NTCIR 17 Conference: Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, December 12-15, 2023, Tokyo, Japan
DOI: https://doi.org/10.20736/0002001298

Forst: A Challenge to the NTCIR-17 QA Lab-PoliInfo-4 Task

Daiki Iwayama
Yokohama National University
Japan
iwayama-daiki-kc@ynu.jp

Hideyuki Shibuki
Besna Institute Inc.
Japan
shib@besna.institute

Tatsunori Mori
Yokohama National University
Japan
tmori@ynu.ac.jp

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

Abswer Verification, Fact Check, ChatGPT

TEAM NAME

Forst

SUBTASKS

Answer Verification(Japanese)

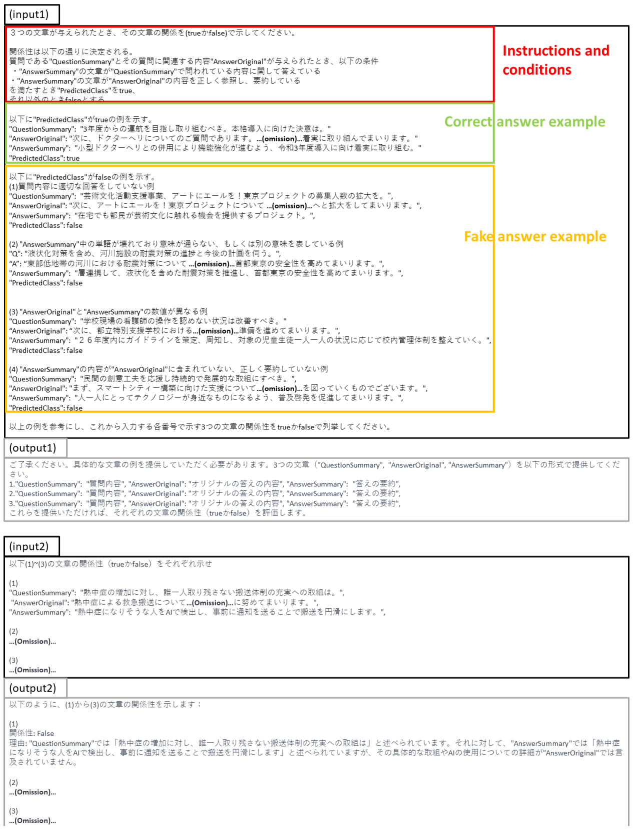
1 INTRODUCTION

We worked on the Answer Verification (AV) subtask in the NTCIR-17 QA Lab PoliInfo-4task[1]. The purpose of the AnswerVerification 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.



(input1)

３つの文章が与えられたとき、その文章の関係を(trueかfalse)で示してください。

関係性は以下の通りに決定される。
質問である"QuestionSummary"とその質問に関連する内容"AnswerOriginal"が与えられたとき、以下の条件
・"AnswerSummary"の文章が"QuestionSummary"で問われている内容に関して答えている
・"AnswerSummary"の文章が"AnswerOriginal"の内容を正しく参照し、要約している
を満たすとき"PredictedClass"をtrue、
それ以外のときfalseとする。

以下に"PredictedClass"がtrueの例を示す。
"QuestionSummary": "3年度からの運航を目指し取り組むべき。本格導入に向けた決意は。"
"AnswerOriginal": "次に、ドクターヘリについてのご質問であります。**...(omission)...**着実に取り組んでまいります。"
"AnswerSummary": "小型ドクターヘリとの併用により機能強化が進むよう、令和3年度導入に向け着実に取り組む。"
"PredictedClass": true

以下に"PredictedClass"がfalseの例を示す。
(1)質問内容に適切な回答をしていない例
"QuestionSummary": "芸術文化活動支援事業、アートにエールを！東京プロジェクトの募集人数の拡大を。",
"AnswerOriginal": "次に、アートにエールを！東京プロジェクトについて **...(omission)...**へと拡大をしてまいります。",
"AnswerSummary": "在宅でも都民が芸術文化に触れる機会を提供するプロジェクト。",
"PredictedClass": false

(2) "AnswerSummary"中の単語が壊れており意味が通らない、もしくは別の意味を表している例
"Q": "液状化対策を含め、河川施設の耐震対策の進捗と今後の計画を伺う。",
“A”: “東部低地帯の河川における耐震対策について **...(omission)...**首都東京の安全性を高めてまいります。",
"AnswerSummary": "層連携して、液状化を含めた耐震対策を推進し、首都東京の安全性を高めてまいります。",
"PredictedClass": false

