Annotating Hallucinations in Question-Answering using Rewriting

Xu Liu♣, Guanyi Chen♣*, Kees van Deemter♡, Tingting He♣
♣Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning,
National Language Resources Monitoring and Research Center for Network Media,
School of Computer Science, Central China Normal University
♡Department of Information and Computing Sciences, Utrecht University
liuxu@mails.ccnu.edu.cn, {g.chen, tthe}@ccnu.edu.cn, c.j.vandeemter@uu.nl

Abstract

Hallucinations pose a persistent challenge in open-ended question answering (QA). Traditional annotation methods, such as span-labelling, suffer from inconsistency and limited coverage. In this paper, we propose a rewriting-based framework as a new perspective on hallucinations in open-ended QA. We report on an experiment in which annotators are instructed to rewrite LLM-generated answers directly to ensure factual accuracy, with edits automatically recorded. Using the Chinese portion of the Mu-SHROOM dataset, we conduct a controlled rewriting experiment, comparing fact-checking tools (Google vs. GPT-4o), and analysing how tool choice, annotator background, and question openness influence rewriting behaviour. We find that rewriting leads to more hallucinations being identified, with higher inter-annotator agreement, than span-labelling.

1 Introduction

With the rapid advancement of Large Language Models (LLMs), hallucination has emerged as a central research topic, spanning annotation (Vázquez et al., 2025), detection (Mickus et al., 2024), mitigation (Ji et al., 2023b), and interoperability (Zhang et al., 2024). Hallucinations are typically defined as content in a model's output that is not supported by the input (Dušek and Kasner, 2020; Ji et al., 2023a; van Deemter, 2024). To capture such phenomena, prior work (Thomson and Reiter, 2020; Thomson et al., 2023; Vázquez et al., 2025) proposed fine-grained annotation schemas that treat hallucination identification as a *span-labelling* task: annotators are asked to highlight spans in model outputs that constitute hallucinations, sometimes followed by labelling the error types or suggesting corrections.

More recently, research on hallucination has shifted focus from classical Natural Language Generation (NLG) tasks (e.g., summarisation and paraphrasing) to Question Answering (QA). Most newly introduced benchmarks (e.g., Li et al. (2023), Vázquez et al. (2025)) and mitigation techniques (e.g., Ji et al. (2023b)) now target QA. Compared to NLG, QA outputs can be significantly more open-ended. On one hand, this shift necessitates a rethinking of how hallucinations are defined, as QA responses are often grounded not in well-specified structured information (as is typically the case in Data-to-Text NLG) but in world knowledge.

On the other hand, the nature of QA presents challenges for existing span-labelling annotation schemas. These challenges include: **First**, the absence of a unified criterion for annotating factual errors in QA. Annotators may label different spans for the same error. For instance, some highlight only the head noun, others the full noun phrase, or even the entire sentence. **Second**, span-labelling often involves quickly scanning the text and marking visible errors without immediate feedback on how the corrected output would look. This can lead to overlooked minor mistakes as attention wanes. **Third**, QA answers frequently contain multiple factual claims, many of which are outside the annotators' domain knowledge. With limited access to external resources, collecting reliable annotations becomes nearly impossible.

To address these limitations and build a corpus that better facilitates the understanding of hallucinations in QA, we designed an annotation experiment based on the idea of *rewriting*. Rewriting has proven effective in evaluating the quality of model outputs, especially complex ones, as it encourages annotators to focus more carefully on errors (Chen et al., 2024). Specifically, using the Chinese portion[1] of the Mu-SHROOM corpus (a multilingual

*Corresponding Author

[1]We chose Chinese because it was the only language for

Proceedings of the 18th International Natural Language Generation Conference, pages 823–832
October 29–November 2, 2025.

Question	安德烈·克拉夫库克的电影作品有哪些? What are some of Andrei Konchalovsky's film works?
Answer	安德列·克拉克夫（Andrei Konchalovsky）是一位俄罗斯导演、编剧和制片人，他的作品包括：《俄罗斯方舟》（2011年）、《悲悯世界》(1991)、《莫斯科不相信眼泪》(18% 白人) (Moskva slezam ne verit, 1% blondynki)（10%的白人）等。 Andrei Konchalovsky is a Russian director, screenwriter, and producer. His works include: "Russian Ark" (2011), "A Pitying World" (1991), and "Moscow Does Not Believe in Tears" (18% White) (Moskva slezam ne verit, 1% blondes) (10% White), among others.

