LLM は ASD 小児と定型発達小児が作成したストーリーを識別できるか？

河野真有香 [1]　平尾悠太朗 [1]　ペルスキアエルナンデスモニカ [1]
内山英昭 [1]　上垣外英剛 [1]　清川清 [1]
[1] 奈良先端科学技術大学院大学
{kono.mayuka.ki4}@is.naist.jp

概要

ASD の小児の言語や事物の認識を理解することは，彼らに対する支援の適切性の評価に重要である．当事者やその家族を対象とした研究や生物学的研究のほかに，工学的観点からヒトの模倣可能性が注目されている LLM を活用した研究がある．我々はこれらを組み合わせて ASD の小児における言語認知メカニズムの解明を目指す．そのために，LLM に対して ASD の小児のペルソナを与え，LLM に ASD 様の振る舞いを再現させることを目指す．本稿では，LLM に ASD の小児と定型発達の小児のストーリーを識別する能力があるかを調査した結果を報告する．ASD の小児が作ったストーリーを回答するという 5 択の QA タスクにより，現状の LLM による識別精度は 22%であることを明らかにした．

1　はじめに

2022 年現在，世界の小児のうち 1/100 が ASD（Autism Spectrum Disorder，自閉スペクトラム症）と診断されている [1]．ASD は，社会的コミュニケーション障害と反復的な感覚運動行動を特徴とする [2]．早期発見と早期介入が，長期的な成果の向上に寄与する [3, 4]．ASD の症例ではコミュニケーションの困難が顕著であることから，言語スキルに焦点を当てた介入が行われることが一般的である [5]．また，小児の家族や教師など，直接的に関わる人々の支援の仕方が，介入の成否に大きく影響を与える [2]．しかし，ASD の小児がどのように言語や事物を認識しているかを理解することは容易ではなく，支援の適切性を評価するための課題である．

従来，ASD に関する研究は当事者やその家族を対象とする研究や生物学的研究を中心に行われてきた [6, 7, 8]．しかし，ヒトや生物を対象とするため，

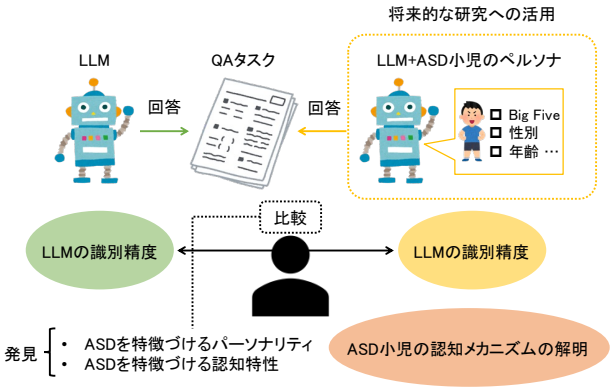

図 1　LLM に ASD 小児のペルソナを与える研究の概要．

人的コストや倫理的制約などの課題が存在する．一方、急速に進歩を遂げている LLM（Large Language Model、大規模言語モデル）は、工学的観点からヒトの模倣可能性が注目を集めている [9]．特に，LLM にペルソナを付与することで，人間の行動や振る舞いを模倣する試みが進められており，この手法がさまざまな応用につながる可能性がある．LLM はヒトや生物とは異なるものの，人的コストや倫理的制約が伴わない点で優位性を持つ．実際，ASD とは異なるものの，LLM のパラメータを削減することで，統合失調症の言語・思考障害をシミュレーションした研究がある [10]．

我々はこの研究に着想を得て，従来の当事者を対象とする手法や生物学的手法と LLM を活用した手法を組み合わせることで，ASD の小児における言語認知メカニズムの解明を目指す．言語認知メカニズムが明らかになれば，ASD の小児と親や教師，また ASD 以外の小児とのコミュニケーションの向上に寄与できる．さらに，ASD の小児の学習支援デバイスや，ASD の小児との関わり方を助言するチャットボットの開発などへ応用が可能である．

