競技クイズにおける LLM と人間の誤答傾向の分析と比較

杉浦 尚弥 [1]　小川 泰弘 [2]　外山 勝彦 [1]　山田 康輔 [1]　笹野 遼平 [1]
[1] 名古屋大学　[2] 名古屋市立大学
sugiura.naoya.e7@s.mail.nagoya-u.ac.jp, kosyamada0526@gmail.com
{sasano, toyama}@i.nagoya-u.ac.jp, ogawa@ds.nagoya-cu.ac.jp

概要

本稿では，競技クイズを題材として，問題文や正解[1)]の特性が LLM と人間の解答の正答率にもたらす影響について調査し，人間と LLM における「難しさ」の違いを明らかにする．人間の正答率が付与されたクイズデータを収集し，クイズの正解の Wikipedia エントリが存在するか，正解の文字種が何であるか，問題文中で解答候補が列挙されているか，という 3 つの観点で分類する．GPT-4o など 6 つの LLM にクイズを解答させ，各観点による分類ごとの LLM と人間の正答率の差を明らかにする．

1　はじめに

大規模言語モデル（LLM）は質問応答など様々な NLP タスクにおいて，人間と同等以上の性能を示している．クイズに対しても，競技クイズのコンペティションである AI 王[2)]において 9 割以上の正答率を達成するシステムが登場するなど，人間を超える性能を達成している．しかし，LLM と人間の誤答傾向は異なっており，人間が正答しやすい容易な問題であっても，LLM が誤答するケースが散見される．LLM の誤答例を表 1 に示す．この問題は本研究で収集した人間の正答率付きクイズデータの中の 1 問である．人間の正答率は 7 割を超えているが，GPT-4o は全体としては非常に高性能であるにもかかわらず，「サイ」という誤った解答を出力する．このような現象はしばしば確認され，人間にとっての「難しさ」は LLM のそれとは異なる可能性が示唆される．これを明らかにすることは，LLM の特性を理解する手がかりとなる．

先行研究では，問題形式が LLM の推論に及ぼす影響 [1, 2] や，クイズの解答時に人間がどのように思考しているか [3] など，様々な認知的側面から分析されている．これに対して本研究では，人間と LLM の誤答傾向を比較・分析することにより，クイズにおける「人間にとって難しい問題」と「LLM にとって難しい問題」の違いを明らかにする．

表 1　人間には容易だが GPT-4o が誤答する例

問題文: 元々は賭博用語で「ゾロ目」を意味する言葉で，今では同い年を指すときに用いられるカタカナ 2 文字の言葉は何？
人間の正答率: **0.73**　正解: タメ　解答: **サイ**

2　競技クイズの特性とデータセット

本節では，競技クイズの自動解答に関する既存研究を紹介した後，本研究で使用する人間の正答率付きクイズデータについて述べる．

2.1　自然言語処理における競技クイズ

競技クイズはオープンドメイン質問応答 (ODQA) タスクとして研究が行われており，LLM を用いたシステム [4] や，ベクトル検索 [5] と生成を複合したシステム [6] などが開発されている．競技クイズは ODQA タスクの中でも，質問文が短文で一問一答形式という特徴を持つタスクである．しかし，「ですが問題」といった特殊な出題形式を取る問題も存在することが知られており，最近では，「ですが問題」を題材として LLM に対し認知的な分析を行う研究 [2] や，特殊な出題形式を持つ問題を自動作成する研究が行われている [7, 8, 9]．

早押しクイズに焦点を当てた研究も存在しており，不完全文に対し，LLM を用いて解答する研究 [10] や，情報が不十分な中で判断を下すという視点から人間の判断や方略変化などを分析する研究が存在する [3]．しかし，これらの研究はクイズの自動解答や，人間によるクイズの解答プロセスの分析を目的としており，人間と LLM を比較していない．

1)　本稿では，クイズの答えのことは「正解」と呼び，「正答」は人間や LLM の解答が正しかった場合を指す語として用いる．
2)　https://sites.google.com/view/project-aio/home