Table 1: An example item from the Chinese portion of Mu-SHROOM. The hallucinations are coloured red.

hallucination detection dataset annotated via span-labelling (Vázquez et al., 2025)), we asked annotators to directly rewrite LLM-generated answers to make them factually correct using WPS Word. The revision mode of WPS Word was used to record each rewriting operation. To mitigate issues related to external knowledge access, participants were provided with one of two tools: Google or GPT-4o. Using the revision logs, we built a corpus that records both the deleted content, deemed erroneous by annotators, and the insertions made to generate factually correct answers.

We hope this corpus can advance our understanding of hallucinations in QA. This study presents a series of statistical analyses exploring how tool usage, participant background, and question openness influence rewriting behaviours. We also compare our annotations with the original span-labelling-based annotations in Mu-SHROOM to assess whether rewriting helps annotators identify more errors and leads to higher inter-annotator agreement.

We begin by briefly reviewing the Mu-SHROOM dataset and describing the annotation process for hallucinations in its Chinese portion. Next, we present our rewriting experiment along with the corresponding hypotheses regarding rewriting behaviours. Finally, we test these hypotheses and discuss our findings.

2 The Mu-SHROOM Dataset

Mu-SHROOM (Vázquez et al., 2025) is a multilingual hallucination dataset designed to support research on detecting hallucinations produced by LLMs in QA across 14 languages. The dataset comprises questions and answers to these questions that were generated by various LLMs. Each answer is annotated with two types of token-level labels derived from human judgments: a *soft label*, indicating the proportion of annotators who identified the token as hallucinatory, and a *hard label*, indicating whether the token was marked as hallucinatory by the majority of annotators. An example QA pair in Chinese with hard labels is shown in Table 1.

To construct Mu-SHROOM, Vázquez et al. first manually selected 200 Wikipedia pages. For each page, they wrote one question that could be answered using information contained within the page. Then, they used LLMs to generate answers for each question. Among these LLM-generated answers, those that were fluent and relevant but appeared to contain hallucinations were manually selected and added to the corpus. [2]

Subsequently, annotators were hired to label hallucinations in the LLM-generated answers as a span-labelling task. During annotation, they were allowed to consult Wikipedia, not only the page associated with each question, but the entire encyclopedia. For the Chinese portion, each item was annotated by up to six annotators, with each annotator labelling 20 items.

Despite this carefully constructed process, the aforementioned limitations of span labelling sometimes lead to annotations that appear inconsistent or unintuitive. For example, in one case, only the first two characters of "制片人" (producer)[3] are marked as hallucinations. One additional contributing factor is the degeneration exhibited by the LLMs that generate these answers (Holtzman et al., 2019), which makes it difficult for annotators to determine precisely which parts should be labelled as hallucinations. For instance, in a degenerated segment such as "(18% 白人) (Moskva slezam ne verit, 1% blondynki)（10%的白人）", which is completely

which we were able to recruit enough native speakers for an in-lab experiment.

[2] Although Vázquez et al. wrote questions based on specific Wikipedia pages, the answers were not grounded in these pages, as the LLMs generated answers without accessing the original content. In other words, Mu-SHROOM still targets an open-ended QA task.

[3] In Chinese, "制片" means "to produce", whereas "人" is a noun that means "person"?

nonsensical in Chinese, some tokens are marked as hallucinations, while others are not. In addition, the annotations exhibit inconsistencies in handling symbols: characters like '%' and ')' are sometimes labelled as hallucinations and sometimes left unmarked.

3 Method

In this section, we elaborate on experimental settings for our rewriting experiments.

3.1 Materials and Design

We used the test set of the Chinese portion of Mu-SHROOM, which consists of 150 items, as the materials for our experiment.

As mentioned in the introduction, we aimed to give participants access to external knowledge as freely and conveniently as possible. While Wikipedia, used in Mu-SHROOM, contains ample information, it is not particularly user-friendly in this context. First, Wikipedia lacks a powerful search engine, making it difficult for participants to locate the relevant pages. Second, even when participants do find the appropriate pages, it remains challenging to extract the specific information they need from lengthy articles.

Instead, in this study, we explored two more powerful alternatives: Google and GPT-4o. When using Google for fact-checking, it often returns the most relevant Wikipedia article among the top results, with the desired content highlighted, making it easier to locate specific information. Powerful LLMs have been shown to be effective in assisting with factual claim annotation (Ni et al., 2024). To further enhance this approach, we employed GPT-4o with Retrieval-Augmented Generation (RAG), which not only gives participants access to one of the most capable LLMs to date, but also presents a list of reference articles upon which the model's answers are based, enabling more reliable and interpretable fact-checking. One potential concern is that, with access to GPT-4o, participants might simply replace the original answers with those generated by GPT-4o, rather than engaging in genuine rewriting. To minimise this risk, we explicitly instructed participants not to directly copy answers from GPT-4o. As a result, our experiment included two conditions: rewriting answers with assistance from Google and rewriting with assistance from GPT-4o.

