Effective Prompt-tuning for Correcting Hallucinations in LLM-generated Japanese Sentences

Haruki Hatakeyama, Masaki Shuzo, Eisaku Maeda
Tokyo Denki University
{23amj18@ms, shuzo@mail, maeda.e@mail}.dendai.ac.jp

Abstract

We propose a method to efficiently correct hallucinations occurring in Large Language Models (LLMs) using LLMs themselves. Previous studies have used a pipelined method, multiple prompts (MP) to correct hallucinations, but this approach had the problem of requiring significant calculation cost. Therefore, in this study, we use a single prompt (SP) that integrates the process to detect and correct hallucinations. In the proposed method, we instruct the LLM using SP to generate a corrected sentence if a hallucination is present, and not to modify the text if no hallucination is occurring. We compare SP with MP in terms of calculation time and correcting accuracy. Additionally, we examine the effectiveness of hallucination correcting with Chain-of-Thought (CoT). Experimental results show that SP achieves correcting with reduced calculation time compared with MP. Furthermore, we revealed that while correcting with CoT decreases the correcting accuracy of MP, it improves that of SP.

1 Introduction

The evolution of Large Language Models (LLMs) has become more prominent through models such as GPT-4 (Achiam et al., 2023) and Claude[1]. These models are capable of generating more natural text. As a result, LLMs are being put into practical use in a wide range of applications such as ChatGPT[2] and Perplexity AI[3].

However, LLMs have the potential to generate hallucinations, which poses a significant challenge in practical use. It has been reported that in open-domain text generation, GPT-3.5-turbo generates hallucinations at a rate of 17.7%, while GPT-4 does so at 15.7% (Mündler et al., 2024).

As a method to suppress hallucination, RAG (Lewis et al., 2020b) is mentioned. RAG retrieves information from an external database and uses that information as a reference for the LLM to generate text. Since it retrieves information externally, it is reported to be capable of generating information that has not been learned and suppressing hallucination (Shuster et al., 2021). However, it has been reported that LLMs can add new information to the text generated by the search-provided external information (Dziri et al., 2022) or prioritize internal knowledge over external information (Longpre et al., 2021; Xie et al., 2024). Therefore, there is still a possibility of hallucination occurring even when using RAG.

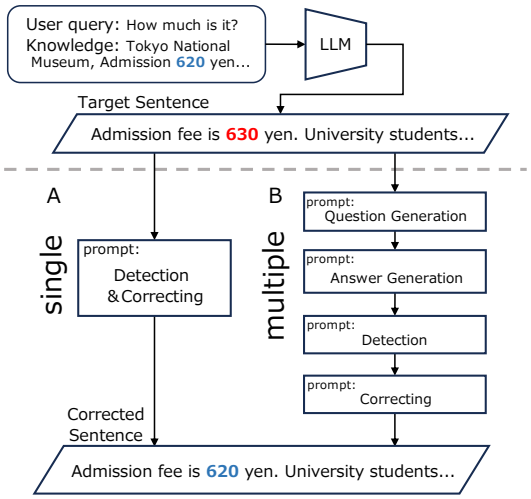

Figure 1: Processing flow of methods for correcting hallucinations contained in the target sentence. A: correcting using single prompt (proposed method). B: correcting using multiple prompt (Dhuliawala et al., 2024).

To solve the problem of hallucination, methods have been proposed to detect and correct hallucinations. Correcting can be applied to LLMs that have implemented RAG. Furthermore, methods have been proposed to use LLMs themselves for correcting.

Many of these methods perform correcting through multiple prompts (MP) (Zhao et al., 2023;

[1]https://www.anthropic.com/news/introducing-claude
[2]https://openai.com/blog/chatgpt
[3]https://www.perplexity.ai/

Table 1: Actual prompts used in the proposed method. The system estimates the contradiction implication relationship between the knowledge and the target sentence, and gives instructions to correct the contradiction relationship. The red characters indicate the strings of characters used when using CoT. The text shown here is the English translation of the Japanese original.

#Tasks

• You can infer contradictions and implication relationships between knowledge and target sentences.
• If there is a contradiction between knowledge and the target sentence, you can correct the target sentence.
• Output the reasoning for determining whether there are contradictions between knowledge and the target sentence, and based on this reasoning, output a judgment label and a corrected sentence.
• Outputting a 0 label means the knowledge and target sentence have an implication relationship.
• Outputting a 1 label means the knowledge and target sentence have a contradiction relationship.
• If you output a 0 label, since the knowledge and target sentence have an implication relationship, output " correcting: None ".
• If you output a 1 label, since the knowledge and target sentence have a contradiction relationship, correct the target sentence based on the reasoning that indicates which part of the target sentence should be modified.
• If the target sentence contains information not present in the knowledge or information that contradicts the knowledge, output the reasoning for why it's considered a contradiction, and correct the target sentence based on the knowledge and reasoning.
• Maintain the format of the target sentence.

#Instructions

• Always follow the rules.
• Strictly adhere to the output format.
• Make judgments based on the reasoning.
• Detect any contradictions between the knowledge and target sentence.
• Output 0 if the knowledge and target sentence have an implication relationship.
• Output 1 if there are contradictions between the knowledge and target sentence.
• Carefully examine the knowledge and target sentence to determine if there's a contradiction or implication relationship and output the label.
• If you determine implication, output " correcting: None ".
• If you determine contradiction, correct the target sentence based on the knowledge and reasoning.
• When correcting, faithfully revise the target sentence based on the knowledge.
• If information not present in the knowledge exists in the target sentence, delete it.
• Please refer to the following specific examples.

