ABSTRACT
In this paper, we describe our work on Answer Verification. We submitted one result to Answer Verification. The method used for the submitted data is to input the "AnswerSummary," "AnswerOriginal," and "QuestionSummary" items together into ChatGPT in order to classify them. As a result, an Accuracy of 0.5800 for the Answer Verification was obtained.

KEYWORDS
Answer Verification, Fact Check, ChatGPT

TEAM NAME
Forst

SUBTASKS
Answer Verification (Japanese)

1 INTRODUCTION
We worked on the Answer Verification (AV) subtask in the NTCIR-17 QA Lab PoliInfo-4 task[1]. The purpose of the Answer Verification subtask is to perform fact checking on the short answers automatically generated by the Question and Answer (QA) task. The first stage is a training data extension to generate false answers, including both those that appear to be true but are false and those that appear to be false but are true. In the second stage, participants build a classifier that can correctly judge the test data set. We worked on the second stage and experimented with ChatGPT to see if it would work effectively for text classification in this task.

In this paper, we describe the methods we worked on for our Answer Verification subtask. Section 2 describes the related studies we use in our method. Section 3 describes our proposed method and its prompts to be entered into ChatGPT. Section 4 presents our experimental results. Section 5 discusses our method and this task based on the results.

2 RELATED WORK
ChatGPT 1, released by OpenAI in 2022, has attracted a great deal of attention in NLP for its usefulness not only in the community but also in society at large. ChatGPT and other recent large-scale language models have acquired the versatility to perform a variety of tasks according to instructions given in natural language. Among them, ChatGPT has been shown to perform well on many tasks that emphasize inference ability[2]. In particular, ChatGPT has shown a high ability to classify text that matches the facts in the Recognizing Textual Entailment (RTE) task. ChatGPT also has the ability to interactively provide instructions and questions in addition to traditional language models. This has the advantage of making it easily accessible to a segment of the population that is not familiar with large-scale language models. By analyzing the results of using ChatGPT for the AV subtask, we hope to gain insight into how to create an environment that will facilitate fact-finding for a large audience.

Figure 1: Few-shot prompt with all 3 items

3 METHODS
We tried three methods using three main items: "QuestionSummary" (QS), which is a summary of the question from the Q and A session that appeared in the minutes from each of the data used for the AV subtask; "AnswerSummary" (AS), which is a few lines of correct and fake answers to this question; and "AnswerOriginal" (AO), which is the text segment in the minutes that corresponds to this question. GPT4 is used in our all methods.
3.1 Method 1: Few-shot prompting with all three items (QS+AO+AS)

This method inputs the three aforementioned items together into ChatGPT. The condition indicates what is to be output to ChatGPT. If the content of AS is an answer regarding what was asked in QS, and if it correctly summarizes the content of AO, ChatGPT outputs True. Otherwise, it outputs False. First, the prompt to ChatGPT consists of the instructions, conditions, and some example answers selected from the training dataset, as shown in input1 of Figure 1. One correct answer and four fake answers were selected from the training dataset, taking into account the limit of input length for ChatGPT. The four examples of answers chosen as fake answers were manually selected from the training dataset, each with a different reason for being fake. After inputting these sentences, ChatGPT outputs the sentences shown in output1 of Figure 1. The second and subsequent test data are entered as shown in input 2 of Figure 1 while taking into account the input length for ChatGPT. An example of the output obtained from this is shown in output 2 of Figure 1.

![Figure 2: Prompt with "AnswerOriginal" and "AnswerSummary"](image)

3.2 Method 2: Combination method with two zero-shot subtasks ([AO+AS], [QS+AS])

In Section 3.1, three items were processed together, but here two items are used to formalize them as the combination of two zero-shot subtasks.

![Figure 3: Prompt with "QuestionSummary" and "AnswerSummary"](image)

3.2.1 Subtask 1: Examine entailment relations between "AnswerOriginal" and "AnswerSummary"([AO+AS]). First, we input the instructions and conditions shown in input1 of Figure 2. In input 1, AO is "T" and AS is "H", instructing ChatGPT to answer True or False for the entailment relation of the two texts. After inputting these sentences, ChatGPT outputs the sentences shown in output1 of Figure 2. The second and subsequent test data are entered as shown in input 2 of Figure 2 while taking into account the input length for ChatGPT. An example of the output obtained from this is shown in output 2 of Figure 2.

3.2.2 Sub-task 2: Examine whether the content of the answer "AnswerSummary" is appropriate to the question "QuestionSummary". It is expected that Subtask 1 will tell us whether the content of AS is included in AO or not. However, it is not possible to determine whether the content of AS is appropriate for QS, even if there is an entailment relationship between the AO and AS. The purpose of subtask 2 is to check whether AS is appropriate for the content of the question.