To record participants' rewriting operations (specifically, which parts of the original answer were deleted and what new content was inserted), we used WPS Word[4] with the revision mode enabled during the experiment. This feature tracks deletions and insertions, storing them in a separate XML file.

Your task is to identify and correct incorrect information in a question-answering system, including but not limited to fabricated, false, inaccurate, or factually inconsistent content (such as dates, locations, people, events, etc.). Given a question and its corresponding system-generated answer, you are to: - Identify the errors in the answer and directly revise them in the original text to ensure factual accuracy; - Ensure the revised answer is concise and accurate, while preserving the original structure and style as much as possible. Avoid major changes unless necessary. - During the revision process, you may use ChatGPT/Google to consult relevant information as a reference, but you **MUST** not copy and paste directly. Please synthesise information from multiple sources to ensure the accuracy and reliability of the final answer.

Table 2: The translated instruction.

3.2 Procedure

We first conducted a small pilot study with five participants to evaluate our instructions and determine how many items each participant should rewrite. We observed that participants typically spent 3–5 minutes revising each answer. Based on this, we divided the 150 items into 10 groups, each consisting of 15 items.

The experiment began with a form requesting demographic information from each participant. We collected data on participants' educational background, age, and gender. They were informed that this information would be used solely for scientific purposes and would not be made publicly available.

Participants were asked to complete the experiment following the instructions provided in Table 2. They were informed that they would receive a set of questions paired with answers generated by LLMs, and their task was to identify and correct any in-

[4] https://zh-hant.wps.com/

accurate information in the answers. Specifically, the instructions stated that participants should: (1) directly revise the answers within the original text; (2) ensure that the revised answers are accurate with respect to the questions; and (3) use ChatGPT[5]/Google (depending on the experimental condition) for fact-checking, without directly copying and pasting content from these sources.

Additionally, to prevent unnecessary alterations or the introduction of over-specific information, participants were instructed to preserve the structure and style of the original answers as much as possible and to avoid making major changes unless clearly necessary.

Each participant revised answers from a specific group of 15 items under one experimental condition, either using Google or GPT-4o. Throughout the session, we recorded the total time each participant spent completing the task.

3.3 Participants

To ensure that each answer was rewritten by two participants under each condition (i.e., a total of four rewritings per answer), we recruited $10 \times 4 = 40$ participants. Among them, 25 were male and 15 were female, with a mean age of 22.75 years. 25 participants had educational backgrounds in computer science–related fields (e.g., artificial intelligence, software engineering), while the remaining 15 came from other disciplines. On average, participants spent 61.45 minutes completing the experiment. Each participant was compensated with 40 RMB (approximately $\$5.5$), which is nearly double the minimum hourly wage in China.

4 Hypotheses

We aim to understand hallucinations in QA by examining how humans correct answers that contain factual errors. As an initial step, we are primarily interested in *the factors that influence how people revise answers produced by LLM-based QA systems*. We are also interested in *the effectiveness of rewriting as a method for understanding hallucinations, in comparison to span-labelling*. To answer these two research questions, we form the following hypotheses.

First, we gave participants access to two external tools: Google and GPT-4o. As noted by Ni et al. (2024), LLMs are effective tools for assisting in hallucination annotation. Compared to search engines like Google, LLMs have a stronger ability to interpret annotators' queries, reducing the need for careful query formulation, and to summarise the most relevant information from retrieved sources, thereby streamlining the fact-checking process. Based on this, our first hypothesis (referred to as hypothesis $\mathcal{H}_1$) was that, compared to participants using Google for fact-checking, those using GPT-4o would: (1) correct answers more quickly, as measured by the time spent revising a given number of answers; (2) identify more hallucinations, as indicated by the number of tokens deleted from the original answers; and (3) add more information, as measured by the number of tokens inserted into the revised answers.

Second, intuitively, participants with a background in computer science or artificial intelligence may be likely have a better understanding of hallucinations and greater proficiency with tools such as Google and GPT-4o. Accordingly, we expected these participants to: (1) complete the task more quickly, (2) identify more hallucinations, and (3) add more information to the revised answers (referred to as hypothesis $\mathcal{H}_2$).