(3) "AnswerOriginal"と"AnswerSummary"の数値が異なる例
"QuestionSummary": "学校現場の看護師の操作を認めない状況は改善すべき。"
“AnswerOriginal”: “次に、都立特別支援学校における**...(omission)...**準備を進めてまいります。",
"AnswerSummary": "２６年度内にガイドラインを策定、周知し、対象の児童生徒一人一人の状況に応じて校内管理体制を整えていく。",
"PredictedClass": false

(4) "AnswerSummary"の内容が"AnswerOriginal"に含まれていない、正しく要約していない例
"QuestionSummary": "民間の創意工夫を応援し持続的で発展的な取組にすべき。",
"AnswerOriginal": "まず、スマートシティー構築に向けた支援について**...(omission)...**を図っていくものでございます。",
"AnswerSummary": "人一人にとってテクノロジーが身近なものになるよう、普及啓発を促進してまいります。",
"PredictedClass": false

以上の例を参考にし、これから入力する各番号で示す3つの文章の関係性をtrueかfalseで列挙してください。

(output1)

ご了承ください。具体的な文章の例を提供していただく必要があります。3つの文章（"QuestionSummary", "AnswerOriginal", "AnswerSummary"）を以下の形式で提供してください。
1."QuestionSummary": "質問内容", "AnswerOriginal": "オリジナルの答えの内容", "AnswerSummary": "答えの要約",
2."QuestionSummary": "質問内容", "AnswerOriginal": "オリジナルの答えの内容", "AnswerSummary": "答えの要約",
3."QuestionSummary": "質問内容", "AnswerOriginal": "オリジナルの答えの内容", "AnswerSummary": "答えの要約",
これらを提供いただければ、それぞれの文章の関係性（trueかfalse）を評価します。

(input2)

以下(1)~(3)の文章の関係性（trueかfalse）をそれぞれ示せ

(1)
"QuestionSummary": "熱中症の増加に対し、誰一人取り残さない搬送体制の充実への取組は。",
"AnswerOriginal": "熱中症による救急搬送について**...(Omission)...**に努めてまいります。",
"AnswerSummary": "熱中症になりそうな人をAIで検出し、事前に通知を送ることで搬送を円滑にします。",

(2)
...(Omission)...

(3)
...(Omission)...

(output2)

以下のように、(1)から(3)の文章の関係性を示します：

(1)
関係性: False
理由: "QuestionSummary"では「熱中症の増加に対し、誰一人取り残さない搬送体制の充実への取組は」と述べられています。それに対して、"AnswerSummary"では「熱中症になりそうな人をAIで検出し、事前に通知を送ることで搬送を円滑にします」と述べられていますが、その具体的な取組やAIの使用についての詳細が"AnswerOriginal"では言及されていません。

(2)
...(Omission)...

(3)
...(Omission)...

Figure 1: Few-shot prompt with all 3 items

3 METHODS

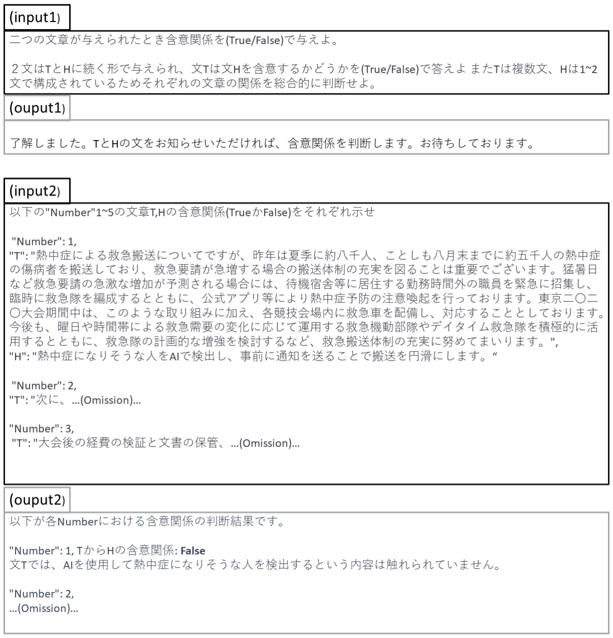
We tried three methods using three main items: "QuestionSummary"(QS), which is a summary of the question from the Q and A session that appeared in the minutes from each of the data used for the AV subtask; "AnswerSummary"(AS), which is a few lines of correct and fake answers to this question; and "AnswerOriginal"(AO),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

---

[1] https://chat.openai.com/

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 show 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.