2.2 人間の正答率付きクイズデータ

本研究では，人間と LLM の誤答傾向を比較するため，人間の正答率付きクイズデータを利用する．そのため，クイズアプリ「みんなで早押しクイズ」[3] で出題された過去問と正答率を収集した[4][5]．「みんなで早押しクイズ」では，プレイヤー自身が問題を投稿することができ，運営側が問題の質を確認した上で採用しているため，多岐にわたる良質なクイズが出題されている．このアプリでは，各問題に対するプレイヤーの正答率 (解答者数に対する正答者の割合) を確認できる．ただし，プレイヤーは基本的に早押しで解答をしているため，この正答率は問題全文が与えられた場合に比べて低い値になっていると考えられる．

3 分析・比較手法

本節では，LLM の解答データの収集とその評価方法，および分析・比較の観点について説明する．

3.1 LLM による自動解答

本研究では，LLM によるクイズ解答データを収集するため，6 つの LLM にクイズを解答させる．具体的には，英語を中心に事前学習を行ったモデルである GPT-4o，日本語で継続事前学習を行ったモデルである Swallow 70B，日本語を中心に事前学習を行ったモデルである Sarashina2 70B，また，それぞれの軽量版である GPT-4o mini，Swallow 13B，Sarashina2 13B の計 6 モデルを使用する．各モデルには，few-shot 学習，QLoRA[11] による fine-tuning の 2 種類による追加学習を実施し，正答率が最も高い学習法を採用する．具体的には，GPT-4o は 5-shot 学習，Swallow は fine-tuning により学習する．Sarashina2 については，追加学習によって正答率が低下したため，追加学習せずに使用する．

2.2 節で述べたように，人間の正答率は早押しが可能な設定で算出されている．問題文の途中で回答した場合の性能を調査するため，最も正答率が高かった GPT-4o については，問題文の先頭から 25%, 50%, 75%のみを与える設定でも実験を行う．

3.2 正誤判定

モデル出力の正誤は以下の手順で判定する．

手順 1 モデル出力と正解から「(，)，[，]，・，=」の 6 種類の記号を機械的に取り除き，文字列が一致した場合は正答と判定

手順 2 手順 1 で正答と判定されなかった場合，GPT-4o を用いて作成した表記揺れ訂正器 (付録 B 参照) に入力し，正解の表記揺れと判定された場合は正答と判定

ただし，表記揺れ訂正器の影響を調査するため，手順 1 のみ行った場合の正答率も算出する．

3.3 分析・比較の観点

本稿では，正解の Wikipedia エントリの有無，正解の文字種，問題文中における解答候補の列挙という 3 つの観点でクイズを分類し，LLM と人間の誤答傾向の分析を行う．

Wikipedia エントリの有無 LLM は膨大なウェブデータを学習資源としており，その中でも Wikipedia は知識源として大きな存在である．このため，Wikipedia にエントリが存在する単語が正解となる問題は，LLM にとって正答しやすい問題である可能性が高い．一方，人間が Wikipedia から得た知識の割合は限定的であり，正解が Wikpedia のエントリとして存在するかどうかは人間の正答率に大きくは影響しないと考えられる．本分析では，2024 年 10 月 1 日における日本語 Wikipedia のダンプファイルから Wikipedia エントリのタイトルをすべて取得し，前節で述べた 6 種類の記号を取り除く処理を行った上でタイトルと正解を比較することで，正解の Wikipedia エントリが存在するかを判定する．また，Wikipedia エントリが存在した場合において，それを説明する記事が直接存在するパターンと，他の記事に転送させるパターンが存在するため，この 2 種は区別して分類する．

正解の文字種 LLM は事前学習で用いられているテキストデータの主言語がモデルごとに異なるため，学習される知識の傾向も異なる可能性がある．例えば，英語を中心に学習されたモデルは，英語圏の知識を豊富に持つ一方で，他の言語圏の知識は乏しい可能性がある．競技クイズは言語圏を限定しない広範な知識を問うが，日本語のクイズでは多くの場合，外来語はカタカナ，和語はひらがなや漢字で表記するなど，言葉の由来と使用される文字種には関連性がある．また，クイズの中には正解が数字であるものも存在するが，これは算術や数え上げなど他タスクとの複合問題であることが多く，LLM に

3) https://minhaya.com/
4) https://mininome.com/
5) https://raityo.com/

とって難しい問題である可能性がある．このように，正解の文字種は LLM の正答率に影響を及ぼすことが考えられることから，本稿では，正規表現を用いて，正解の文字種が数字，アルファベット，カタカナ，ひらがな・漢字の 4 種類のいずれに該当するか判定し，その結果に基づき問題を分類する．