#Specific Examples

##Specific Example 1 (Implication Relationship)
##Input
Knowledge: Business hours: 10 AM to 10 PM (10 AM to 9 PM from January to March), admission until 20 minutes before closing, Open: Every day
Target sentence: Admission is until 20 minutes before closing.
##Output
Reasoning: The target sentence "Admission is until 20 minutes before closing." does not contradict the knowledge "admission until 20 minutes before closing,".
Judgment: 0
correcting: None
• • •

##Specific Example 7 (Contradiction Relationship)
##Input
Knowledge: Operating hours: 10:00 AM to 7:00 AM the next day, Closed: Never
Target sentence: It's from 10 AM to 7 PM.
##Output
Reasoning: The target sentence "It's from 10 AM to 7 PM." states 7 PM, but the knowledge "Operating hours: 10:00 AM to 7:00 AM the next day, Closed: Never" indicates 7:00 AM the next day. Therefore, the target sentence contradicts the knowledge. As a result, the target sentence should be corrected to "It's from 10 AM to 7 AM the next day."
Judgment: 1
correcting: It's from 10 AM to 7 AM the next day.

#input/output

##Input
Knowledge: {knowledge} Target sentence: {target sentence}
##Output

To reiterate, you should complete the following tasks: #Tasks You can infer contradictions and implication relationships between knowledge and target sentences. If there is a contradiction relationship between knowledge and the target sentence, you can correct the target sentence. Compare the knowledge and target sentence. If you determine that the target sentence implies the knowledge, output the target sentence without correcting. However, if the target sentence contains information (contradictions) not present in the knowledge, please correct the target sentence.

Mündler et al., 2024; Dhuliawala et al., 2024). They design prompts that break down tasks into phases such as query generation for external knowledge search, hallucination detection, and correcting, and incorporate these in a pipeline to tackle hallucination correcting.

These existing studies have developed models using MP to perform correcting of hallucinations in complex tasks by subdividing the tasks. However, the challenge with MP is that they involve multiple processes, which increases computational costs.

Applications like ChatGPT and Perplexity AI have made LLMs more accessible by utilizing them in a conversational format. In such dialogue-based interactions, real-time responsiveness becomes crucial. Therefore, there is a demand for hallucination correcting methods that can minimize the calculation time as much as possible.

We propose a hallucination correcting method constructed using only a single prompt (SP) to address the challenges in existing studies. Our proposed method is characterized by its ability to simultaneously detect and correct hallucinations.

In this study, we conducted a comparative evaluation of methods for correcting hallucinations using SP and MP, focusing on calculation time and correct capability. Furthermore, we applied the Chain-of-Thought (CoT) (Wei et al., 2022), which has been reported to be effective in various tasks, and analyzed its effects in detail.

The analysis yielded the following findings:

- It was confirmed that SP could significantly reduce calculation time while maintaining correcting accuracy equal to or better than MP.
- The effect of CoT in hallucination correcting was found to be strongly dependent on prompt design. In particular, the combination of SP

and CoT was shown to be most effective in hallucination correcting.

- It was revealed that SP is a method that minimizes the reduction in recall observed with CoT.
- In the case of MP, it was suggested that the reduction in recall caused by the application of CoT could lead to a decrease in correct capability.

2 Related Work

2.1 Correcting by Fine-tuning

Methods have been proposed to perform hallucination correcting using a pipeline approach, training BERT (Devlin et al., 2019) as a detector and models such as BART (Lewis et al., 2020a) and T5 (Raffel et al., 2020) as correctors, using hallucination data (Thorne et al., 2021; Lee et al., 2022). However, these methods face the challenge of error propagation due to the combination of multiple models.

Addressing this issue, Moriwaki et al. (2022) proposed a joint learning method that shares part of the loss function, reporting improved accuracy of the corrector. Conversely, Cao et al. (2020) proposed a method to correct hallucinations using a single model, enabling correcting without constructing a separate detector.

While these existing studies use fine-tuning, they require new training data to handle hallucinations across various domains and tasks. However, preparing data that corresponds to the diverse domains in the real world is challenging. Therefore, there is a demand for methods that can address hallucinations occurring in various domains without being dependent on specific domains.

2.2 Correcting by Prompt-tuning

LLMs can perform tasks without fine-tuning by using In-context Learning (ICL) with few-shot prompts (Brown et al., 2020). As ICL provides better generalization accuracy than fine-tuning (Awadalla et al., 2022; Si et al., 2023), it is considered suitable for hallucination correcting across various domains.

Existing hallucination correcting methods using LLMs adopt MP approach to break down tasks into smaller subtasks. Dhuliawala et al. (2024) proposed a method that uses different prompts for question generation, answer generation, detection, and correcting stages to detect and correct hallucinations in list-based QA and long-form text generation tasks. Their method follows the flow shown in Figure1:B. Zhao et al. (2023) also developed a method that uses multiple prompts to generate intermediate steps of CoT, detect hallucinations by calculating agreement rates, and perform correcting using external knowledge with another prompt. Mündler et al. (2024) constructed a framework that uses different prompts in three stages - generation, detection, and correcting - to address self-contradictions in LLMs.

These existing studies use MP for correcting hallucinations in complex tasks, but this results in high computational costs. Therefore, there is a need for prompt designs that integrate MP and enable more efficient detection and correcting of hallucinations.

Table 2: Results of manually annotating 50,000 outputs obtained by "hobbyist" (Sugiyama et al., 2021). The numbers not enclosed in brackets are the numbers that were found to be valid by filtering. The specific filtering method is described in Section 4.2. Con. stands for Contradiction, and Imp. stands for Implication.