(input1)

二つの文章が与えられたとき含意関係を(True/False)で与えよ。

２文はTとHに続く形で与えられ、文Tは文Hを含意するかどうかを(True/False)で答えよ またTは複数文、Hは1~2
文で構成されているためそれぞれの文章の関係を総合的に判断せよ。

(ouput1)

了解しました。TとHの文をお知らせいただければ、含意関係を判断します。お待ちしております。

(input2)

以下の"Number"1~5の文章T,Hの含意関係(TrueかFalse)をそれぞれ示せ

"Number": 1,
"T": "熱中症による救急搬送についてですが、昨年は夏季に約八千人、ことしも八月末までに約五千人の熱中症の傷病者を搬送しており、救急要請が急増する場合の搬送体制の充実を図ることは重要でございます。猛暑日など救急要請の急激な増加が予測される場合には、待機宿舎等に居住する勤務時間外の職員を緊急に招集し、臨時に救急隊を編成するとともに、公式アプリ等により熱中症予防の注意喚起を行っております。東京二〇二〇大会期間中は、このような取り組みに加え、各競技会場内に救急車を配備し、対応することとしております。今後も、曜日や時間帯による救急需要の変化に応じて運用する救急機動部隊やデイタイム救急隊を積極的に活用するとともに、救急隊の計画的な増強を検討するなど、救急搬送体制の充実に努めてまいります。",
"H": "熱中症になりそうな人をAIで検出し、事前に通知を送ることで搬送を円滑にします。“

"Number": 2,
"T": "次に、...(Omission)...

"Number": 3,
"T": "大会後の経費の検証と文書の保管、...(Omission)...

(ouput2)

以下が各Numberにおける含意関係の判断結果です。

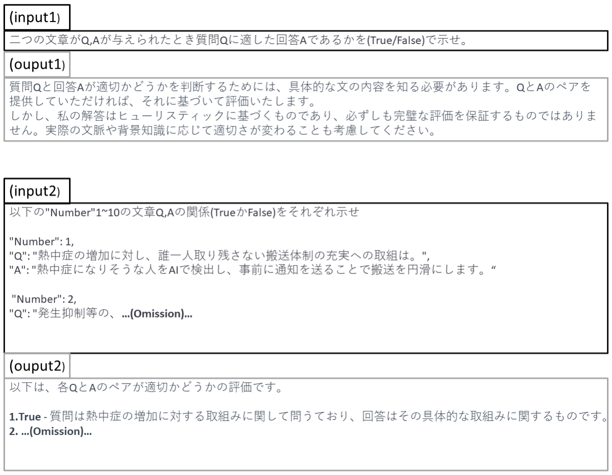
"Number": 1, TからHの含意関係: **False**
文Tでは、AIを使用して熱中症になりそうな人を検出するという内容は触れられていません。

"Number": 2,
...(Omission)...

Figure 2: Prompt with "AnswerOriginal" and "AnswerSummary"

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.


(input1)

二つの文章がQ,Aが与えられたとき質問Qに適した回答Aであるかを(True/False)で示せ。

(ouput1)

質問Qと回答Aが適切かどうかを判断するためには、具体的な文の内容を知る必要があります。QとAのペアを提供していただければ、それに基づいて評価いたします。
しかし、私の解答はヒューリスティックに基づくものであり、必ずしも完璧な評価を保証するものではありません。実際の文脈や背景知識に応じて適切さが変わることも考慮してください。

(input2)

以下の"Number"1~10の文章Q,Aの関係(TrueかFalse)をそれぞれ示せ

"Number": 1,
"Q": "熱中症の増加に対し、誰一人取り残さない搬送体制の充実への取組は。",
"A": "熱中症になりそうな人をAIで検出し、事前に通知を送ることで搬送を円滑にします。“

"Number": 2,
"Q": "発生抑制等の、**...(Omission)...**

(ouput2)

以下は、各QとAのペアが適切かどうかの評価です。

1.True - 質問は熱中症の増加に対する取組みに関して問うており、回答はその具体的な取組みに関するものです。
2. ...(Omission)...