First, we input the instructions and conditions shown in input 1 of Figure 3. In input 1, QS is "Q", AS is "A", and ChatGPT is instructed to answer whether the answer is appropriate for the question with True or False. After inputting these sentences, ChatGPT outputs the sentences shown in output1 of Figure 3. The second and subsequent test data are entered as shown in input 2 of Figure 3 while taking into account the input length for ChatGPT. An example of the output obtained from this is shown in output 2 of Figure 3.

The results obtained from the two subtasks were combined to obtain the final judgment results. Basically, the results from subtask 1 were used. For data that were judged True in subtask 1 and False in subtask 2, the results of subtask 2 were used.
3.3 Method 3: Add items for additional information to each of the subsystems in

3.3.1 Add "Date" to Subtask 1 in Section 3.2. AS may contain date expressions. We would like to improve the classification accuracy for these cases. For example, AO may contain vague date expressions such as "next month", while AS has specific date expression like "March, [Reiwa] 2). For this reason, we added 'Date', which is date and time information, to subtask 1 in section 3.2.

Additional instructions, indicated by the orange squares in Figure 4, were added to the instructions and conditions used in subtask 1 of Section 3.2.

3.3.2 Add "SubTopic" to Subtask 2 in Section 3.3. The QS had sentences that briefly asked questions such as "How should we tackle this?". These sentences are not specific enough, and ChatGPT might judge them as fake answers due to the lack of relevance to the answers. Therefore, a "SubTopic" was added to Subtask 2 of Section 3.2 to indicate the subject of the target Q and A session. Additional instructions, indicated by the orange squares in Figure 5, were added to the instructions and conditions used in subtask 2 of Section 3.2.

As in section 3.2, the results obtained from the two subtasks were combined to obtain the final judgment results.

4 EXPERIMENTS

Table 1 shows the experimental results for methods described in Section 3. The accuracy for identifying correct answers is also shown in the table, as well as the accuracy for identifying fake answers. The method described in Section 3.1 was used for submission ID 223.

<table>
<thead>
<tr>
<th>ID:Method</th>
<th>Accuracy</th>
<th>Correct</th>
<th>Fake</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z23:[Q5+AO]+AS</td>
<td>58.00%</td>
<td>97.30%(37/37)</td>
<td>34.92%(22/63)</td>
</tr>
<tr>
<td>[AO]+[AS],[Q5+AS]</td>
<td>77.00%</td>
<td>89.19%(33/37)</td>
<td>69.84%(44/63)</td>
</tr>
<tr>
<td>[AO]+[AS]+Date,[AS]+ST</td>
<td>78.00%</td>
<td>100.00%(37/37)</td>
<td>65.08%(41/63)</td>
</tr>
</tbody>
</table>

5 DISCUSSIONS

5.1 Method 1(QS+AO+AS)

Table 1 shows that fake answers are misclassified far more often than correct answers. The cases in which the correct classification of fake answers was made are those in which the content of AS is unrelated to the other two items ((1) in Figure 6) or different (2) in Figure 6).

It is consider that these can correctly classify fake answers because phrases appearing in AS often differ from those in AO.
content of AS is not fake. In subtask 2 of Method 2, we would like to explore whether we correctly classified AS summarizing these incorrect parts.

Figure 11: Example of AS date information differing from that of AO.

Furthermore, there were cases where ChatGPT made incorrect judgments due to differences in date information. In the example shown in Figure 11, ChatGPT cannot determine whether "来年度" in AO corresponds to the correct date information "30年度" in AS. However, it appears that the rationale behind ChatGPT’s output does not take into consideration the validity of the date information within AS.

It is also possible from these misclassified fake answers that the first example of fake answers inputted into ChatGPT was not appropriate, and that the prompt was not in the correct format. This is because some of the misclassified fake answers had the same reason that AS was fake answer as the first example of fake answer inputted into ChatGPT.

5.2 Method 2 ([AO+AS],[QS+AS])

Table 2 shows that Method 2 was able to correctly classify more erroneous fake answers than Method 1. This can be thanks to the performance of ChatGPT for the recognition of entailment relations. Table 2 shows the classification results of the test data using ChatGPT.
Table 2: The results of only sub-task 1(AO+AS)

<table>
<thead>
<tr>
<th>ID:Method</th>
<th>Accuracy</th>
<th>Correct</th>
<th>Fake</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub-task1</td>
<td>77.00%</td>
<td>100.00%(37/37)</td>
<td>63.49%(40/63)</td>
</tr>
</tbody>
</table>

only SubTask 1. Table 2 shows that ChatGPT’s ability to recognizing textual entailment is high not only in classifying fake answers, but also in classifying correct answers.