解答候補の列挙の有無　競技クイズには，解答の存在範囲を明示し，解答候補を列挙することで残りの 1 つを答えさせる出題形式 (列挙型問題) がある．例えば，「世界三大珍味とは，キャビア，トリュフと何？」という問題では，「世界三大珍味」という範囲と，「キャビア」「トリュフ」という候補が提示され，最後の候補「フォアグラ」を導かせる．LLM にとって，列挙された解答候補は正解を想起させるヒントとなるため，通常の問題より正答率が高くなる可能性が考えられる．一方，列挙型問題は解答候補が揃うまで正解が確定しないため，早押しクイズのように問題文の途中で解答すべき状況では，正答率が低下する可能性がある．そのため，人間の正答率や早押し設定の GPT-4o の正答率は，列挙型問題において減少すると考えられる．本分析では，few-shot で学習した GPT-4o を用いて分類する．

4　分析結果と考察

本節では，各モデルの正答率や分類の観点ごとの分析結果を示し，それらを考察する．また，観点ごとの結果である表 2，表 3，表 4 において，見出し行の括弧内の値は該当する分類のデータ数を表し，他の括弧内の値は表記揺れ訂正前の正答率 (%) を表す．Human は人間の平均正答率 (%) を示す．また，GPT-4o の早押し設定の結果は，問題文の先頭 25% を与えた場合，GPT-4o (25%) のように表記する．

4.1　正答率

表 2 の最右列に各モデルの全体としての正答率，および人間の平均正答率を示す．モデルごとに比較すると，GPT-4o が最も高い正答率となり，Sarashina2 70B，Swallow 70B が続いた．GPT-4o は早押し設定においても高い正答率となり，問題文の 50%を与えた場合には，全文を与えた Swallow 13B に匹敵するスコアを，25%のみを与えた場合においても，人間の平均正答率を超えるスコアを示した．より詳細な人間との比較は付録 A に掲載する．

表 2　Wikipedia エントリの有無による分類ごとの正答率

モデル	記事有 (3,042)	転送有 (486)	存在無 (440)	全体 (3,968)
Swallow 13B	84.4 (78.0)	63.2 (40.1)	60.3 (26.3)	79.1 (67.6)
Swallow 70B	91.7 (84.6)	72.6 (42.6)	70.7 (33.1)	87.0 (73.7)
Sarashina2 13B	87.5 (79.7)	82.1 (62.1)	68.9 (44.2)	84.7 (73.6)
Sarashina2 70B	93.8 (84.7)	88.9 (62.8)	81.6 (48.8)	91.8 (78.0)
GPT-4o mini	74.8 (66.3)	68.1 (53.1)	54.9 (30.8)	71.7 (60.7)
GPT-4o	96.6 (88.7)	92.2 (71.8)	87.3 (61.5)	95.0 (83.6)
GPT-4o (25%)	47.0 (41.6)	42.6 (27.2)	26.3 (13.2)	44.2 (36.6)
GPT-4o (50%)	78.1 (69.1)	75.5 (53.9)	61.7 (38.8)	76.0 (63.8)
GPT-4o (75%)	93.9 (85.2)	89.3 (66.9)	82.3 (54.2)	92.0 (83.6)
Human	41.9	41.1	40.5	41.6

4.2　Wikipedia エントリの有無による影響

表 2 に Wikipedia エントリの有無に基づく分類ごとの正答率を示す．すべてのモデルで共通して，記事有，転送有，存在無の順で正答率が低下し，その差は性能の低いモデルであるほど顕著であった．これに対し，人間の正答率の差は 1 ポイント程度であり，モデルの正答率の差に比べて非常に小さかった．これは 3.3 節で示した予想の通り，LLM の事前学習において Wikipedia が重要な知識源となっている可能性を示唆している．ただし，Wikipedia エントリが存在する単語は，比較的使われやすい単語が多く，存在しない単語に比べて学習コーパス中の出現頻度が高く，正答率が高まった可能性も考えられる．

次に，表記揺れ訂正前後の正答率の上昇幅を比較すると，正解の Wikipedia エントリが存在しない問題は，存在する問題に比べて 2 倍以上上昇しており，表記揺れが非常に多いことが分かる．Wikipedia エントリが存在する問題において表記揺れが少ないのは，Wikipedia エントリが存在する単語は，代表的な表記が定まっていることが多く，さらにそれが広く普及しているかな漢字変換や，予測変換のシステム辞書に登録されていることが多いためであると考えられる．