Figure 3: Prompt with "QuestionSummary" and "AnswerSummary"

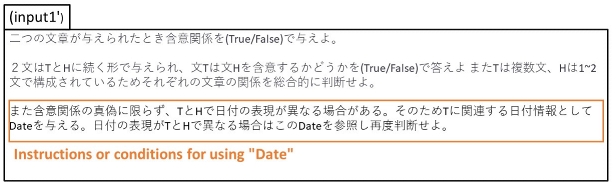
3.2.1 Subtask 1: Examine entailment relations between "AnswerOriginal" and "AnswerSummary"([AO+AS]). First, we input the instructions and conditions shown in input1 of Figure 2. In input 1, AO is "T" and AS is "H", instructing ChatGPT to answer True or False for the entailment relation of the two texts.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


(input1')

二つの文章が与えられたとき含意関係を(True/False)で与えよ。

２文はTとHに続く形で与えられ、文Tは文Hを含意するかどうかを(True/False)で答えよ またTは複数文、Hは1~2文で構成されているためそれぞれの文章の関係を総合的に判断せよ。

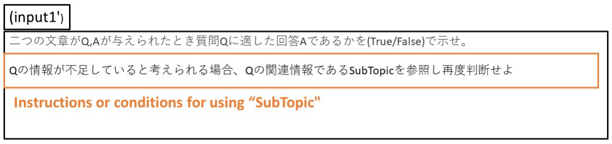
また含意関係の真偽に限らず、TとHで日付の表現が異なる場合がある。そのためTに関連する日付情報としてDateを与える。日付の表現がTとHで異なる場合はこのDateを参照し再度判断せよ。

Instructions or conditions for using "Date"

Figure 4: Prompting by "AnswerOriginal" and "AnswerSummary" with "Date" information added

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 vage date expressions such as "来月"(next month), while AS has specific date expression like "2 年 3 月"(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.


(input1')

二つの文章がQ,Aが与えられたとき質問Qに適した回答Aであるかを(True/False)で示せ。

Qの情報が不足していると考えられる場合、Qの関連情報であるSubTopicを参照し再度判断せよ

Instructions or conditions for using “SubTopic"

Figure 5: Prompting by "QuestionSummary" and "AnswerSummary" with "SubTopic" information added

3.3.2 Add "SubTopic" to Subtask 2 in Section 3.3. The QS had sentences that briefly asked questions such as "どう取り組む。"(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.

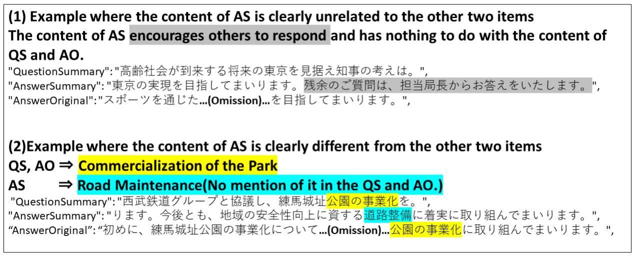
5 DISCUSSIONS

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).

Table 1: The results of Answer Verification

ID:Method	Accuracy	Correct	Fake
223:(QS+AO+AS)	58.00%	97.30%(36/37)	34.92%(22/63)
:([AO+AS],[QS+AS])	77.00%	89.19%(33/37)	69.84%(44/63)
:([AO+AS+Date],[QS+AS+ST])	78.00%	100.00%(37/37)	65.08%(41/63)


(1) Example where the content of AS is clearly unrelated to the other two items
The content of AS encourages others to respond and has nothing to do with the content of QS and AO.
"QuestionSummary":"高齢社会が到来する将来の東京を見据え知事の考えは。",
"AnswerSummary":"東京の実現を目指してまいります。残余のご質問は、担当局長からお答えをいたします。",
"AnswerOriginal":"スポーツを通じた**...(Omission)...**を目指してまいります。",

(2)Example where the content of AS is clearly different from the other two items
QS, AO ⇒ Commercialization of the Park
AS ⇒ Road Maintenance(No mention of it in the QS and AO.)
"QuestionSummary":"西武鉄道グループと協議し、練馬城址公園の事業化を。",
"AnswerSummary":"ります。今後とも、地域の安全性向上に資する道路整備に着実に取り組んでまいります。",
“AnswerOriginal”:“初めに、練馬城址公園の事業化について**...(Omission)...**公園の事業化に取り組んでまいります。",

Figure 6: Examples where the method 1 (QS+AO+AS) correctly determined fake answers

It is consider that these can correctly classify fake answers because phrases appearing in AS often differ from those in AO.

