation of factual precision based on Fact Score.

TimeChara (Ahn et al., ACL 2024) is a contemporary work that tackles temporal hallucination. We did not know about this paper at the submission time of our paper since it was not publicly available as a refereed publication and even the arXiv preprint was only two weeks old. We have two key contributions compared to the TimeChara paper.

1. Our dataset has more diverse characters and a much larger dataset size. The TimeChara benchmark has 14 characters and 10,895 time-sensitive task samples. Our SGR dataset has 2000 diverse characters and 36,000 time-sensitive task samples (as mentioned in Section 3).
2. Our dataset supports significantly higher temporal granularity and flexibility of extension in terms of time-sensitive role-play. For example, in the Harry Potter series, the TimeChara dataset presents 25 unique points in time for evaluating hallucinations. For the same storyline, SGR presents around 1022 unique time points for scene-grounded interviews and 448 unique time points for the dialogue completion task. Additionally, the dataset construction process of TimeChara includes hand-crafted time patterns for each storyline (e.g. Halloween time, Christmas time for Harry Potter). Our dataset annotation process is story-agnostic. The knowledge base for each storyline has timestamps for each speech and non-speech event. Our annotations allow the research community to change the behavior of role-playing characters to a new point in time simply by changing the timestamp or create new time-sensitive tasks at any given timestamp without any human intervention. This flexibility may help the research community to explore temporal hallucination more thoroughly.

Evaluation metrics and empirical results

Fact Score (Min et al., EMNLP 2023) is an LLM-based evaluation metric for factual precision. Each response is decomposed into a set of atomic facts and each atomic fact is checked against a knowledge base for deciding whether the fact is supported or not. The Fact Score paper already demonstrates that LLM-generated Fact Score has high correspondence with human-generated factual precision. Subsequent works in the area of LLM hallucinations have demonstrated Fact Score to be a reliable fine-grained metric for evaluating hallucinations.

Reference:

1. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation (Min et al., EMNLP 2023)

Details on prompt anonymization

During the prompt anonymization process, we replace each character name and story name in the constructed prompt with anonymous story IDs and character IDs.

Multi-turn conversation and instruction-tuning method

Multi-turn hallucinations and instruction-tuning-based character role-play are beyond the scope of this work. However, we hope the large-scale diverse dataset we are releasing will be helpful for the research community to explore this direction in the future.

Inference cost and THR of remaining baselines

As per the suggestion of the reviewer, we will add the inference cost in terms of token consumption, and temporal hallucination performance of the other two baselines in our paper.

Add: **Author-Editor Confidential Comment**

Respone to Rebuttal

Official Comment Reviewer R364 29 Jul 2024, 00:29 (modified: 22 Aug 2024, 15:39)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer R364, Commitment Readers

Revisions (/revisions?id=VUByfMqVtw)

Comment:
Thank you for responding to my questions, especially clarifying the distinctions between SGR and related papers.

But I have some remaining concerns:

- Dataset: The authors haven't provided a quality check for the proposed SGR dataset. Are the questions (e.g., adversarial, open-ended, and scene-grounded) generated by GPT-4 reliable? Similarly, are the attributes (in Table 1) generated via script segmentation and character identification using GPT-3.5 trustworthy? GPT-4 and GPT-3.5 might generate misinformation, especially when processing less popular characters.
- Evaluation: While the Fact Score paper demonstrated the reliability of evaluation metrics (as the authors mentioned above), it was only evaluated in a "Wikipedia-retrieval" scenario. The authors, however, used the Fact Score to evaluate outputs by retrieving atomic facts from "story-specific scripts," which are more fine-grained and should be considered a different retrieval domain. Moreover, evaluating facts related to less popular characters would be even more challenging. So, it is still necessary to demonstrate the reliability of the Fact Score in this new dataset.

Given these considerations, I am maintaining the original score.

Add: **Author-Editor Confidential Comment**

Rebuttal (dataset quality, evaluation metric, etc.)

Official Comment

Authors (jay_kang@intuit.com (/profile?id=jay_kang@intuit.com), prarit_lamba@intuit.com (/profile?id=prarit_lamba@intuit.com), Xiang Gao (/profile?id=~Xiang_Gao4), Julian McAuley (/profile?id=~Julian_McAuley1), +2 more (/group/info?id=aclweb.org/ACL/ARR/2024/June/Submission4477/Authors))

30 Jul 2024, 10:43 (modified: 22 Aug 2024, 15:39)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer R364, Commitment Readers

Revisions (/revisions?id=efX2DLuqKh)

Comment:

We are glad that we were able to clarify the prior concerns of the reviewer. We want clarify the remaining two issues here:

Dataset Quality

The following components of the SGR dataset are manually verified (and corrected when appropriate):

1. The validation set which consists of 1000 samples from each task category (mentioned in line 240) is manually verified. The verified task samples constitute 5.5% of our task dataset. We have found that GPT-4 is a reliable interviewer for role-playing agents and there is precedence of using GPT-4 in both single-turn and multi-turn interview setting (Character-LLM, Shao et al., EMNLP 2023).
2. Story knowledge metadata such as story name, the list of characters, and role profiles are manually verified. The grouping and ordering of multiple stories from the same series is done manually as we have mentioned in line 193.

Associating character names with speech event can be challenging for LLMs when they are relying on their parametric memory, especially for less popular characters. However, this is not the case in our dataset construction process. All required annotations such as character names, speech utterances, action events, scene descriptions already exist in the collected script. LLMs do not create new annotations rather they convert unstructured story snippet into a structured format. This is less challenging and less noisy compared to LLM-as-annotator or LLM-as-judge setup where LLMs have to create annotations without a non-parametric reference knowledge. As requested by reviewer R364, we will add quality analysis in our paper.

Adapting Fact Score for Role-play

The process of adapting Fact Score for role-play is already addressed in the paper (evaluation metric section in lines 259 - 262). From manual inspection, we found one particular challenge of adapting Fact Score to role-playing domain. This is associated with atomic fact decomposition. Character role-play involves a lot of utterances with pronouns. If the decomposed atomic fact do not carry the name of the character, it can be difficult to verify this fact against reference knowledge during evaluation. In order to mitigate this problem, we convert all atomic facts into name only third person format. The prompt for atomic fact decomposition including the instruction of name-only third person conversion is shown in the paper in line 159.

Add: **Author-Editor Confidential Comment**

-
=
≡

Replying to Rebuttal (dataset quality, evaluation metric, etc.)

Response to Rebuttal

Official Comment Reviewer R364 30 Jul 2024, 17:01 (modified: 22 Aug 2024, 15:39)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer R364, Commitment Readers

Revisions (/revisions?id=xstc8b096P)

Comment:

Thank you for the reply and further clarifications.

Still, I'm worried about the reliability issue of the Fact Score metric.

- For instance, L259-262 merely explains that the authors adapted the Fact Score for role-play without thorough justification.
- While the authors addressed the pronoun problem through 'name-only third person conversion', other potential issues remain unexplored: (1) evaluation of facts about more vs. less popular characters, (2) other unidentified problems specific to role-playing scenarios, (3) the authors did not

quantify the reliability of the Fact Score metric by comparing it to human evaluations in the role-playing context.

- Again, **ensuring the reliability of the Fact Score metric in the role-playing domain is crucial for this paper, as it's an unexplored area that deserves more attention**. The authors should further explore and validate the reliability of the Fact Score metric in the context of role-playing evaluations in this dataset.

Given these considerations, I slightly increased the overall assessment from 2.5 to 3.

Add: **Author-Editor Confidential Comment**