生成 AI は含意関係認識ができるのか

荒沢 康平 [1]　狩野 芳伸 [1]
[1] 静岡大学
{karasawa,kano}@kanolab.net

概要

含意関係認識は論理的過程の基盤であり、たとえばファクトチェックなど様々な重要タスクの下敷きとなる。ファインチューン等で表層的なパターンを学ばせれば含意関係認識の評価スコアは向上するが、必ずしも含意関係認識そのものが解けて論理的な演算一般ができることにはならない。本研究では、ファインチューンや few shot prompting 抜きに GPT-4 がどのような場合に含意関係認識に失敗するかを詳細に分析し、特に時制、数量詞・数詞の認識に課題があること、またわずかな単語や語順の違いで結果が大きく変動することを示した。現在の大規模言語モデルの限界を示すとともに、今後の言語モデル改良に向けた手がかりになると期待される。

1　はじめに

大規模言語モデル（LLM）で実装された生成 AI の登場によって、AI によるコーディングや文章生成など業務の効率化が期待されている。中でも OpenAI 社の提供する GPT-4 は、大量の学習データとパラメータを用いて訓練され、その能力が高く評価されているモデルの一つである。しかし、生成結果に「ハルシネーション」（事実に反する出力）[1] を含むことが問題となっている。さらに、機械学習全般にその推論過程を説明することは困難で、生成 AI に説明を生成させたとしてもその説明が実際の推論過程と一致するという舗保証はない。こうしたことは、LLM が自然言語の論理的側面をどの程度理解しているかという疑問を引き起こしている。

本研究では、論理的過程の基盤である含意関係認識タスクを通じて LLM の論理的推論能力を調査する。含意関係認識とは、前提文と仮定文のペアに対し、前提文が真の場合に仮定文の真偽を判断するもので、含意、中立、矛盾の 3 値に分類することが多い。たとえば、前提文を「事実」とし、仮定文を「SNS などに投稿された文章」とすれば、ファクトチェックに応用できる。

既存の含意関係認識タスクの評価スコアは高いが、ファインチューンや few shot prompting により表層的パターンを学んでいるだけで、含意関係そのものを扱えていない可能性がある。本研究では分類器のファインチューンと LLM の few shot prompting により既存の LLM の性能を検証したうえで、LLM の zero shot prompting でさまざまな含意関係認識データのどのような場合に間違うのか分析を行い、現在の LLM の性能の限界を推測する。

2　関連研究

2.1　含意関係認識データセット

日本語の含意関係認識データセットには、JSNLI[2]、JSICK[3]、JNLI[4] などがある。

JSNLI[2] は英語の大規模な含意関係認識データセットである SNLI[5] を日本語に機械翻訳したもので、事前学習済み BERT[6] をファインチューンし Accuracy 0.929 の分類性能を報告している。JSICK[3] は様々な言語現象を含んだ英語の含意関係認識データセットである SICK[7] を人手で日本語に翻訳したもので、事前学習済み BERT をファインチューンしたモデルで Accuracy 0.84 の正答率を達成している。JNLI[4] は、日本語理解ベンチマーク JGLUE[4] に含まれる含意関係認識データセットで、翻訳を介さずに作成された。JNLI はクラウドソーシングで構築され、画像の内容を文章で表現させることで、含意と中立のラベルを持つ文ペアを作成している。矛盾ラベルの文ペアは、表現した文章に対して矛盾する内容をクラウドソーシングで作成させた。事前学習済み BERT をファインチューンし Accuracy 0.906 を達成している。

2.2　LLM を用いた含意関係認識

GPT は Transformer[8] の Encoder 部分を活用したモデルで、OpenAI によって提供されている GPT-4[9]

は、GPT に加え InstructGPT で追加的な学習をした LLM である。GPT-4 は多言語処理能力を備えており、さまざまなベンチマークにおいて人間に匹敵しうるパフォーマンスを示している。現時点では、GPT-4 を代表モデルとして性能検証することは妥当と考えられる。Nejumi LLM リーダーボード Neo[1)] で公開されている、gpt-4-1106-preview で JNLI を学習・評価した結果では、Accuracy 0.77 を達成している。