However, it also misjudged implicational relationships in data with different date expressions, as mentioned in Section 5.1, such as the data represented in Figure 11. In particular, differences in date information are crucial elements in the context of fake news. Therefore, it is necessary for the first instructions input to ChatGPT in Subtask 1 to add conditions such as non-entailment relation when AS and AO date expressions are not in correspondence. Additionally, in Subtask 1 of Method 3, we will consider whether these data can be correctly classified when date information is added as input.

Figure 13: Example of QS content being abstract; ChatGPT output shows no relevance to AS content.

Table 2 also shows that it was the influence of subtask 2 that misclassified the correct answers. An example of a misclassified correct answer is when the QS was an abstract sentence such as "What is your approach?", as shown in Figure 13. In this case, ChatGPT does not find a concrete relationship between AS and QS and incorrectly classifies it as a fake answer.

Figure 14: Examples of misclassifying correct answers in Subtask 2

In some cases where the correct answer was misclassified, the omission of QS content was assumed to be the cause of the misclassification. As shown in Figure 14, “僕の解釈です。”(Express your views.) in QS omits the target content. This may have caused ChatGPT to misinterpret QA question as asking for the respondent’s views on politicians. This omitted content is one of the data items, SubTopic, which is "江戸城再建" (Reconstruction of Edo Castle). In Method 3, we would like to explore whether adding SubTopic can help alleviate the abstraction of QS that lead to such misclassifications.

Figure 15: The results of evaluating the data from Figure 10 for Subtask 2

However, the number of correctly classified fake answers increases when the results of Subtask 2 are taken into account. The fake answers correctly classified by Subtask 2 are those for which the relationship between QS and AS is not appropriate, such as Figure 10 in Section 5.1. Figure 15 shows the output results of ChatGPT when classifying the data shown in Figure 10 for Subtask 2. Therefore, it is necessary to consider whether the results of Subtask 2 should be taken into account, depending on whether the correct or fake answers are more important.

It is also possible that the initial instructions provided to ChatGPT for Subtask 2 may not have been appropriate. This is because many of the misclassified false answers in Method 2 are data for which the implication relationship between AO and AS is established, but the question-answer relationship is not suitable, and such data are misclassified in Subtask 2. ChatGPT outputs the basis for each answer based on heuristics related to the initial instructions. It is necessary to consider incorporating finer conditional specifications in the instructions to establish the correct relationship between the question and the answer.

Similarly to Method 1, Method 2 was not able to correctly classify fake answers with unnaturally broken sentences in the AS or broken proper nouns.

5.3 Method 3([AO+AS+Date],[QS+AS+ST])

From Table 1, it can be seen that Method 3 reduces the misclassification of correct answers compared to Method 2.

Figure 16: Examples where adding "Subtopic" allowed for correct classification

As shown in Figure 16, abstract sentence for QS as shown in Section 5.2 was completed by “Subtopic”. With this completion,
we expected ChatGPT to improve its correct answer classification performance. However, as with Method 2, there were many data that were misjudged as having a relationship with data for which the relationship between question and answer was not suitable. Therefore, as in section 5.2, it is necessary to reconsider the conditions of the directive.

Figure 17: Example of implication recognition results with the addition of "Date" information

The recognition of implication relations with the addition of "Date" information was not directly improved the results without "Date" information. Figure 17 is the output of ChatGPT when the same data as shown in Figure 11 in Section 5.2 is used for implication recognition with the "Date" information added. The fact that the output results in Figure 17 do not mention date information suggests that ChatGPT does not consider date information important for implication recognition. Therefore, it is necessary to verify the date information of AO and AS separately for the comparison of date information.

Finally, the addition of "Date" and "SubTopic" in Method 3 did not have a significant impact on the classification of misclassified answers in which the sentences in the AS were unnaturally broken or proper nouns were broken. It is a common problem to all methods that ChatGPT misclassifies fake answers with AS containing many of the same phrases as AO.

6 CONCLUSIONS

In this paper we described our effort to the Answer Verification task. The result was an Accuracy of 0.5800 in the submitted data. The method used for the submitted data was to input the "AnswerSummary," "AnswerOriginal," and "QuestionSummary" items together and have ChatGPT classify them.

However, there were some data in the results that ChatGPT did not consider all three items. Therefore, the classification was performed as a combination of two zero-shot subtasks using two items. As a result, many more fake answers were correctly classified compared to the method that ChatGPT considered all three items together.

REFERENCES