	Con.	Imp.	Total
Access	3,396 (4,967)	12,528 (14,978)	15,924 (19,945)
Fee	2,135 (3,424)	5,442 (6,611)	7,577 (10,035)
Business hours	7,082 (8,709)	9,761 (11,311)	16,843 (20,020)
Total	12,613 (17,100)	27,731 (32,900)	40,344 (50,000)

3 Proposed Method

We propose a method to correct hallucinations using only SP, enabling more efficient detection and correcting of hallucinations. As shown in Figure1:A, the SP approach performs hallucination correcting with SP. Therefore, we design the prompt to generate a corrected sentence when hallucination occurs in the LLM's output, and not to modify the text when no hallucination is present. Table 1 shows the actual prompt used. Although Table 1 is translated into English, we use Japanese prompts in practice.

In #Tasks, we provide instructions to detect hallucinations and correct them when detected. We also instruct to output not only the corrected text but also a hallucination detection label. The instruction is to output 1 if a hallucination is detected and 0 if not. When using CoT, we provide the text

string shown in red in Table 1, instructing to output the reasoning and correct the hallucination based on this reasoning.

Next, #Instructions describes the items to be observed when performing the task. This is important information for the LLM to accurately execute the task according to instructions. We expect that providing this information will stabilize the LLM's output format.

In #Specific Examples, we provide Few-shot examples. We provided 3 examples of entailment relations and 4 examples of contradiction relations, and gave instructions for the outputs when hallucinations were not detected and when they were detected. As shown in red in Table 1, when using CoT, we implement it by providing examples that output reasoning. We realize CoT by having the output explain where in the target sentence there are contradictions with the knowledge, and how these contradictory parts should be corrected.

Finally, by repeatedly providing #Tasks, we ensure that the LLM follows the task instructions more faithfully. This is based on reports that when long input is given to an LLM, information in the middle is less likely to be referenced, while information at the beginning and end is more easily referenced (Liu et al., 2024).

By designing such prompts, we can give clear instructions to the LLM and make it faithfully follow these instructions. As a result, it becomes possible to effectively correct hallucinations using only a SP.

4 Dataset

To effectively correct hallucinations, it is crucial to use hallucinations actually generated by LLMs. Cao et al. (2020); Kryscinski et al. (2020) have conducted correct tests using datasets that include artificially created hallucinations.

However, it has been reported that such datasets have a different distribution from hallucinations actually generated by LLMs (Balachandran et al., 2022). Therefore, it is difficult to evaluate whether LLMs can detect and correct actual hallucinations using artificially created ones. Considering this issue, we use a dataset constructed with hallucinations actually generated by LLMs.

4.1 Generated Hallucination Data Using LLM

Moriwaki et al. (2022) fine-tuned "hobbyist," a Transformer-based LLM with 1.6 billion parameters (Sugiyama et al., 2021), and extracted hallucination data from the generated texts to construct a Japanese dataset of hallucinations.

The corpus used for training is a travel agency dialogue corpus constructed by Kaneda et al. (2022). This corpus contains dialogues between two people, a travel agent and a customer, collected using crowd workers. The agent responds to the customer's questions while referencing knowledge about tourist destinations. Therefore, this corpus includes the customer's questions, the agent's responses, and the knowledge used to create these responses.

They trained the LLM to generate response sentences by inputting questions and reference knowledge using the travel agency dialogue corpus. For reference knowledge, they used the tourist destination database in "Rurubu DATA"[4] provided by JTB Publishing Co., Ltd. This database contains information on business hours, fees, access, tourist destination names, overviews, and reviews.

Based on the finding that LLMs are prone to hallucinations regarding numerical values and proper nouns, they focused on three categories: business hours, fees, and access. They input knowledge from these categories and prepared question sentences (e.g., "What time should I go?" "How much is the fee?") to the LLM, generated response sentences, and collected hallucination data.

To collect hallucination data more efficiently, they generated responses five times for each input. Through this process, they collected 50,000 outputs from 10,000 inputs.

4.2 Manual Hallucination Judgment

Moriwaki et al. (2022) conducted manual annotations on the 50,000 data points described in Section 4.1 to determine whether they were "hallucination" or "non-hallucination." Each data point consists of a question, knowledge, and text generated by an LLM. Annotators were asked to make relation and contradiction-implication judgments by examining the knowledge and generated text. Each data point was evaluated by 5 annotators.

In the relation judgment, annotators determined whether the information contained in the generated text was included in the knowledge. In the contradiction-implication judgment, they assessed whether the generated text contradicted or was implied by the provided knowledge. If both "contra-

[4] https://solution.jtbpublishing.co.jp/service/domestic/

diction" and "implication" could be selected, annotators were instructed to choose "contradiction."

From the collected judgment results, only data where 4 or more people selected "related" in the relation judgment, and 4 or more people selected either "implication" or "contradiction" in the contradiction-implication judgment were extracted as valid data.

The number of extracted data points is shown in Table 2. The 12,613 contradiction relations and 27,731 implication relations not enclosed in brackets indicate the number of valid data points.

5 Experiment

5.1 Experimental Overview

In this study, we conduct the following three comparative experiments to verify the effectiveness of hallucination correcting using SP:

1. Calculation time
2. Correcting accuracy
3. Effects of CoT on correcting accuracy

The experiments deal with hallucinations in a Japanese knowledge-grounded dialogue generation task. In this task, hallucination refers to the inclusion of content in the LLM-generated text (target sentence) based on reference knowledge that contradicts that knowledge. Therefore, the experiments verify whether the method can detect parts of the target sentence that contradict the knowledge and appropriately correct them based on the reference knowledge.

5.2 Correcting with SP

In this experiment, we use GPT-3.5-turbo as the LLM. We use the prompt shown in Table 1. Also, to reduce output randomness and obtain more focused results, we set the temperature to 0. This setting follows Li et al. (2023).

5.3 Correcting with MP

Similar to SP, MP also uses GPT-3.5-turbo as the LLM, with the temperature set to 0. The correcting process with MP follows these steps. The actual prompts used are shown in Appendix A.