OpenAI の GPT-4 以外にも様々な LLM がある。llm-jp-eval リーダーボード [10] に掲載されている、JNLI の性能上位 2 件[2)] の LLM は、それぞれ Accuracy が 0.916, 0.91 である。

なお、GPT-4 よりも後者二つの LLM の性能が大きく上に見えるのは、後者二つのインストラクション訓練データに JNLI が含まれているためと考えられる。

3 含意関係認識問題の事前分析

含意関係認識タスクに出現する問題は、多様なパターンがある。その中で分析対象を絞るために、GPT-4 の few shot prompting および BERT とその亜種のファインチューンを行い性能評価と失敗事例の分析を行った。各モデルの実験結果を表 3 に示す。実験詳細は付録 A を参照されたい。

失敗事例の分析の結果、図 1 に示すように GPT-4 は「時制」と「数量詞・数詞」を不得意とする可能性が示唆された。数量詞とは「たくさん」など、数詞とは「1 つ、2 つ」などを指す。次節以降で、この二種を対象に分析を行う。

この結果から、4 章において「時制」「数量詞・数詞」をどの程度見極められているかを、それぞれに特化したデータセットを作成し GPT-4 の性能を調査する。

4 GPT-4 zero shot の含意関係認識

前述のように、ファインチューンないし few shot prompting を行えばその表層パターンを学習し対応できる可能性があるが、現実の利用では含意関係認識は中間的なステップであり、含意関係認識に特化したチューニングを前提にすることはできない。また、含意関係認識のパターンも表層的、深層的にもさまざまであり、few shot prompting でカバーするのは困難である。そこで、チューニングを施さない「素の」GPT-4 を対象に、すなわち zero-shot で性能を測り分析する。

具体的には、現実的な設定として「ファクトチェック」を想定したプロンプトを用いる。そのうえで、前節で述べた時制判別と数量詞・数詞判別を試すデータセットを作成し、失敗事例を分析する。

・時制で間違えた問題例
前提文：路上にある赤い消火栓を使って消火活動を**しています**。
仮定文：消火栓からの消火活動が**終わった**。
正解：矛盾
GPT-4 の回答：中立

・数量詞・数詞で間違えた問題例
前提文：**たくさん**の人が凧を見上げています。
仮定文：一**人**が凧を見上げている。
正解：矛盾
GPT-4 の回答：含意

図 1 GPT-4 が不正解だった問題例

4.1 データセット

時制データセット 基本となる文の時制を変化させることで含意・矛盾文ペアを作成したのが時制データセットである。

具体的には図 2 に示すように、GPT-4 で生成した基本となる文を用意し、その文末を変える事によって時制を変化させる。本データセットでは現在進行形・過去形・未来形を使用し、前提文と仮定文の時制を揃えた場合に含意、時制が異なる場合に矛盾ラベルを付与した。中立はなく、2 値ラベルである。含意 300 件、矛盾 300 件の計 600 件を作成した。

基本文： 私は本を読んでいます。

含意関係
前提文: 私は本を読んでいます。(現在進行形)
仮定文: 私は本を読んでいる。(現在進行形)

矛盾関係
前提文: 私は本を読んでいます。(現在進行形)
仮定文: 私は本を読みました。(過去形)

図 2 時制データセットの例

数量詞・数詞データセット 基本となる文の数量詞を変化させることで含意・中立・矛盾の 3 値ラベルからなる文ペアを作成したのが数量詞・数詞デー

1) https://wandb.ai/wandb-japan/llm-leaderboard/reports/Nejumi-LLM-Neo--Vmlldzo2MTkyMTU0

2) llm-jp/elyza-ELYZA-japanese-Llama-2-7b-fast-instruct-full-jaster)[11][12] および llm-jp/llm-jp-13b-instruct-full-jaster-dolly-oasst-v1.0(https://huggingface.co/llm-jp/llm-jp-13b-instruct-full-jaster-dolly-oasst-v1.0)