4.3　正解の文字種による影響

表 3 に正解の文字種による分類ごとの正答率を示す．すべてのモデル・条件において，正解が数字，ひらがな・漢字である問題の正答率は，正解がアルファベット，カタカナである問題の正答率と比べ低い値となった．次に，人間の平均正答率と比較すると，正解がアルファベットと数字の場合は LLM と同じ傾向が確認できたが，カタカナとひらがな・漢

表 3 正解の文字種による分類ごとの正答率

モデル	数字 (141)	アルファベット (153)	カタカナ (1,835)	ひらがな・漢字 (1,838)
Swallow 13B	68.8 (35.5)	57.5 (45.8)	81.9 (71.4)	79.1 (68.3)
Swallow 70B	78.0 (43.3)	69.3 (51.0)	90.8 (78.9)	85.5 (73.0)
Sarashina2 13B	68.1 (44.7)	91.5 (73.8)	85.9 (75.4)	74.1 (84.4)
Sarashina2 70B	83.7 (49.6)	93.5 (69.9)	93.7 (80.6)	90.6 (78.5)
GPT-4o mini	76.6 (44.7)	86.9 (73.2)	78.3 (65.3)	63.8 (56.5)
GPT-4o	92.2 (67.4)	98.7 (81.0)	96.6 (84.7)	93.4 (84.1)
GPT-4o (25%)	18.4 (11.3)	52.9 (40.5)	49.8 (40.9)	39.9 (34.1)
GPT-4o (50%)	66.7 (41.8)	83.0 (65.4)	81.0 (68.1)	71.2 (61.2)
GPT-4o (75%)	92.2 (65.2)	95.4 (75.2)	94.6 (81.5)	89.2 (79.1)
Human	39.7	45.6	40.1	42.9

字の場合においては逆の傾向が確認された．

まず，正解が数字である場合の傾向について考察する．このような問題の多くは「都道府県の中で，海に面していないのはいくつある？」のように，数え上げを求めるものである．これを解く際には，47個の都道府県すべてに対し，海に面しているかの知識を有する必要がある．求められる知識量が大幅に増加するため，LLM，人間ともに正答率が低下したと考えられる．加えて，LLM においては，答えが数であるとき，すべての数が解答候補となるため，他の場合よりも難しい問題となると考えられる．

次に，アルファベット，カタカナ，ひらがな・漢字間の傾向について考察する．正解がカタカナ，アルファベットのとき，そのほとんどは外来語であり，ひらがな・漢字のときは和語である．そのため，事前学習に使用した言語によって正答率が変化すると思われたが，その傾向は見られなかった．これは，LLM が日本語学習をする際，知識が学習されにくい可能性を示している．これに対し，人間の回答者のほとんどは日本語話者であるため，人間の平均正答率ではひらがな・漢字がカタカナを上回ったと考えられる．アルファベットについては，略称が正解となっている問題が多く，正解を推測できる場合が多いため，正答率が他に比べて高くなったと考えられる．

4.4 解答候補の列挙の有無による影響

表 4 に，列挙の有無による分類ごとの正答率を示す．予想に反して，列挙の有無は正答率に影響をほとんど及ぼさなかった．これは，列挙によって LLM の内部状態は変化したものの，正しいトークンの出力確率のみが上昇するわけではなく，誤ったトークンの出力確率も高まったからだと考えられる．早押しクイズの状況下においては，解答のタイミングが早いほど正答率が低下した．これは，解答候補の列挙が問題文の後半で行われるため，正解が確定しない状態での解答が多くなったことが原因だと考えられる．しかし，25%や 50%の早押し設定では，候補がすべて列挙される前に解答を行うパターンが多いものの，正答できている例が複数確認された．この結果は，LLM は人間と同じく，問われやすい単語を推測している可能性を示唆している．具体的な正答例は付録 C に示す．

表 4 解答候補の列挙の有無による分類ごとの正答率

モデル	列挙有 (158)	列挙無 (3,810)
Swallow 13B	78.5 (72.2)	79.2 (67.5)
Swallow 70B	83.5 (75.9)	87.2 (73.7)
Sarashina2 13B	85.4 (78.5)	84.8 (73.4)
Sarashina2 70B	89.9 (81.0)	91.9 (78.0)
GPT-4o mini	72.2 (63.9)	71.8 (60.6)
GPT-4o	94.9 (88.0)	95.0 (83.5)
GPT-4o (25%)	27.2 (15.2)	44.9 (37.5)
GPT-4o (50%)	68.4 (40.5)	76.3 (64.8)
GPT-4o (75%)	89.2 (79.1)	92.2 (79.5)
Human	40.2	41.7