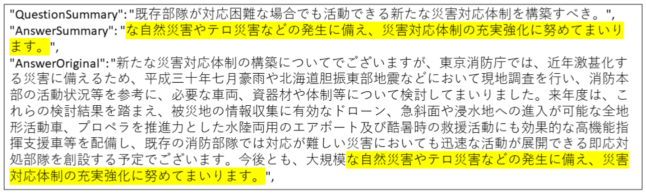

"QuestionSummary":"既存部隊が対応困難な場合でも活動できる新たな災害対応体制を構築すべき。",
"AnswerSummary": "な自然災害やテロ災害などの発生に備え、災害対応体制の充実強化に努めてまいります。",
"AnswerOriginal":"新たな災害対応体制の構築についてでございますが、東京消防庁では、近年激甚化する災害に備えるため、平成三十年七月豪雨や北海道胆振東部地震などにおいて現地調査を行い、消防本部の活動状況等を参考に、必要な車両、資器材や体制等について検討してまいりました。来年度は、これらの検討結果を踏まえ、被災地の情報収集に有効なドローン、急斜面や浸水地への進入が可能な全地形活動車、プロペラを推進力とした水陸両用のエアボート及び酷暑時の救援活動にも効果的な高機能指揮支援車等を配備し、既存の消防部隊では対応が難しい災害においても迅速な活動が展開できる即応対処部隊を創設する予定でございます。今後とも、大規模な自然災害やテロ災害などの発生に備え、災害対応体制の充実強化に努めてまいります。",

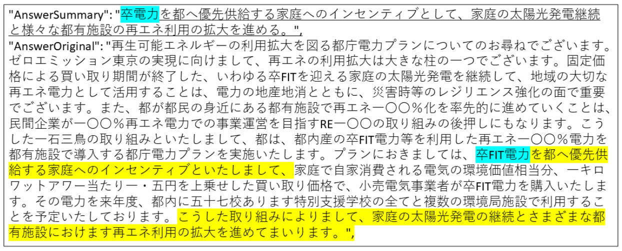
Figure 7: Example of fake answer that is misclassified although the sentence is broken in the middle.

Here are some examples of misclassified fake answers.

First, as shown in Figure 7, the sentence of AS is broken in the middle. Such AS contains the same phrase as AO, as shown in the yellow part of Figure 7. Therefore, ChatGPT may mistakenly conclude that there is a relationship between the two items.

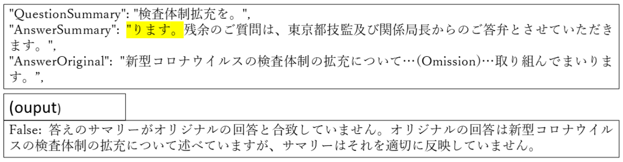
Also, as shown in the blue part of Figure 8, there is a case in which an AS sentence with broken proper nouns is mistakenly judged as a correct answer. In this case, as in the previous case, the same phrase AO appears, so the ChatGPT may mistakenly judge that there is an entailment with the fake answer data.

As shown in Figure 9, As shown in Figure 9, even fake answers that could have been correctly classified may not be mentioned by ChatGPT as evidence of broken sentences or proper nouns. Based



"AnswerSummary": "卒電力を都へ優先供給する家庭へのインセンティブとして、家庭の太陽光発電継続と様々な都有施設の再エネ利用の拡大を進める。",
"AnswerOriginal": "再生可能エネルギーの利用拡大を図る都庁電力プランについてのお尋ねでございます。ゼロエミッション東京の実現に向けまして、再エネの利用拡大は大きな柱の一つでございます。固定価格による買い取り期間が終了した、いわゆる卒FITを迎える家庭の太陽光発電を継続して、地域の大切な再エネ電力として活用することは、電力の地産地消とともに、災害時等のレジリエンス強化の面で重要でございます。また、都が都民の身近にある都有施設で再エネ一〇〇%化を率先的に進めていくことは、民間企業が一〇〇%再エネ電力での事業運営を目指すRE一〇〇の取り組みの後押しにもなります。こうした一石三鳥の取り組みといたしまして、都は、都内産の卒FIT電力等を利用した再エネ一〇〇%電力を都有施設で導入する都庁電力プランを実施いたします。プランにおきましては、卒FIT電力を都へ優先供給する家庭へのインセンティブといたしまして、家庭で自家消費される電気の環境価値相当分、一キロワットアワー当たり一・五円を上乗せした買い取り価格で、小売電気事業者が卒FIT電力を購入いたします。その電力を来年度、都内に五十七校あります特別支援学校の全てと複数の環境局施設で利用することを予定いたしております。こうした取り組みによりまして、家庭の太陽光発電の継続とさまざまな都有施設におけます再エネ利用の拡大を進めてまいります。",

Figure 8: Example of fake answer that is misclassified although the sentence contains partially missing proper noun