1. Generate questions for which the target sentence is the answer, based on the knowledge and target sentence (A.1)
2. Generate answers based on the generated questions and knowledge (A.2)
3. Based on the knowledge, the target sentence, and the answer generated in Step 2, perform hallucination detection on the target sentence and output the detection label(A.3)
   (a) with CoT: Input the knowledge, target sentence, and the answer generated in step 2, and output the reasoning and hallucination detection label
4. If a hallucination is detected, input the knowledge, answer, and target sentence to correct the target sentence (A.4)
   (a) with CoT: If a hallucination is detected, input the knowledge, answer, reasoning output by the detector, and target sentence to correct the target sentence

5.4 Evaluation

To evaluate each method, we prepare hallucination data and non-hallucination data. Hallucination data is obtained from 12,613 Contradiction cases shown in Table 2, from which 150 cases in each category (access, fee, business hours) are randomly extracted, for a total of 450 cases. Non-hallucination data are taken from the 27,731 Implications shown in Table 2, with a total of 450 cases randomly extracted from 150 cases in each category. Then, to evaluate the reproducibility of the output, we apply the 900-item dataset five times repeatedly for each method and obtain the results.

The outputs of each method for the evaluation data are manually annotated. The outputs are annotated according to the Correcting Type shown in Table 11.

Then, using the annotation results, we calculate Faithfulness (Parikh et al., 2020). Faithfulness represents the proportion of outputs that are non-hallucination. We use this Faithfulness to comparatively evaluate the correcting accuracy of SP and MP. The formula for calculating Faithfulness differs depending on whether the input is non-hallucination or hallucination.

$$\text{Faithfulness} = \begin{cases} \dfrac{\#\text{NN} + \#\text{NCN}}{\#\text{N}} & (\text{Input} \in \text{N}) \\ \dfrac{\#\text{HCN}}{\#\text{H}} & (\text{Input} \in \text{H}) \end{cases}$$

#NN represents the number of cases where no correcting was made for non-hallucinations. #NCN and #HCN represent the number of instances

that were corrected from hallucination to non-hallucination, and from non-hallucination, respectively.

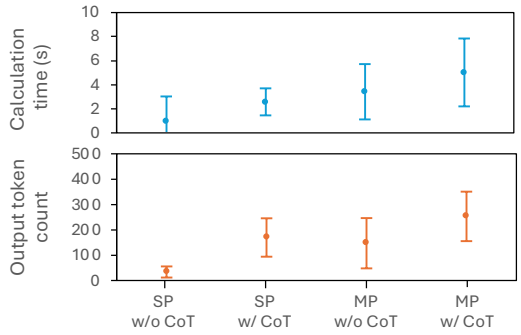

Figure 2: Compare each method in terms of calculation time and output token count. Paired t-tests revealed statistically significant differences at the 1% significance level for all comparisons.

Table 3: Comparison of mean Faithfulness between two methods with and without CoT. Faithfulness is calculated based on the output obtained by inputting non-hallucination and hallucination, using the formula defined in Section 5.4.

Input	hallucination	non-hallucination
SP w/o CoT	0.39	0.93
SP w/ CoT	0.61	0.93
MP w/o CoT	0.60	0.96
MP w/ CoT	0.49	0.96

6 Results

6.1 Calculation Time

Figure 2 shows the average calculation time and average number of output tokens for each method when a dataset of 900 instances was inputted five times. Comparing SP and MP and conducting a paired two-sided t-test revealed that SP has a shorter calculation time. SP with CoT had a shorter average calculation time than MP without CoT, and this difference was statistically significant ($t(4) = -6.14$, $p < 0.01$).

It was also revealed that correcting without CoT results in shorter calculation times. The difference with and without CoT in SP was statistically significant ($t(4) = -12.35$, $p < 0.01$). Similarly, the difference with and without CoT in MP was also statistically significant ($t(4) = -9.08$, $p < 0.01$).

When CoT is applied, the average output token count increases as it outputs the reasoning. This suggests that using CoT increases the number of output tokens, leading to longer calculation times. Furthermore, it was suggested that by integrating and unifying prompts, calculation time can be shortened, enabling efficient correcting.

6.2 Correcting Accuracy

Table 3 shows the comparison of correcting results using Faithfulness for the two methods with and without CoT. Table 3 results was the mean of five runs, with standard deviations ranging from 0.00 to 0.02. A corresponding two-sided t-test was performed for Faithfulness to confirm statistical significance.

In SP, with CoT led to a statistically significant improvement in Faithfulness for hallucinations. However, while Faithfulness in non-hallucination contexts decreased, this decrease was not statistically significant ($t(4) = 1.91$, $p = 0.13$).

In MP, with CoT resulted in a statistically significant decrease in Faithfulness in hallucinations ($t(4) = 21.58$, $p < 0.01$). Faithfulness in non-hallucinations also decreased, but this was not statistically significant ($t(4) = 1.00$, $p = 0.37$).

The results in Table 3 reveal that the effects of CoT vary significantly depending on the method used. It was found that while CoT is effective in SP, it is not effective in MP. As a result, SP outperformed MP in terms of Faithfulness in hallucinations.

7 Discussion

7.1 Correcting Accuracy Improvement of SP with CoT

The effectiveness with CoT in SP was analyzed using Table 4. This table shows the average results of annotations based on the Correcting Type defined in Section 5.4 for the corrected sentences generated by each method. Using this data, a paired two-tailed t-test was conducted to verify the correcting accuracy with CoT in SP, and to assess for any statistically significant differences.