Third, the openness of a question may influence how participants correct the answer. Closed-ended questions (e.g., "What countries are there in the Americas?") typically require concrete, fact-based answers, such as lists of entities, that are relatively easy to verify. In contrast, open-ended questions (e.g., "What is the history of America?") tend to elicit narrative responses that integrate diverse types of information, making fact-checking more complex. Thus, it seems plausible to hypothesise that, compared to closed-ended questions, rewritings in response to open-ended questions would involve: (1) more identified hallucinations, (2) more added information, and (3) lower inter-annotator agreement (referred to as hypothesis $\mathcal{H}_3$).

Finally, as discussed in the introduction, we hoped that, compared to conventional span–labelling–based hallucination annotation, rewriting would help annotators remain more engaged and focused. Based on this assumption, we anticipated that rewriting would lead annotators to: (1) identify more hallucinations, (2) add more information, and (3) achieve higher inter-annotator agreement (referred to as hypothesis $\mathcal{H}_4$).

[5]For simplicity, we used the term "ChatGPT", though GPT-4o was actually used.

```
"qid": 1,
"uid": 1,
"question": "What is David Sandberg's directorial debut?"
"original_answer": "David Sandberg's directorial debut was the 1993 film Sex, Lies, and
Videotape. This film won the Palme d'Or at the Cannes Film Festival and became a milestone
in independent cinema at the time.",
"modified_answer": "David Sandberg's directorial debut was the 2016 film Lights Out.",
"original_with_deletions": "David Sandberg's directorial debut was the <del>1993</del>
film <del>Sex, Lies, and Videotape</del>. <del>This film won the Palme d'Or at the Cannes
Film Festival and became a milestone in independent cinema at the time.</del>",
"modified_with_insertions": "David Sandberg's directorial debut was the <ins>2016</ins>
film <ins>Lights Out</ins>."
```


Figure 1: An example of a post-processed data entry (translated from Chinese).

5 Data Post-processing

To test the hypotheses in Section 4 and enable further research based on our dataset, we post-processed the recorded rewriting operations and computed Inter-Annotator Agreements (IAA). The processed data is available at: https://github.com/a-quei/hallucination-rewriting.git.

Extracting Rewriting Operations. To identify which tokens were considered hallucinations (i.e., deleted by participants) and what information was added to produce a corrected answer, we developed scripts to extract deletion and insertion operations from the revision records generated by WPS Word's revision mode.

All extracted information was saved into a JSON file, with each entry, an example of which is shown in Figure 1, representing the complete set of rewriting operations performed by one participant on a single answer. Each entry contains the following seven fields: (1) "qid": the question ID; (2) "uid": the participant ID; (3) "question": the question text; (4) "original_answer": the answer before rewriting; (5) "modified_answer": the answer after rewriting; (6) "original_with_deletions": the original answer with deleted tokens marked using <del> </del> tags; (7) "modified_with_insertions": the revised answer with inserted tokens marked using <ins> </ins> tags. There are 600 entries in total.

Annotating Openness. In hypothesis $\mathcal{H}_3$, we suggested that rewriting behaviour may be strongly influenced by whether a question is open-ended or closed-ended. In this study, based on the questions in Mu-SHROOM, we operationally defined closed-ended questions as those that can be answered with a concrete, finite set of entities or numbers, for example, a list of movies or countries. All other questions were categorised as open-ended.

The first two authors of this paper independently annotated all 150 questions and resolved any disagreements through discussion. This process resulted in 59 open-ended questions and 91 closed-ended questions.

Although the questions in Mu-SHROOM were designed to be "closed" to specific Wikipedia pages (i.e., the answer to the question is contained in the page; see Section 2), they are not necessarily closed-ended. For example, the question "What is the history of America?" can indeed be answered using content from the Wikipedia page on America, but it still qualifies as an open-ended question.

Inter-Annotator Agreement. Both hypotheses $\mathcal{H}_3$ and $\mathcal{H}_4$ referred to IAA.

Given an answer, our focus was on annotators' agreement regarding which content should be considered hallucinations (and thus subject to rewriting) rather than on what information should be added in the revised answer. Therefore, we measured IAA based solely on deletion operations, not insertion operations.