タセットである。

含意の文ペアは、前提文と仮定文に意味が同じ数量詞を用いるパターンと、前提文に数量詞・数詞を、仮定文に基本文を用いるパターンの大きく分けて 2 パターンで構成した。中立の文ペアは、前提文に基本文を、仮定文に数量詞・数詞を用いるパターンで構成した。矛盾の文ペアは、前提文に数量詞を、仮定文にそれと矛盾した数量詞・数詞を用いるパターンで構成した。含意・中立・矛盾各 200 件、計 600 件の文ペアを作成した（付録の図 6）。

4.2 プロンプト

前述の通り、zero shot プロンプトでの推論を行った。現実的な設定としてファクトチェックを想定し、図 3 に示すように前提文が事実である場合に仮定文のニュースが事実であるか否かを答えさせるプロンプトを使用した。

```
# 指示 #
#前提文#が事実であるとした場合に、#仮定文#の内容が事実であるか、事実でないかを判定せよ。
#仮定文#の内容が事実であると判定した場合は「0」、事実であるか事実でないか#前提文#からは判定できない場合は「1」、事実でないと判定した場合は「2」と出力せよ。
次の項目のみを返答せよ。
判定:[判定]


#前提文#
人々が防寒具を求めて、衣料品店を訪れていました。
#仮定文#
人々が防寒具を求めて、衣料品店を訪れていた。
```

図 3 事実か事実でないかを尋ねるプロンプトの例

2 値分類である「時事・時制データセット」では「指示」の中の「事実であるか事実に反するか判断がつかない場合は「中立」」の文言を削除する。

4.3 実験と結果

gpt-4-1106-preview を使用して zero shot prompting を行い、出力の一貫性を極力保つため temperature を 0.1 に設定した。

「時制」、「数量詞・数詞」と JNLI データセットを GPT-4 zero shot で実行した結果を表 1 に示す。表 1 の「事実」ラベルはデータセットの「含意」ラベルに、「事実でない」ラベルはデータセットの「矛盾」ラベルにそれぞれ対応している。Accuracy が「時制」で 0.52、「数量詞・数詞」は 0.74、「時制」については、ランダムベースラインの 0.50 と大差ない値で、総じて実用的とは言い難い性能であった。

表 1 データセット毎の性能比較 (p : precision , r : recall)

	事実	中立	事実でない
時制	p : 0.51	p : —	p : 1.00
	r : 1.00	r : —	r : 0.03
Accuracy:0.52	f1 : 0.67	f1 : —	f1 : 0.06
数量詞	p : 0.64	p : 0.78	p : 0.87
	r : 0.86	r : 0.75	r : 0.61
Accuracy:0.74	f1 : 0.74	f1 : 0.77	f1 : 0.72

5 失敗事例の分析・考察

5.1 時制の含意関係認識

表 1 の「事実でない」ラベルに注目すると、極端に precision が高く recall が低く、「事実でない」と回答する数が極端に少なかったことがわかる。

プロンプトの検証 プロンプトによって出力、特にこうした分類タスクのバランスが大きく変わる可能性がある。特に回答に関連の深い単語の影響を調べるため、プロンプトの「事実である」「事実でない」という表現を「正しい」「正しくない」に変更して実行した。結果、「事実でない」と回答する数が増加したものの、極端に少ないという結果は変わらなかった。表 2 にこの 2 種のプロンプトの比較結果を示す。表 2 内の「使用時制」とは問題の前提文と仮定文に用いた時制のペアである。また、分数の分母は各時制ペアの総数を表している。「現在・未来」ペア (前提文に現在形、仮定文に未来形の問題) に関しては両プロンプトにおいて正解数が 0 件であった。全体に、時制にかかわる含意関係認識の性能が低いことを示唆している。