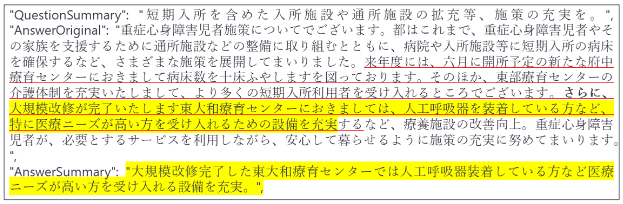
"QuestionSummary": "検査体制拡充を。",
"AnswerSummary": "ります。残余のご質問は、東京都技監及び関係局長からのご答弁とさせていただきます。",
"AnswerOriginal": "新型コロナウイルスの検査体制の拡充について…(Omission)…取り組んでまいります。”,

(ouput)

False: 答えのサマリーがオリジナルの回答と合致していません。オリジナルの回答は新型コロナウイルスの検査体制の拡充について述べていますが、サマリーはそれを適切に反映していません。

Figure 9: Example of correctly classified fake answer. No mention of broken sentence.

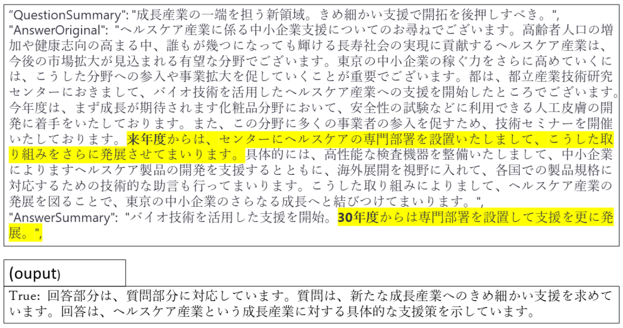
on these facts, it is considered that ChatGPT is unable to recognize fatal flaws as a summary, such as broken sentence and proper noun that cannot be established as correct summaries when AS has many phrases in common with AO.



"QuestionSummary": "短期入所を含めた入所施設や通所施設の拡充等、施策の充実を。",
"AnswerOriginal": "重症心身障害児者施策についてでございます。都はこれまで、重症心身障害児者やその家族を支援するために通所施設などの整備に取り組むとともに、病院や入所施設等に短期入所の病床を確保するなど、さまざまな施策を展開してまいりました。来年度には、六月に開所予定の新たな府中療育センターにおきまして病床数を十床ふやしますを図っております。そのほか、東部療育センターの介護体制を充実いたしまして、より多くの短期入所利用者を受け入れるところでございます。**さらに、**大規模改修が完了いたします東大和療育センターにおきましては、人工呼吸器を装着している方など、特に医療ニーズが高い方を受け入れるための設備を充実するなど、療養施設の改善向上。重症心身障害児者が、必要とするサービスを利用しながら、安心して暮らせるように施策の充実に努めてまいります。
",
"AnswerSummary": "大規模改修完了した東大和療育センターでは人工呼吸器装着している方など医療ニーズが高い方を受け入れる設備を充実。",

Figure 10: Example of AS content not giving a response suitable for QS. An answer suitable for QS are mentioned not only in the yellow part (same content as AS) in AO, but also in the red underlined part.

Next, Figure 10 shows that ChatGPT misclassified the fake answer as the correct answer, when AS did not give a suitable answer to QS. In this case, as in the previous case, AS contains the same phrase as AO. The difference from the previous case is that the 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.



“QuestionSummary": "成長産業の一端を担う新領域。きめ細かい支援で開拓を後押しすべき。",
"AnswerOriginal": "ヘルスケア産業に係る中小企業支援についてのお尋ねでございます。高齢者人口の増加や健康志向の高まる中、誰もが幾つになっても輝ける長寿社会の実現に貢献するヘルスケア産業は、今後の市場拡大が見込まれる有望な分野でございます。東京の中小企業の稼ぐ力をさらに高めていくには、こうした分野への参入や事業拡大を促していくことが重要でございます。都は、都立産業技術研究センターにおきまして、バイオ技術を活用したヘルスケア産業への支援を開始したところでございます。今年度は、まず成長が期待されます化粧品分野において、安全性の試験などに利用できる人工皮膚の開発に着手をいたしております。また、この分野に多くの事業者の参入を促すため、技術セミナーを開催いたしております。**来年度**からは、センターにヘルスケアの専門部署を設置いたしまして、こうした取り組みをさらに発展させてまいります。具体的には、高性能な検査機器を整備いたしまして、中小企業によりますヘルスケア製品の開発を支援するとともに、海外展開を視野に入れて、各国での製品規格に対応するための技術的な助言も行ってまいります。こうした取り組みによりまして、ヘルスケア産業の発展を図ることで、東京の中小企業のさらなる成長へと結びつけてまいります。",
"AnswerSummary": "バイオ技術を活用した支援を開始。**30年度**からは専門部署を設置して支援を更に発展。",