We did not adopt the Intersection over Union (IoU) metric used by Vázquez et al. (2025), which roughly measures the overlap between two annotations. This is because IoU disregards unmarked tokens (i.e., tokens that are not labelled as hallucinations) when computing IAA, and it does not generalise well to settings with more than two annotators. Instead, we computed the token-level *observed agreement*, which is in fact the binary version of Fleiss' κ (Fleiss, 1971). Formally, for the token i, the agreement a_i is defined as:

$$a_i = 1 - \frac{k_i(n - k_i)}{n(n-1)/2} \tag{1}$$

where k_i is the number of annotators who annotate token i, and n is the total number of annotators.

In this way, the agreement is 1 (i.e., 100%) when all annotators agree to mark or not mark the token. Then, the agreement of annotating an answer A_j is the average of the agreements of all tokens in it:

$$A_j = \frac{1}{N_j} \sum_{i=1}^{N_j} a_i \tag{2}$$

where N_j is the number of tokens in answer j. Using this approach, the average agreement of our experiment is 79.61%.

6 Results

Here we report on our testing of the four hypotheses we put forward in Section 4.

6.1 The impact of Tool Use and Background

Our first two hypotheses (i.e., $\mathcal{H}_1$ and $\mathcal{H}_2$) proposed that two key factors influence how humans correct hallucinations in QA: (1) tool use, specifically whether participants use Google or GPT-4o for fact-checking assistance; and (2) background, referring to whether participants have a background in computer science or artificial intelligence (hereafter CS) or not (hereafter non-CS).

Figure 2 shows how tool use and participant background affect three aspects of rewriting behaviour: (1) the time spent completing the task; (2) the amount of hallucinations identified, measured by the number of deletions; and (3) the amount of added information, measured by the number of insertions. We used independent-samples t-tests to assess whether these effects were statistically significant.

Contrary to our expectations in $\mathcal{H}_1$, the results for tool use showed no significant differences. Participants spent a comparable amount of time completing the rewritings using GPT-4o (M=61.10, SD=15.29) and Google (M=61.80, SD=11.62); the difference was not statistically significant (t=-0.163, p=0.874). Similarly, GPT-4o did not lead to significantly more identified hallucinations (t=-0.573, p=.570) or more added information (t=-0.419, p=.677) compared to Google. In fact, participants using Google identified slightly more hallucinations (M=3666.75, SD=1310.11) than those using GPT-4o (M=3441.85, SD=1170.09), and also added slightly more information (Google: M=1694.80, SD=999.42; GPT-4o: M=1574.85, SD=797.75). One possible explanation is that participants were more accustomed to using search engines like Google as knowledge acquisition tools than LLMs such as GPT-4o.

Regarding participant background, the results again ran counter to our expectations in $\mathcal{H}_2$. Participants with a CS background identified significantly fewer hallucinated tokens (M=3246.16, SD=1226.28) than those without a CS background (M=4067.87, SD=1092.51, t=-2.198, p=.035). A possible explanation is that most non-CS participants had a background in the humanities, while many of the questions in Mu-SHROOM pertain to topics such as history or art—areas more closely aligned with the humanities. In other words, although non-CS participants may have been less familiar with computational tools, they were likely more familiar with the subject matter of the questions. We also found that CS and non-CS participants spent a similar amount of time completing the task (t=0.670, p=.507) and inserted a comparable number of tokens into their rewritings (t=-0.042, p=.967).

6.2 Open-ended vs. Close-ended Questions

Hypothesis $\mathcal{H}_3$ mentioned the impact of question types on the rewriting behaviours: it argued that rewritings for open-ended questions may have more identified hallucinations, more added information and lower IAA. Since IAA scores computed using observed agreement are not normally distributed, we used the Mann–Whitney U for this hypothesis.

In line with our expectation, corrections to answers for open-ended questions contained significantly more insertions than those for closed-ended questions (U=1786.5, p<.001), suggesting that participants tended to add more information when addressing open-ended questions, likely due to their greater complexity.

We did not observe a significant effect of question openness on either inter-annotator agreement (U=2824, p=.593) or the number of identified hallucinations (U=2509, p=.501). This may be because, compared to closed-ended questions, answers to open-ended questions do not necessarily contain more factual errors, and therefore do not inherently pose greater challenges for fact-checking.