表 2 使用時制毎の正解数

使用時制	「事実」プロンプト	「正しい」プロンプト
現在・未来	0 / 50	0 / 50
未来・現在	0 / 50	4 / 50
過去・未来	8 / 50	4 / 50
未来・過去	0 / 50	5 / 50
現在・過去	0 / 50	3 / 50
過去・現在	2 / 50	4 / 50

失敗の具体例 時制が「現在・未来」ペアの失敗例（図 4 の時制の不正解例）を見ると、「現時点ですでに地元住民が参加している」ことが事実であるにもかかわらず、「現時点ではまだ地元住民が参加していない」のが事実であると、GPT-4 が出力している。このような問題は、フェイクニュースの検出をはじめイベント発生の有無にかかわる根本的な部分であり、LLM の出力の信頼性に疑問が生じる。

・時制の不正解例
前提文：地域の老人ホームの文化祭には、地元住民が参加**しています。(現在形)**
仮定文：地域の老人ホームの文化祭には、地元住民が参加**するでしょう。(未来形)**
正解：**事実でない**
GPT-4 の回答：**事実**

・数量詞の不正解例 1
前提文：**大量**の人々が防寒具を求めて、衣料品店を訪れています。**(数量詞あり)**
仮定文：**1 人**の人が防寒具を求めて、衣料品店を訪れています。**(数詞あり)**
正解：**事実でない**
GPT-4 の回答：**事実**

・数量詞の不正解例 2
前提文：地域の老舗劇場が再開したことで、演劇ファンが公演を楽しんでいます。**(数量詞なし)**
仮定文：地域の老舗劇場が再開したことで、**数多く**の演劇ファンが公演を楽しんでいます。**(数量詞あり)**
正解：**中立**
GPT-4 の回答：**事実**

図 4　時制および数量詞・数詞で不正解だった例

5.2　数量詞・数詞の含意関係認識

失敗の具体例　正解ラベルが「事実でない」問題を「事実である」と回答するパターンが多かった。図 4 の数量詞の不正解例 1 を見ると、「大量の人が衣料品店を訪れている」ことが事実であるにもかかわらず、「1 人が衣料品店を訪れている」という情報が事実であると出力されており、数量詞の大小関係を認識できない可能性を示唆している。

正解ラベルが「中立」の問題を「事実」と回答するパターンも多かった。図 4 の数量詞の不正解例 2 に示すように、前提文に数詞がなく仮定文に数詞がある場合に誤答が多かった。この例では前提文に数量詞・数詞がないため、仮定文に数詞がある場合その真偽を判断出来ず正解は中立であるが、「事実」と回答した。形容する内容によっては必ずしも正しいとは言えない場合に対応できず、たとえば虚偽に近い誇張した表現の検出が難しい可能性がある。

出力の信頼性　問題文のうち語彙がほとんど共通しているのに、正解できたり不正解であったりする例がいくつもあった。そこでどの要素が出力の変化に影響するか、動詞や語順を変えて確認した。図 5 の例は、「元の問題文 1」の動詞「減少する」を「増加する」におきかえた場合、節の順序を入れ替えた場合、「元の問題文 2」の「ロシア」を「アメリカ」に置き換えた場合である。いずれも同じ入力を 10 回試行したが、置き換え前後で大きく回答のバランスが変わっており、「増加」「減少」や「ロシア」「アメリカ」はそれぞれ単語の意味合いに応じた単語の出現頻度と事実かどうかの相関が、訓練データにおいて偏っていた可能性が考えられる。

・元の問題文 1
前提文：時間外労働の上限規制によって、山ほどの運送ドライバーの収入が減少する
仮定文：時間外労働の上限規制によって、少しの運送ドライバーの収入が減少する
GPT-4 の回答：**10 回とも「事実」と回答**

・動詞を変化させた場合
前提文：時間外労働の上限規制によって、山ほどの運送ドライバーの収入が増加する
仮定文：時間外労働の上限規制によって、少しの運送ドライバーの収入が増加する
GPT-4 の回答：
「中立」を 9 回、「事実でない」を 1 回出力