5 おわりに

本研究では，競技クイズにおける人間と LLM の誤答傾向の違いについて分析・比較し，人間と LLM における「難しさ」の違いについて調査した．Wikipedia に関する分析の結果，Wikipedia エントリが存在しない単語を問う問題は，存在する単語を問う問題に比べて LLM の正答率は大きく低下したが，人間の正答率には大きく影響しなかった．この結果から，LLM の知識源として Wikipedia の存在が大きいと考えられる．正解の文字種による分析では，解答がひらがな・漢字の場合の正答率は相対的に低く，LLM の事前学習において，日本語の知識が獲得されにくい可能性が示唆された．解答候補の列挙の有無による分析では，早押しにおける列挙の影響は大きいものの，GPT-4o は列挙される前から問われる単語を予想できており，人間と同等の推論を行っている可能性が示された．本研究では，人間の正答率付きデータを収集することで比較を行ったが，算出におけるプレイヤーの数や実力が不明であるため，指標として不完全な部分が存在する．より精密な比較のためには，被験者実験を行い，実際に正答率を算出する必要があり，これは今後の課題である．

参考文献

[1] 杉山宏輝, 角康之. 早押しクイズの名数問題における解の妥当性を考慮した解答を するための cot プロンプトの構築. 言語処理学会第 30 回年次大会 (NLP2024), pp. 832–837, 2024.

[2] 山下陽一郎, 原田宥都, 大関洋平. 人間らしい予測処理機構を取り入れた質問応答モデルの提案： 早押しクイズのパラレル問題を題材として. 言語処理学会第 29 回年次大会 (NLP2023), pp. 2891–2896, 2023.

[3] 白砂大, 小坂健太. 早押しクイズに見る不確実性下の判断：クイズ大会の行動データに基づく事例研究. 認知科学, Vol. 31, No. 2, pp. 352–361, 2024.

[4] 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, 2020.

[5] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wentau Yih. Dense passage retrieval for open-domain question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing**, pp. 6769–6781, 2020.

[6] Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics**, pp. 874–880, 2021.

[7] 島田克行, 折原良平, 森岡慎太, 市川尚志. 嘘がなく、面白いクイズの自動生成. 言語処理学会第 30 回年次大会 (NLP2024), pp. 436–440, 2024.

[8] 小林俊介, 河原大輔. 多様なクイズを自動生成する手法およびその検証. 言語処理学会第 30 回年次大会 (NLP2024), pp. 24–29, 2024.

[9] 橋元佐知, 佐藤理史, 宮田玲, 小川浩平. 早押しクイズのパラレル問題の自動生成. 言語処理学会第 28 回年次大会 (NLP2022), pp. 832–837, 2022.

[10] Naoya Sugiura, Kosuke Yamada, Ryohei Sasano, Koichi Takeda, and Katsuhiko Toyama. Building a buzzer-quiz answering system. In Vishakh Padmakumar, Gisela Vallejo, and Yao Fu, editors, **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics**, pp. 194–199, 2023.

[11] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized llms. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, **Advances in Neural Information Processing Systems**, Vol. 36, pp. 10088–10115, 2023.