6.3 Role-labelling vs. Rewriting

We proposed using rewriting as a method for annotating hallucinations with the expectation that it would help annotators remain more focused. Accordingly, in $\mathcal{H}_4$, we hypothesised that rewriting

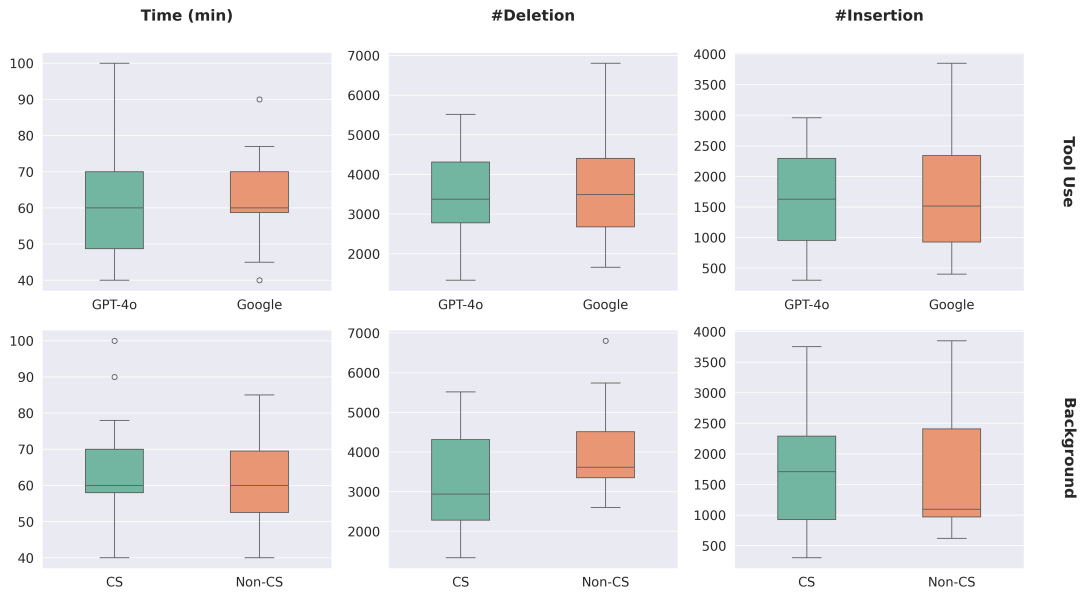

Figure 2: Relationships between Tool Use/Background and (1) time spent to complete the task; (2) the number of deleted tokens; and (3) the number of inserted tokens.

would lead to higher IAA and more identified hallucinations compared to conventional span-labelling (as in Mu-SHROOM). To test this hypothesis, we compared the two approaches along these two dimensions.

In Mu-SHROOM, a token is considered hallucinated if the proportion of annotators who label it as such exceeds a specified agreement threshold. By default, this threshold is set to 50% (see Section 2 for details). We adopted this approach in our rewriting experiment: a token is considered a hallucination if the proportion of annotators who chose to edit it exceeds a given agreement threshold. To enable a systematic comparison between hallucinations annotated via our rewriting approach and those identified through span-labeling in Mu-SHROOM, we evaluated three agreement thresholds, 50%, 75%, and 100%, where 100% indicates that a token is considered hallucinated only if all annotators agreed.

Table 3 presents the average number of hallucinations (averaged over the 150 answers) identified using either span-labelling or rewriting under three agreement thresholds. In line with our expectation in $\mathcal{H}_4$, paired t-tests confirmed that rewriting led to significantly more identified hallucinations at all thresholds: 100% (t = -7.997, p < .001), 75% (t = -9.239, p < .001), and 50% (t = -6.312, p < .001). These results suggest that rewriting helps annotators remain more focused, thereby enabling them to detect more hallucinations.

	100%	**75%**	**50%**
Span-labelling	30.20	87.25	195.10
Rewriting	156.75	225.66	269.18

Table 3: Average number of tokens identified as hallucinations using span-labelling and rewriting, evaluated under different agreement thresholds.

Additionally, the results show that the number of identified hallucinated tokens is less sensitive to changes in the agreement threshold when using rewriting compared to span-labelling. Even under the strictest threshold of 100%, participants using rewriting still identified an average of 156.75 hallucinated tokens, substantially higher than the 30.20 tokens identified through span-labelling. This suggests that rewriting not only helps participants identify more hallucinations overall, but also leads to more consistent annotation across annotators.

To further analyse IAA quantitatively, we adapted the IAA computation described in Section 5 to the Mu-SHROOM dataset using its soft labels. Specifically, the soft label of token i represents the probability p_i of this token being marked as hallucinated, i.e., $p_i = k_i/n$. Equation 1 can be written as:

$$a_i = 1 - \frac{2np_i(1-p_i)}{n-1} \tag{3}$$

Equation 2 was used to measure the IAA of each answer in Mu-SHROOM.