In SP, it is believed that correcting with CoT makes it easier to perform corrections simultaneously with hallucination detection, which leads to an improvement in Faithfulness. The number of #HCH and #HH was higher for SP without CoT, with statistical significance (#HCH: $t(4) = 9.95$, $p < 0.01$, #HH: $t(4) = 33.09$, $p < 0.01$). On the other hand, the number of #HCN was higher for SP with CoT, with statistical significance ($t(4) = -50.48$, $p < 0.01$). This implies that with CoT, through

Table 4: Results of annotating the corrected sentences obtained by inputting 900 data points into each method 5 times and calculating the mean for each category, following Table 11. (Unit: count) In the output row, "N" indicates that no hallucination was included in the corrected sentence, and "H" indicates that a hallucination was included.

Category	HCN	HCH	HH	NCN	NCH	NN
Input	Hallucination (H)			Non-hallucination (N)		
Corrected	Yes	Yes	No	Yes	Yes	No
Output	N	H	H	N	H	N
SP w/o CoT	174.4	146.6	129.0	96.6	30.6	322.8
SP w/ CoT	274.8	125.2	50.0	114.0	33.4	302.6
MP w/o CoT	270.2	106.8	73.0	162.4	16.6	271.0
MP w/ CoT	220.2	108.2	121.4	54.2	18.6	377.4

Table 5: Among the data annotated with #HH, this shows the rate at which the target sentence was outputted verbatim. This represents the rate at which the model detected a hallucination but was unable to correct it.

	Verbatim Output Rate
SP w/o CoT	0.80
SP w/ CoT	0.15
MP w/o CoT	0.50
MP w/ CoT	0.90

Table 6: Accuracy comparison of different hallucination detection methods using various evaluation metrics (acc. = accuracy, rec. = recall, prec. = precision). Results are shown for Detection only, SP, and MP Methods, both with and without CoT.

		acc.	rec.	prec.	f-1
Detect only	w/o CoT	0.76	0.81	0.74	0.77
	w/ CoT	0.79	0.68	0.88	0.76
SP	w/o CoT	0.74	0.94	0.69	0.78
	w/ CoT	0.78	0.91	0.72	0.80
MP	w/o CoT	0.69	0.86	0.65	0.74
	w/ CoT	0.79	0.76	0.81	0.78

the generation of intermediate steps, suggests an improvement in the correcting accuracy of hallucinations.

This is also suggested by the proportion of #HH in SP without CoT. #HCH refers to data where correcting was performed on hallucination data but hallucination occurred in the corrected text, and #HH refers to data where "none" was output due to failure to detect hallucination, or where the input target sentence was output as is. Therefore, #HH contains a mix of data where hallucination was not detected and data where correcting could not be performed.

In #HH, the proportion of the target sentences that were output unchanged without being corrected is shown in Table 5. In SP without CoT, it accounts for 80.0%, and this difference is statistically significant when compared with SP with CoT. This result indicates that in SP without CoT, although hallucinations can be detected, it is difficult to correct them simultaneously, leading to a tendency to output the target sentences as they are.

7.2 Correcting Accuracy Decrease of MP with CoT

In MP, the introduction of CoT is thought to have lowered the detection metrics for hallucinations, which in turn made corrections difficult, leading to a decrease in Faithfulness.

The observed trend for each category showed that the counts of #HCN and #NCN were higher in MP without CoT, with a statistically significant difference (#HCN: $t(4) = 21.72$, $p < 0.01$; #NCN: $t(4) = 28.38$, $p < 0.01$). On the other hand, MP with CoT had a higher number of #HH and #NN cases, with statistical significance (#HH: $t(4) = -12.00$, $p < 0.01$, #NN: $t(4) = -22.45$, $p < 0.01$). The higher counts of #HH and #NN suggest that it becomes easier to detect non-hallucination instances.

This implies that detection with CoT in MP shifts the discrimination boundary towards hallucination, making it more likely to mistakenly detect non-hallucination instances as hallucinations. Consequently, it is suggested that with CoT, corrections become less feasible, resulting in fewer instances of #HCN and #NCN compared to without CoT.

The hallucination correcting accuracy can be evaluated by comparing the numbers of #HCH and #NCH cases. As a result, no statistical significance was found in the difference in the numbers of #HCH and #NCH cases with and without CoT ($t(4) = -1.00$, $p = 0.37$). Therefore, it is likely that CoT does not have a significant impact on the hallucination correcting accuracy itself.

From the above analysis, it was suggested that in MP, while CoT does not significantly change the hallucination correcting accuracy, it does make hallucinations more difficult to detect in terms of detection metrics. MP separates the detection and correcting processes, and if hallucination is not detected, it does not transition to the correcting process, making correcting impossible. Therefore, it is thought that with CoT, it became easier to misidentify hallucinations as non-hallucinations, resulting in an inability to correct hallucinations and a decrease in Faithfulness.

7.3 Effect of CoT on Hallucination Detection

To analyze the effectiveness of CoT in hallucination detection, we prepared a detection-only prompt. We analyzed the effectiveness of CoT using the detection labels output when using the detection-only prompt, SP, and the detection prompt in MP.

From the hallucination detection metrics results of each method shown in Table 6, it was confirmed that with CoT in all methods, accuracy and precision improve while recall decreases. This result suggests that while CoT contributes to improving the accuracy of hallucination detection, it tends to decrease recall.

While recall decreases in all methods, it became clear that SP is the method that can most effectively suppress the decrease in recall with CoT among the three methods. This result suggests that by unifying detection and correcting, more careful detection becomes necessary as it needs to consider the correcting process, enabling a more attentive detection and thus suppressing the decrease in recall that occurs with CoT.

In MP, where the detection and correcting processes are separated, it is important to increase recall and prevent hallucinations from being overlooked. However, as mentioned earlier, with CoT for hallucination detection tends to decrease recall, and in MP, this might lead to overlooking hallucinations. Consequently, this inability to correct hallucinations may result in a decrease in Faithfulness, as suggested.