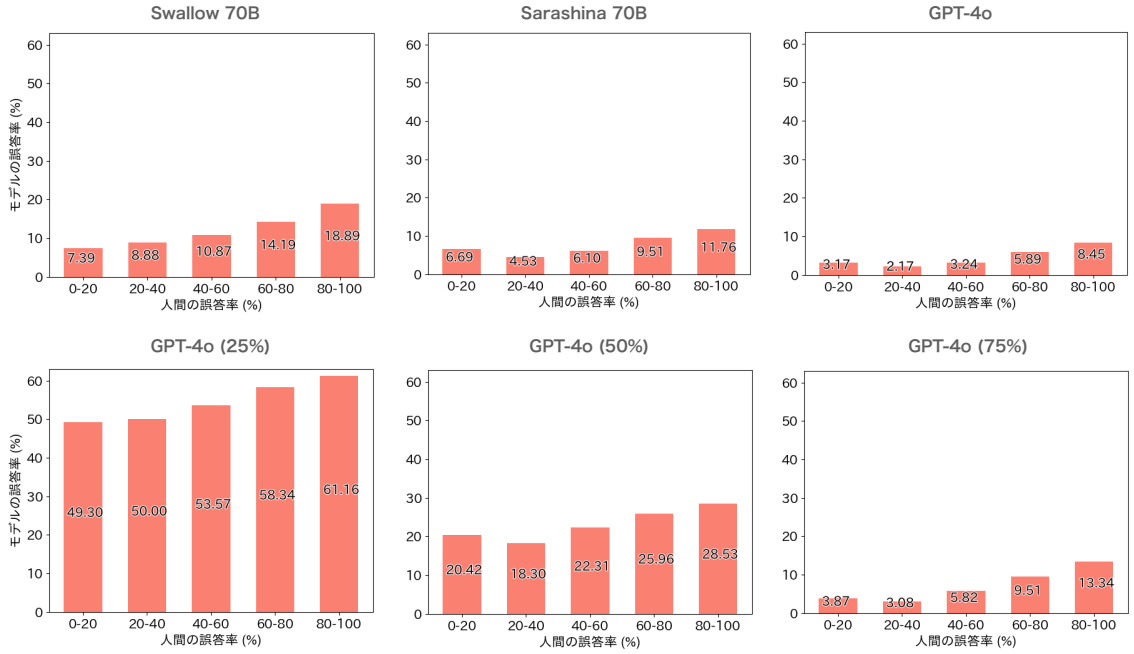

図 1　人間の誤答率に対するモデルの誤答率

A　人間の誤答率とモデルの誤答率

図 1 に人間の誤答率に対するモデルの誤答率のグラフを掲載する．まず，全文入力設定においては，基本的に人間の誤答率とモデルの誤答率には正の相関が確認された．しかし，Sarashina 70B, GPT-4o といった，全体での正答率が 9 割を超えているモデルにおいては，人間の誤答率が 20-40%の問題に対して不自然な正答率の落ち込みが見られた．早押し設定においても，全体的に正の相関が見られたが，全文入力設定に比べて相関は小さい結果となった．

B　表記揺れ訂正器

表記揺れ訂正器は，問題文と正解，モデル解答を入力し，表記揺れによって不当に誤答とされていると判断した場合は True，単なる誤答だと判断した場合 False を返す二値分類器である．GPT-4o を用いて，few-shot 学習を行い構築した．shot 数に関しては，人手で作成されたテストデータを用いてテストを行い，最良であった 6-shot とした．6-shot におけるテスト結果の混同行列を表 5 に，出力例を表 6 に掲載する．

C　列挙における早押し正答例

表 7 に，列挙型問題においてすべての候補が列挙される前にモデルが正答した例を掲載する．黒字部分が実際に入力された問題文を示す．

表 5　6-shot における表記揺れ訂正器のテスト結果

		Actual	
		True	False
Predict	True	228	6
	False	1	265

表 6　表記揺れ訂正器の出力例

問題文: 慣用句で，不測の事態に備えて必要な準備をしておくことを，「転ばぬ先の何」という？
正解: 杖　**モデル解答:** 転ばぬ先の杖
訂正器出力: True

問題文: 自動車の給油所を指す「SS」とは，何という英語の略でしょう？
正解: サービス・ステーション　**モデル解答:** Service Station
訂正器出力: True

表 7　列挙が行われる前にモデルが正答した例

問題文:1949 年に起きたいわゆる「国鉄三大事件」に数えられる 3 つの事件とは，7 月 6 日に起きた下山事件，7 月 15 日に起きた三鷹事件と，8 月 17 日に起きた何事件でしょう？
正解: 松川事件　解答: 松川事件

問題文: テニスや卓球の試合形式を大きく 2 つに分けると，1 人対 1 人の「シングルス」と，2 人対 2 人で行う何でしょう？
正解: ダブルス　解答: ダブルス

問題文: 日本の競馬における 4 つの馬場状態とは，「良」，「稍重」，「重」，もう一つは何でしょう？
正解: 不良　解答: 不良

問題文: 日本における建築士の資格を大きく 3 つに分けると，一級建築士，二級建築士ともう一つは何でしょう？
正解: 木造建築士　解答: 木造建築士