	Rewriting	Span-labelling
100%	安德列·克拉克夫（Andrei Konchalovsky）是一位俄罗斯导演、编剧和制片人，他的作品包括：《俄罗斯方舟》（2011年）、《悲悯世界》(1991)、《莫斯科不相信眼泪》(18% 白人) (Moskva slezam ne verit, 1% blondynki)（10%的白人）等。	安德列·克拉克夫（Andrei Konchalovsky）是一位俄罗斯导演、编s剧和制片人，他的作品包括：《俄罗斯方舟》（2011年）、《悲悯世界》(1991)、《莫斯科不相信眼泪》(18% 白人) (Moskva slezam ne verit, 1% blondynki)（10%的白人）等。
75%	安德列·克拉克夫（Andrei Konchalovsky）是一位俄罗斯导演、编剧和制片人，他的作品包括：《俄罗斯方舟》（2011年）、《悲悯世界》(1991)、《莫斯科不相信眼泪》(18% 白人) (Moskva slezam ne verit, 1% blondynki)（10%的白人）等。	安德列·克拉克夫（Andrei Konchalovsky）是一位俄罗斯导演、编剧和制片人，他的作品包括：《俄罗斯方舟》（2011年）、《悲悯世界》(1991)、《莫斯科不相信眼泪》(18% 白人) (Moskva slezam ne verit, 1% blondynki)（10%的白人）等。
50%	安德列·克拉克夫（Andrei Konchalovsky）是一位俄罗斯导演、编剧和制片人，他的作品包括：《俄罗斯方舟》（2011年）、《悲悯世界》(1991)、《莫斯科不相信眼泪》(18% 白人) (Moskva slezam ne verit, 1% blondynki)（10%的白人）等。	安德列·克拉克夫（Andrei Konchalovsky）是一位俄罗斯导演、编剧和制片人，他的作品包括：《俄罗斯方舟》（2011年）、《悲悯世界》(1991)、《莫斯科不相信眼泪》(18% 白人) (Moskva slezam ne verit, 1% blondynki)（10%的白人）等。

Table 4: Hallucination annotations of the example QA pair in Table 1 using Rewriting and Span-labelling with different agreement threshold. See Table 1 for the English translations.

A Wilcoxon Signed-Rank Test confirms that rewriting (M=0.756) leads to a significantly higher IAA than span-labelling (M=0.462, p<.001), which is again consistent with our hypothesis $\mathcal{H}_4$.

A Case Study. To further illustrate the differences in annotation outcomes, Table 4 presents how the example in Table 1 was annotated using both rewriting and span-labelling, under varying agreement thresholds.

Focusing on the annotations from rewriting, we observed that the results remained relatively stable when increasing the agreement threshold from 50% to 100%. Only two major changes emerged: (1) Some participants did not edit the phrase "安德列·克拉克夫（Andrei Konchalovsky）" due to disagreement over the correctness of the transliteration. This arose because the question (see Table 1) used a slightly different version: "安德烈·克拉夫库克". Some participants considered the discrepancy a hallucination, while others did not. (2) One participant failed to identify the phrase "《俄罗斯方舟》（2011年）" as hallucinated, resulting in its exclusion from the annotation when the agreement threshold was raised to 100%. Encouragingly, all participants consistently identified the hallucinations associated with the degenerated content (i.e., the text following "(18% 白人)"), suggesting a strong shared understanding in such cases.

In contrast, when using span-labeling for hallucination annotation, we observed several limitations: (1) under a high agreement threshold (i.e., 100%), only a small number of tokens were annotated as hallucinated; (2) under a low threshold (i.e., 50%), some factual content was incorrectly marked, for example, "制片" (indicating that Andrei Konchalovsky was a film producer), which is actually accurate; (3) there were numerous annotation inconsistencies, such as whether to label symbols like "%" and ")"; and (4) a few annotators failed to recognize hallucinations caused by degenerated text, resulting in their exclusion when the agreement threshold was set to 100%.

This comparison highlights that, compared to span-labelling, rewriting yields hallucination annotations that are more accurate, more consistent, and more comprehensive.

7 Conclusion

This paper proposes rewriting as a new lens through which to observe hallucinations in open-ended question answering. By instructing annotators to revise LLM-generated answers directly, while tracking their edits, we obtained a corpus that not only highlights hallucinated content but also reveals how such content is corrected. We then analysed factors that may influence rewriting behaviours. Contrary to expectations, the choice of fact-checking tool (Google vs. GPT-4o) and participants' technical background had limited impact on annotation quality, while question openness primarily affected the amount of added information. Compared to traditional span-labelling-based hallucination anno-

tation, our analyses show that rewriting leads to the identification of more hallucinations and yields higher inter-annotator agreement. These findings suggest that rewriting encourages more engaged and consistent annotation, making it a promising alternative for the creation of datasets that can enhance the research community's understanding of hallucination in NLG, not only in Question Answering but potentially in other NLG tasks as well.