In contrast, SP can minimize the decrease in recall that occurs with CoT. The unification of the detection and correcting processes helped suppress the decrease in recall, which in turn reduced the oversight of hallucinations, leading to an improvement in Faithfulness with CoT. This suggests that SP is an approach that mitigates the problem of decreased recall associated with CoT, thereby enhancing detection metrics and Faithfulness.

8 Conclusion

In this study, we proposed a hallucination correcting method that requires less calculation time and is more accurate than the method using SP, and verified its effectiveness. The core of this method lies in having the LLM simultaneously detect and correct hallucinations with only SP, thereby efficiently achieving hallucination correcting.

To verify the effectiveness of the proposed method, we conducted comparative experiments with MP, focusing on calculation time and Faithfulness. The experimental results yielded the following findings:

1. It was confirmed that SP can significantly reduce calculation time while achieving Faithfulness equal to or better than MP.

2. With CoT, SP's Faithfulness was proven to further improve.

3. SP was found to excel in its ability to minimally suppress the decrease in recall observed with CoT in hallucination detection.

On the other hand, it was suggested that in MP, with CoT, the discrimination boundary of hallucination detection shifted more towards hallucination, making it easier to misidentify non-hallucinations, resulting in a decrease in Faithfulness.

The results of this study reduced the time required for correcting hallucinations and improved correcting accuracy, providing important insights into tasks that require real-time processing. Furthermore, by presenting a new perspective on the effective use of CoT, this study may contribute to the improvement of hallucination detection using LLMs and correcting tasks in general.

Limitation

This research has the following limitations:

- The study focuses on correcting hallucinations in Japanese. Addressing hallucinations in other languages remains a subject for future research.
- The necessary knowledge was already included in the dataset used, eliminating the need to search for or retrieve information from external sources. However, in practical applications, there may be cases where knowledge needs to be acquired externally, and the associated processing costs have not been considered in this study.
- The research addressed the correcting of relatively short hallucinations consisting of 1-3 sentences, confirming that using a single prompt improved accuracy. However, the correcting of more complex hallucinations in longer texts or list formats remains a topic for future research.
- The study focuses solely on hallucinations related to numerical values and proper nouns. Future research should explore the applicability of the proposed method to other types of hallucinations.
- This study limited its verification to hallucinations in dialogue data. Moving forward, it is important to verify the versatility of the proposed method by applying it to hallucinations across various tasks.

Acknowledgements

This research originated from valuable discussions with Mr. Tomoki Morikawa and Mr. Kazuma Enomoto. I am deeply grateful for their significant contributions to this work. I would also like to express my sincere gratitude to Mr. Yoshiki Tomita, Mr. Nozomi Kimata, and Mr. Jundai Suzuki for their invaluable assistance with the annotation work. Their careful and precise work substantially enhanced the quality of this study.

References

Josh Achiam et al. 2023. GPT-4 technical report. *Computation and Language* arXiv:2303.08774. Version 6.

Anas Awadalla et al. 2022. Exploring the landscape of distributional robustness for question answering models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5971–5987, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Vidhisha Balachandran et al. 2022. Correcting diverse factual errors in abstractive summarization via post-editing and language model infilling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9818–9830, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tom Brown et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Meng Cao et al. 2020. Factual error correction for abstractive summarization models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6251–6258, Online. Association for Computational Linguistics.

Jacob Devlin et al. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Shehzaad Dhuliawala et al. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 3563–3578, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Nouha Dziri et al. 2022. On the origin of hallucinations in conversational models: Is it the datasets or the models? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5271–5285, Seattle, United States. Association for Computational Linguistics.

Ryouhei Kaneda et al. 2022. Utterance generation using a dialogue corpus with one-to-many relationships with knowledge sources. In *Proceedings of the 28th Annual Conference of the Natural Language Processing*, pages 191–196. (in Japanese).

Wojciech Kryscinski et al. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages

9332–9346, Online. Association for Computational Linguistics.

Hwanhee Lee et al. 2022. Factual error correction for abstractive summaries using entity retrieval. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics*, pages 439–444, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Mike Lewis et al. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Patrick Lewis et al. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems 34*, page 9459–9474.

Junyi Li et al. 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.

Nelson F. Liu et al. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Shayne Longpre et al. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Keita Moriwaki et al. 2022. Detection and fix of factual inconsistency contained in neural generated sentences. In *Proceedings of the 36th Annual Conference of the Japanese Society for Artificial Intelligence*, 2L1-GS-2-04. (in Japanese).

Niels Mündler et al. 2024. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. In *The Twelfth International Conference on Learning Representations*.

Ankur Parikh et al. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1186, Online. Association for Computational Linguistics.

Colin Raffel et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Kurt Shuster et al. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chenglei Si et al. 2023. Prompting GPT-3 to be reliable. In *The Eleventh International Conference on Learning Representations*.

Hiroaki Sugiyama et al. 2021. Empirical analysis of training strategies of Transformer-based Japanese chit-chat systems. *Computation and Language* arXiv:2109.05217. Version 1.

James Thorne et al. 2021. Evidence-based factual error correction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3298–3309, Online. Association for Computational Linguistics.

Jason Wei et al. 2022. Chain-of-Thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35*, pages 24824–24837.

Jian Xie et al. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Ruochen Zhao et al. 2023. Verify-and-Edit: A knowledge-enhanced chain-of-thought framework. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5823–5840, Toronto, Canada. Association for Computational Linguistics.

A Prompts Used in MP

A.1 Question Generation Phase

Table 7 shows the prompt used in the question generation phase for MP. In this phase, possible questions are generated from the target sentence. For example, for the target sentence "The entrance fee for university students is 630 yen," it generates the question "How much is the entrance fee for university students?." The questions obtained in this phase are used to generate answers in the subsequent answer generation phase.