・語順を変化させた場合
前提文：山ほどの運送ドライバーの収入が時間外労働の上限規制によって、減少する
仮定文：少しの運送ドライバーの収入が時間外労働の上限規制によって、減少する
GPT-4 の回答：
「事実」を 5 回、「事実でない」を 5 回出力

・元の問題文 2
前提文：ロシアがウクライナとの国境線にたくさんの戦車を配備した
仮定文：ロシアがウクライナとの国境線に 1 台の戦車を配備した
GPT-4 の回答：**「事実」を 7 回、「事実でない」を 3 回出力**

・名詞を変化させた場合
前提文：アメリカがウクライナとの国境線にたくさんの戦車を配備した
仮定文：アメリカがウクライナとの国境線に 1 台の戦車を配備した
GPT-4 の回答：**10 回とも「事実でない」と回答**

図 5　出力の信頼性

6　おわりに

含意関係認識における GPT-4 の失敗事例パターンを分析し、その信頼性に限界があるという示唆を得た。含意関係認識においては出現する字面が同じエンティティは同一事物を指すという前提があるが、本来は文脈に応じてエンティティの参照先を解決する必要がある。背景や文脈を取り入れた現実的な設定のタスクで分析を進めたい。

謝辞

本研究は JSPS 科研費 JP22H00804, JP21K18115, JST AIP 加速課題 JPMJCR22U4, およびセコム科学技術財団特定領域研究助成の支援をうけたものです。

参考文献

[1] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, 2023. 2311.05232 https://arxiv.org/abs/2311.05232.

[2] 吉越卓見, 河原大輔, 黒橋禎夫. 機械翻訳を用いた自然言語推論データセットの多言語化. 情報処理学会 第 244 回自然言語処理研究会, Vol. 2020-NL-244, No. 6, pp. 1–8, 7 2020.

[3] 谷中瞳, 峯島宏次. Jsick: 日本語構成的推論・類似度データセットの構築. 人工知能学会全国大会論文集, Vol. JSAI2021, pp. 4J3GS6f02–4J3GS6f02, 2021.

[4] 栗原健太郎, 河原大輔, 柴田知秀. Jglue: 日本語言語理解ベンチマーク. 自然言語処理, Vol. 30, No. 1, pp. 63–87, 2023.

[5] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In Lluís Màrquez, Chris Callison-Burch, and Jian Su, editors, **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**, pp. 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **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)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[7] Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. A SICK cure for the evaluation of compositional distributional semantic models. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, **Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)**, pp. 216–223, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. **Advances in neural information processing systems**, Vol. 30, , 2017.

[9] OpenAI : Josh Achiam and et al. Gpt-4 technical report, 2023. 2303.08774 https://arxiv.org/abs/2303.08774.

[10] llm-jp-eval リーダーボード, (2024-1 閲覧). https://wandb.ai/llm-jp-eval/test-eval/reports/llm-jp-eval---Vmlldzo1NzE0NjA1?accessToken=s09hm7xrqg43ls8i25am6t0r7iwjpninw\zeelqqgbx53zivlm9s04ixfpv3xgiwm.

[11] Akira Sasaki, Masato Hirakawa, Shintaro Horie, and Tomoaki Nakamura. Elyza-japanese-llama-2-7b, 2023. https://huggingface.co/elyza/ELYZA-japanese-Llama-2-7b.

[12] Hugo Touvron and et al. Llama 2: Open foundation and fine-tuned chat models, 2023. 2307.09288 https://arxiv.org/abs/2307.09288.

[13] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, **Advances in Neural Information Processing Systems**, Vol. 33, pp. 1877–1901. Curran Associates, Inc., 2020. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[14] Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In **Proceedings of International Conference on Learning Representations (ICLR) 2023**, 2023.