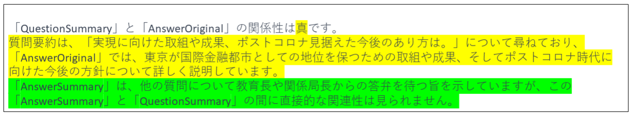
(ouput)

True: 回答部分は、質問部分に対応しています。質問は、新たな成長産業へのきめ細かい支援を求めています。回答は、ヘルスケア産業という成長産業に対する具体的な支援策を示しています。

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.



「QuestionSummary」と「AnswerOriginal」の関係性は真です。
質問要約は、「実現に向けた取組や成果、ポストコロナ見据えた今後のあり方は。」について尋ねており、「AnswerOriginal」では、東京が国際金融都市としての地位を保つための取組や成果、そしてポストコロナ時代に向けた今後の方針について詳しく説明しています。
「AnswerSummary」は、他の質問について教育長や関係局長からの答弁を待つ旨を示していますが、この「AnswerSummary」と「QuestionSummary」の間に直接的な関連性は見られません。

Figure 12: Example of ChatGPT output that may not correctly take account for all of three items.

Finally, ChatGPT may consider only two of the three input items for some data. Figure 12 shows the text output by ChatGPT as a basis for determining the correctness of data. The green part of Figure 12 refers to the relationship between "AnswerSummary" and "QuestionSummary". However, the actual output is considered to be judged only by the relationship between "AnswerSummary" and "AnswerOriginal," which is shown in the yellow part of Figure 12.

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

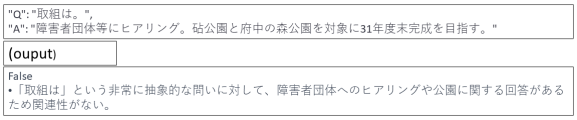
Table 1 show that Method 2 was able to correctly classify more erroneous fake answers than Method1. This can be thanks to the performance of ChatGPT for the recognition of entailment relations. Table 2 shows the classification results of the test data using

Table 2: The results of only sub-task1(AO+AS)

ID:Method	Accuracy	Correct	Fake
:Sub-task1	77.00%	100.00%(37/37)	63.49%(40/63)

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 treat 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.

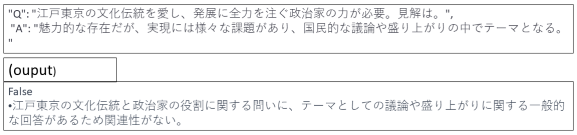

"Q": "取組は。",
"A": "障害者団体等にヒアリング。砧公園と府中の森公園を対象に31年度末完成を目指す。"

(ouput)

False
•「取組は」という非常に抽象的な問いに対して、障害者団体へのヒアリングや公園に関する回答があるため関連性がない。

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 "取組は。"(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.

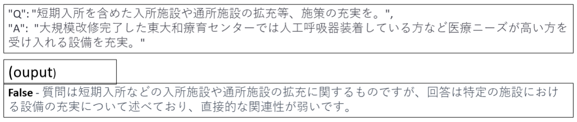

"Q": "江戸東京の文化伝統を愛し、発展に全力を注ぐ政治家の力が必要。見解は。",
"A": "魅力的な存在だが、実現には様々な課題があり、国民的な議論や盛り上がりの中でテーマとなる。
"

(ouput)

False
•江戸東京の文化伝統と政治家の役割に関する問いに、テーマとしての議論や盛り上がりに関する一般的な回答があるため関連性がない。

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, "見解は。"(Please 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.