A.2 Answer Generation Phase

Table 8 shows the prompt used in the answer generation phase for MP. In this phase, answers are generated by referencing knowledge in response to the questions obtained from the previous question generation phase. The answers obtained in this phase are used to detect hallucinations by comparing them with the target sentence in the subsequent detection phase.

A.3 Hallucination Detection Phase

Table 9 shows the prompt used in the hallucination detection phase for MP. In this phase, the answers obtained from the answer phase are compared with the target sentence to detect if there are any contradictions in the target sentence. As shown in red in Table 9, CoT is used. The output determines whether the target sentence and the answer have an implication or contradiction relationship. If there's a contradiction relationship, it outputs the reason why and where the contradiction exists, thus performing hallucination detection. If a hallucination is detected in this phase, it transitions to the subsequent correcting phase. With CoT, both the reasoning and the judgment label are used in the correcting phase. Without CoT, only the judgment label is used in the correcting phase.

A.4 Hallucination Correcting Phase

Table 10 shows the prompt used in the answer generation phase for MP. In this phase, hallucination correcting is performed using the output obtained from the previous phase. With CoT, the target sentence is corrected based on the knowledge, answer sentences, and reasoning. Without CoT, the target sentence is corrected based on the knowledge and answer sentences.

B About Categories

We explain the categories using Table 11. For non-hallucination data, there are three categories: #NN, #NCN, and #NCH. #NN is the category for data where "None" was output, or the input target sentence was output as is. #NCN is the category for data where correcting was performed, and the corrected sentence is non-hallucination. #NCH is the category for data where correcting was performed, and the corrected sentence is hallucination.

For hallucination data, there are three categories: #HH, #HCN, and #HCH. #HH is the category for data where "None" was output, or the input target sentence was output as is. #HCN is the category for data where correcting was performed, and the corrected sentence is non-hallucination. #HCH is the category for data where correcting was performed, and the corrected sentence is hallucination.

Table 7: Prompt for the answer generation phase in MP. The text shown here is the English translation of the Japanese original.

#Tasks
• You should generate questions from the target sentence.
• Generate questions for which the given target sentence would be the answer.
• Absolutely follow the content of the instructions.

#Instructions
• Strictly adhere to the output format.
• You may break down the target sentence and output multiple questions.
• Please refer to the following specific examples.

#Specific Examples
##Example 1
##Input
Target sentence: It's a 5-minute walk from Susukino Station or a 25-minute walk from Sapporo-kita IC (Kita-ku) on the Sasson Expressway.
##Output
Question 1: How long does it take from Susukino Station?
Question 2: How long does it take from Sapporo-kita IC (Kita-ku) on the Sasson Expressway?
• • •

#input/output
##Input
Target sentence: {target sentence}
##Output

To reiterate, you should complete the following tasks: #Tasks You should generate questions from the target sentence. Generate questions for which the given target sentence would be the answer. - Absolutely follow the content of the instructions.

Table 8: Prompt for the answer generation phase in MP. The text shown here is the English translation of the Japanese original.

#Tasks
• You should answer the given questions based on the provided knowledge.

#Instructions
• There may be more than one question given; there could be multiple questions.
• If there are multiple questions, answer all of them.
• Strictly adhere to the output format.
• Do not output the content of the input.
• Only output the answers.
• Please refer to the following specific examples.

#Specific Examples
##Example 1
##Input
Question 1: How much is the fee?
Knowledge: Adults (15 years and older) 1900 yen, Children (Elementary and Junior High School students) 950 yen, Infants (3-5 years old) 300 yen, Seniors (65 years and older) 1100 yen
##Output
Answer 1: For 15 years and older, it's 1900 yen; for elementary and junior high school students, 950 yen; for infants, 300 yen; and for seniors, 1100 yen.
• • •

#input/output
##Input
Question: {question}
Knowledge: {knowledge}
##Output

To reiterate, you should complete the following tasks: #Tasks To repeat, please absolutely follow these rules: There may be more than one question given; there could be multiple questions. If there are multiple questions, answer all of them. Strictly adhere to the output format. Do not output the content of the input. Only output the answers.

Table 9: Prompt for the hallucination detection phase in MP. The text shown here is the English translation of the Japanese original. The red characters indicate the strings of characters used when using CoT.

#Tasks

• You can infer whether the target sentence has a contradiction or implication relationship with the knowledge and answer sentences.
• Output the reasoning for determining if there are contradictions between the knowledge and target sentence, and based on this reasoning, output a judgment label.
• Outputting a 0 label means the knowledge and target sentence have an implication relationship.
• Outputting a 1 label means the knowledge and target sentence have a contradiction relationship.
• If there's a contradiction relationship between the knowledge, answer sentences, and target sentence, output the reasoning for where the target sentence contradicts or contains extra information, and based on this reasoning, output a judgment label.
• Perform the detection based on this reasoning.

#Instructions

• Always follow the rules.
• Strictly adhere to the output format.
• Make judgments based on the reasoning.
• Detect any contradictions between the answer sentences and target sentence based on the knowledge.
• Output 0 if the knowledge, answer sentences, and target sentence have an implication relationship.
• Output 1 if there are contradictions between the knowledge, answer sentences, and target sentence
• Carefully examine the knowledge, answer sentences, and target sentence to determine if there's a contradiction or implication relationship and output the label.
• Please refer to the following specific examples.