A　付録：LLM の含意関係認識性能検証

LLM として GPT-4 と、分類モデルとして BERT とその亜種について、含意関係認識の性能検証を実験条件をそろえて行った。JNLI には評価データが含まれていないため、検証データで評価を行った。

A.1　GPT-4

「含意度」を出力させるプロンプトを使用した。[3] 冒頭に「指示」セクションにおいて含意度を出力させる旨指示した。プロンプト中盤では、Few Shot[13] の場合 JNLI の train-v1.1.json の中から含意・中立・矛盾 1 件ずつを使用し、それぞれの含意度を例示した。プロンプトの最後に推論対象となる前提文と仮定文を提示した。

A.2　BERT 系モデル

cltohoku/bert-base-japanesewhole-word-masking[4]、nlp-waseda/roberta-base-japanese[5]、nlp-waseda/roberta-large-japanese[6]、ku-nlp/deberta-v2-base-japanese)[7]、ku-nlp/deberta-v2-large-japanese[8]、microsoft/deberta-v3-base[14](microsoft/deberta-v3-large)[14] の計 7 つのモデルをファインチューンし使用した。

A.3　実験設定

JNLI の valid-v1.1.json を用いて、含意・中立・矛盾の 3 値分類を行った。

GPT-4 Few Shot, GPT-4 Zero Shot 双方において推論を実施。 GPT-4 のモデルは「gpt-4-1106-preview」を使用。また、temperature を 0.1 に設定し出力の一貫性を保たせた。BERT 系モデルのファインチューニングには JNLI の train-v1.1.json を使用し、学習時ハイパーパラメータは、Epoch 数：5、学習率：2e-5、バッチサイズ：16 とした。

A.4　実験結果

結果を表 3 に示す。DeBERTa v3(microsoft/deberta-v3-base)[14] が最高性能であった。一方、GPT-4 の正答率は 0.73 と、deberta-v3-base の正答率 0.92 よりも 19%低い結果である。さらに、GPT-4 はその他の BERT 系モデルの正答率に及ばない結果となっている。このことから、GPT-4 の含意関係認識性能にはまだ課題があると言えよう。

表 3　モデル性能の比較 (JNLI)

	Accuracy
cltohoku	0.87
roberta-base- japanese	0.86
roberta-large-japanese	0.89
deberta-v2-base-japanese	0.88
deberta-v2- large-japanese	0.88
deberta-v3-base	**0.92**
deberta-v3- large	0.90
GPT-4 Few Shot	0.73
GPT-4 Zero Shot	0.72

基本文 # 空港の通路を人が歩いています。

・含意関係
パターン 1
前提文: 空港の通路**たくさん**の人が歩いています。
(数量詞あり)
仮定文: 空港の通路を**多数**の人が歩いています。
(数量詞あり)
パターン 2
前提文: 空港の通路を **50** 人が歩いています。
(数詞あり)
仮定文: 空港の通路を人が歩いています。
(数詞なし)

・中立関係
パターン 1
前提文: 空港の通路を人が歩いています。
(数量詞なし)
仮定文: 空港の通路を**たくさん**の人が歩いています。
(数量詞あり)
パターン 2
前提文: 空港の通路を人が歩いています。
(数量詞なし)
仮定文: 空港の通路を **5 人**の人が歩いています。
(数詞あり)

・矛盾関係
パターン 1
前提文: 空港の通路を**多数**の人が歩いています。
(数量詞あり)
仮定文: 空港の通路を**少し**の人が歩いています。
(数量詞あり)
パターン 2
前提文: 空港の通路を**たくさん**の人が歩いています。
(数量詞あり)
仮定文: 空港の通路を **1 人**の人が歩いています。
(数詞あり)

図 6　数量詞・数詞特化データセットの例

3)　様々なプロンプトを試行した結果、三値を尋ねるより含意度の方が性能が高かったため。
4)　https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking
5)　https://huggingface.co/nlp-waseda/roberta-base-japanese
6)　https://huggingface.co/nlp-waseda/roberta-large-japanese
7)　https://huggingface.co/ku-nlp/deberta-v2-base-japanese
8)　https://huggingface.co/ku-nlp/deberta-v2-large-japanese