"Q": "短期入所を含めた入所施設や通所施設の拡充等、施策の充実を。",
"A": "大規模改修完了した東大和療育センターでは人工呼吸器装着している方など医療ニーズが高い方を受け入れる設備を充実。"

(ouput)

False - 質問は短期入所などの入所施設や通所施設の拡充に関するものですが、回答は特定の施設における設備の充実について述べており、直接的な関連性が弱いです。

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

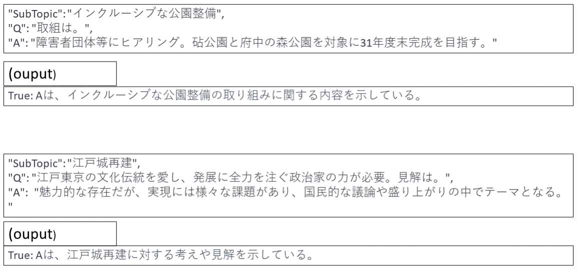
However, the number of correctly classified fake answers increases when the results of Subtask 2 are taken into account. The fake answers correctly classified by Subtask 2 are those for which the relationship between QS and AS is not appropriate, such as Figure 10 in Section 5.1. Figure 15 shows the 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 Method3 reduces the misclassification of correct answers compared to Method2.


"SubTopic": "インクルーシブな公園整備",
"Q": "取組は。",
"A": "障害者団体等にヒアリング。砧公園と府中の森公園を対象に31年度末完成を目指す。"

(ouput)

True: Aは、インクルーシブな公園整備の取り組みに関する内容を示している。

"SubTopic": "江戸城再建",
"Q": "江戸東京の文化伝統を愛し、発展に全力を注ぐ政治家の力が必要。見解は。",
"A": "魅力的な存在だが、実現には様々な課題があり、国民的な議論や盛り上がりの中でテーマとなる。
"

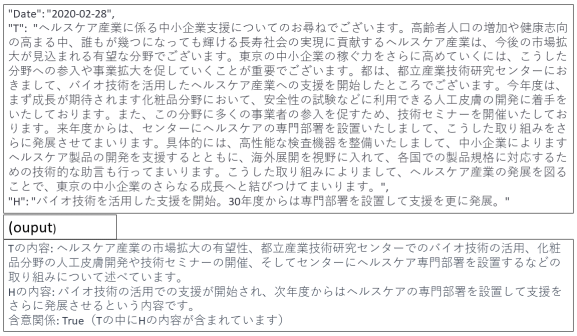
(ouput)

True: Aは、江戸城再建に対する考えや見解を示している。

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

As shown in Figure 16, abstruct 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.



"Date": "2020-02-28",
"T": "ヘルスケア産業に係る中小企業支援についてのお尋ねでございます。高齢者人口の増加や健康志向の高まる中、誰もが幾つになっても輝ける長寿社会の実現に貢献するヘルスケア産業は、今後の市場拡大が見込まれる有望な分野でございます。東京の中小企業の稼ぐ力をさらに高めていくには、こうした分野への参入や事業拡大を促していくことが重要でございます。都は、都立産業技術研究センターにおきまして、バイオ技術を活用したヘルスケア産業への支援を開始したところでございます。今年度は、まず成長が期待されます化粧品分野において、安全性の試験などに利用できる人工皮膚の開発に着手をいたしております。また、この分野に多くの事業者の参入を促すため、技術セミナーを開催いたしております。来年度からは、センターにヘルスケアの専門部署を設置いたしまして、こうした取り組みをさらに発展させてまいります。具体的には、高性能な検査機器を整備いたしまして、中小企業によりますヘルスケア製品の開発を支援するとともに、海外展開を視野に入れて、各国での製品規格に対応するための技術的な助言も行ってまいります。こうした取り組みによりまして、ヘルスケア産業の発展を図ることで、東京の中小企業のさらなる成長へと結びつけてまいります。",
"H": "バイオ技術を活用した支援を開始。30年度からは専門部署を設置して支援を更に発展。"

(ouput)

Tの内容: ヘルスケア産業の市場拡大の有望性、都立産業技術研究センターでのバイオ技術の活用、化粧品分野の人工皮膚開発や技術セミナーの開催、そしてセンターにヘルスケア専門部署を設置するなどの取り組みについて述べています。
Hの内容: バイオ技術の活用での支援が開始され、次年度からはヘルスケアの専門部署を設置して支援をさらに発展させるという内容です。
含意関係: True（Tの中にHの内容が含まれています）

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 implicature 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 implicature 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

[1] Yasuhiro Ogawa, Yasutomo Kimura, Hideyuki Shibuki, Hokuto Ototake, Yuzu Uchida, Keiichi Takamaru, Kazuma Kadowaki, Tomoyoshi Akiba, Minoru Sasaki, Akio Kobayashi, Masaharu Yoshioka, Tatsunori Mori, Kenji Araki, Satoshi Sekine, and Teruko Mitamura. 2023. Overview of the NTCIR-17 QA Lab-PoliInfo-4 Task. *Proceedings of The 17th NTCIR Conference.* https://doi.org/10.20736/0002001326

[2] Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476* (2023).