#Specific Examples

##Specific Example 1 (Implication Relationship)
##Input
Knowledge: Business hours: 10 AM to 10 PM (until 9 PM from January to March), admission until 20 minutes before closing, No regular holidays
Answer 1: The business hours are from 10 AM to 10 PM. Admission is until 20 minutes before closing. We are open every day.
Target sentence: Admission is until 20 minutes before closing.
##Output
Reasoning: The target sentence "Admission is until 20 minutes before closing." does not contradict Answer 1 "The business hours are from 10 AM to 10 PM. Admission is until 20 minutes before closing. We are open every day."
Judgment: 0
• • •

#Specific Example 7(Contradiction Relationship)
##Input
Knowledge: (1) 5-minute walk from JR Onuma-Koen Station (2) 15 minutes from Doo Expressway Onuma-Koen IC
Answer 1: It's a 5-minute walk from JR Onuma-Koen Station.
Answer 2: It's 15 minutes from Doo Expressway Onuma-Koen IC.
Target sentence: It's a 5-minute walk from JR Onuma-Koen Station or a 15-minute walk from Doo Expressway Onuma-Koen IC.
##Output
Reasoning: The part of the target sentence "15-minute walk from Doo Expressway Onuma-Koen IC" contradicts Answer 2 "15 minutes from Doo Expressway Onuma-Koen IC". The information "15-minute walk" is inconsistent. Therefore, it needs to be corrected to "15 minutes from Doo Expressway Onuma-Koen IC."
Judgment: 1

#input/output

##Input
Knowledge: {knowledge} Answer sentence: {answer sentence} Target sentence: {target sentence}
##Output

To reiterate, you should complete the following tasks: #Tasks You can infer contradictions and implication relationships between knowledge and target sentences. Detect any contradictions between the answer sentences and target sentence based on the knowledge. If there are no contradictions between the answer sentences and target sentence, output 0. If there are contradictions between the answer sentences and target sentence, output 1.

Table 10: Prompt for the hallucination correcting phase in MP. The text shown here is the English translation of the Japanese original. The red characters indicate the strings of characters used when using CoT.

#Tasks
• You can correct contradictions in the target sentence based on the knowledge and answer sentences derived from that knowledge.
• Use the detector's output to correct the contradiction parts.
• Make corrects based on the reasoning provided.

#Instructions
• Always follow the rules.
• Strictly adhere to the output format.
• Only output the corrected sentence.
• Correct any contradictions in the target sentence by referring to the knowledge and answer sentences.
• Please refer to the following specific examples.

#Specific Examples
##Specific Example 1 (Contradiction Relationship)
##Input
Knowledge: Operating hours: 10:00 AM to 7:00 AM the next day, Closed: Never
Answer 1: The operating hours are from 10:00 AM to 7:00 AM the next day. It's open every day.
Target sentence: It's from 10 AM to 7 PM.
##Detector Output
Reasoning: The target sentence "It's from 10 AM to 7 PM" states 7 PM, but Answer 1 "The operating hours are from 10:00 AM to 7:00 AM the next day. It's open every day." indicates 7:00 AM the next day. Therefore, the target sentence contradicts the knowledge. As a result, the target sentence needs to be corrected to "It's from 10 AM to 7 AM the next day."
Judgment: 1
##Output
Correcting: It's from 10 AM to 7 AM the next day.
• • •

###Specific Example 4 (Contradiction Relationship)
##Input
Knowledge: (1) 5-minute walk from JR Onuma-Koen Station (2) 15 minutes from Doo Expressway Onuma-Koen IC
Answer 1: It's a 5-minute walk from JR Onuma-Koen Station.
Answer 2: It's 15 minutes from Doo Expressway Onuma-Koen IC.
Target sentence: It's a 5-minute walk from JR Onuma-Koen Station or a 15-minute walk from Doo Expressway Onuma-Koen IC.
##Detector Output
Reasoning: The part of the target sentence "15-minute walk from Doo Expressway Onuma-Koen IC" contradicts Answer 2 "15 minutes from Doo Expressway Onuma-Koen IC". The information "15-minute walk" is inconsistent. Therefore, it needs to be corrected to "15 minutes from Doo Expressway Onuma-Koen IC."
Judgment: 1
##Output
Correcting: It's a 5-minute walk from JR Onuma-Koen Station or 15 minutes from Doo Expressway Onuma-Koen IC.

#input/output
##Input
Knowledge: {knowledge} Answer sentence: {answer sentence} Target sentence: {target sentence} Detector Output: {detector output}
##Output

To reiterate, you should complete the following tasks: #Tasks Only output the corrected sentence. Strictly adhere to the output format. Correct any contradictions in the target sentence by referring to the knowledge and answer sentences.

Table 11: An example of a corrected sentence output by the corrector. The categories represent the types of corrects made to the corrected sentence. We manually annotated the sentences based on the types of corrects.

Knowledge	Target Sentence	Corrected Sentence	Correcting Type	Category
Temple grounds free (Main hall entrance fee is 500 yen)	The temple grounds are free, and the main hall entrance fee is 700 yen.	The temple grounds are free, and the main hall entrance fee is 500 yen.	Corrected hallucination data, no hallucination in the corrected sentence.	HCN
		There is a 500 yen entrance fee for both the temple grounds and the main hall.	Corrected hallucination data, created another hallucination.	HCH
		None or The temple grounds are free, and the main hall entrance fee is 700 yen.	Hallucination data not corrected.	HH
Temple grounds free (Main hall entrance fee is 500 yen)	The temple grounds are free, and the main hall entrance fee is 500 yen.	The temple grounds are free, but there is a 500 yen fee for entering the main hall.	Corrected non-hallucination data, no hallucination in the corrected sentence.	NCN
		The temple grounds are free, but there is a 1500 yen fee for entering the main hall.	Corrected non-hallucination data,, created hallucination.	NCH
		None or The temple grounds are free, and the main hall entrance fee is 500 yen.	Non-hallucination data not corrected.	NN