Limitations

While this paper introduces rewriting as a novel framework for understanding hallucinations in open-ended QA, our focus has primarily been on presenting the experimental design and analysing rewriting behaviour at a high level. We have not yet conducted an in-depth content analysis of what types of information were marked as hallucinated and subsequently deleted, nor what kinds of factual content were added during rewriting. Such qualitative and semantic analyses would be valuable for understanding the nature of hallucinations more precisely. It would also be useful to distinguish hallucination from over-specification (Chen and van Deemter, 2023)—that is, cases where a model introduces unnecessary detail that may be factually correct but exceeds the contextual requirements. Differentiating these phenomena could sharpen our analyses and support the development of a more precise taxonomy of model errors.

Moreover, in our current framework, we approximate hallucinations as the content that annotators deleted during rewriting. This assumption, while practical for analysis, is imperfect. In principle, a hallucination could also be corrected through insertion (e.g., by clarifying or qualifying an existing claim without removing it). However, upon inspecting our corpus, we found no clear instances of hallucinations being corrected solely via insertions, suggesting that our analyses in this work are valid.

Finally, our quantitative analysis focused more heavily on deletions than insertions. While deletions offer a clearer signal of hallucination detection, insertions may capture important nuances in how annotators choose to revise or expand answers. Future work should investigate the interplay between deletions and insertions to develop a more comprehensive understanding of hallucination correction strategies.

Acknowledgments

We are grateful for the comments from reviewers. Guanyi Chen is supported by the start-up funds of Central China Normal University (No.31101232053) and Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning (No.2025AISL002).

References

Guanyi Chen, Fahime Same, and Kees Van Deemter. 2024. Intrinsic task-based evaluation for referring expression generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7220–7231, Bangkok, Thailand. Association for Computational Linguistics.

Guanyi Chen and Kees van Deemter. 2023. Varieties of specification: Redefining over-and under-specification. *Journal of Pragmatics*, 216:21–42.

Ondřej Dušek and Zdeněk Kasner. 2020. Evaluating semantic accuracy of data-to-text generation with natural language inference. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 131–137, Dublin, Ireland. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023a. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023b. Towards mitigating LLM hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843, Singapore. Association for Computational Linguistics.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Timothee Mickus, Elaine Zosa, Raul Vazquez, Teemu Vahtola, Jörg Tiedemann, Vincent Segonne, Alessandro Raganato, and Marianna Apidianaki. 2024.

SemEval-2024 task 6: SHROOM, a shared-task on hallucinations and related observable overgeneration mistakes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1979–1993, Mexico City, Mexico. Association for Computational Linguistics.

Jingwei Ni, Minjing Shi, Dominik Stammbach, Mrinmaya Sachan, Elliott Ash, and Markus Leippold. 2024. AFaCTA: Assisting the annotation of factual claim detection with reliable LLM annotators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1890–1912, Bangkok, Thailand. Association for Computational Linguistics.

Craig Thomson and Ehud Reiter. 2020. A gold standard methodology for evaluating accuracy in data-to-text systems. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 158–168, Dublin, Ireland. Association for Computational Linguistics.

Craig Thomson, Ehud Reiter, and Barkavi Sundararajan. 2023. Evaluating factual accuracy in complex data-to-text. *Computer Speech & Language*, 80:101482.

Kees van Deemter. 2024. The pitfalls of defining hallucination. *Computational Linguistics*, 50(2):807–816.

Raúl Vázquez, Timothee Mickus, Elaine Zosa, Teemu Vahtola, Jörg Tiedemann, Aman Sinha, Vincent Segonne, Fernando Sánchez-Vega, Alessandro Raganato, Jindřich Libovický, Jussi Karlgren, Shaoxiong Ji, Jindřich Helcl, Liane Guillou, Ona de Gibert, Jaione Bengoetxea, Joseph Attieh, and Marianna Apidianaki. 2025. SemEval-2025 Task 3: Mu-SHROOM, the multilingual shared-task on hallucinations and related observable overgeneration mistakes.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. 2024. How language model hallucinations can snowball. In *International Conference on Machine Learning*, pages 59670–59684. PMLR.