先行研究 [10] のように，LLM のネットワーク構造から探索する手法は，プロンプトエンジニアリン

グと比較して透明性が高い反面，計算資源コストが高い．一方，LLM にプロンプトエンジニアリングでパーソナリティを持たせる研究が活発に行われている [11]．我々はこの研究に着想を得て，LLM に対してプロンプトエンジニアリングを用いて ASD の小児のペルソナを与えることで，LLM が ASD 様の振る舞いを再現することを目指している (図 1)．

LLM が ASD の小児に関する知識を既に有している場合にはペルソナの有無に関わらず ASD 様の振る舞いを再現する可能性がある．よって，LLM がペルソナを与える以前に，どの程度 ASD の小児に関する知識を持っているかを調べることは，ペルソナを与えた場合の効果を測定する上で重要である．本稿では，第一歩として，LLM に ASD の小児と定型発達の小児のストーリーを識別する能力があるかを調査する．ASD の小児と定型発達の小児が 6 コマのイラストを見てストーリーを生成するタスクに着目し，5 択の QA データセットを構築して LLM に ASD の小児が作ったストーリーを回答させることで調査した．その結果，現状の LLM における ASD の小児と定型発達の小児のストーリーの識別精度は 22%であることを明らかにした．

本研究の成果は以下の通りである．

- LLM が ASD の小児と定型発達の小児のストーリーを識別する能力を持つ可能性を調査した．
- ASD の専門家が不在の環境でも，定量的に評価可能な QA タスクによる実験手法を提案した．
- 現状の LLM では識別精度は 22%であることを明らかにした．

2 関連研究

2.1 LLM にペルソナを持たせる手法

Jiang らは，GPT-3.5 および GPT-4 に Big Five に基づくパーソナリティを付与する PersonaLLM を提案した [12]．その結果，パーソナリティに適した言語学的特徴を持つストーリーを生成可能であることを示した．さらに，人間とペルソナを持たせた LLM の間には，文章作成時の単語使用に共通点があることも明らかにした．Tu らは，ChatGPT を用いて MBTI (Myers-Briggs Type Indicator) に基づくペルソナ分類を活用し，個人とペルソナが適合するバーチャル対話エージェントをマッチングする CharacterChat を提案した [13]．また，様々なプロファイルを持つバーチャルキャラクターで構成された MBTI-1024 Bank を構築した．

2.2 ASD とパーソナリティの関連

Fortenberry らは，FFM (Five-Factor Model of personality) を改変して作成した質問紙を用い，ASD の小児と定型発達の小児の間でパーソナリティに差があるかを調査した [14]．ASD の小児と定型発達の小児を比較した結果，ASD の小児では Extraversion (外向性)，Conscientiousness (誠実性)，Open to Experience (開放性) のスコアが低いことを示した．

LLM にペルソナを持たせる研究では，成人を対象としたものが多い上，定型発達と ASD の区別は考慮されていない．本研究では，小児に焦点を当て，ASD と定型発達の 2 点に着目する．ASD の小児と定型発達の小児の間で Big Five の傾向に差があることから，Big Five と追加情報をペルソナとして LLM に与えることで，ASD の小児のペルソナを持たせることが可能であると期待される．

LLM 固有の性格特性が存在し，その傾向はモデルごとに異なる [15]．しかし，先行研究では，LLM 固有の性格特性とペルソナ付与後の差分は明確に示されていない．本研究では，その差分を明らかにするために， LLM 自体に ASD の小児と定型発達の小児のストーリーを識別する能力がどの程度あるかを調査する．

3 QA データセット構築

本章では，実験で使用する QA データセットの構築方法を述べる．ASD の専門家が不在の環境でも，定量的に評価可能な手法として，QA タスクによる実験手法を提案する．

3.1 使用したコーパス

コーパスは CHILDES[1)]から入手した．ASD の小児のコーパスは ASDBank Dutch Assymmetries Corpus [16, 17] を使用した．定型発達の小児のコーパスは CHILDES Dutch Asymmetries Corpus [16, 17] の SK-TD を使用した．データ数は ASD の小児に関するデータが 46 名分，定型発達の小児に関するデータが 38 名分である．言語はオランダ語で，いずれも 6–12 歳の小児が 6 コマのイラストを見て作成したストーリーからなる．6 コマのイラストの種類は 4 種類あるが，本研究では欠損データのない

1) https://childes.talkbank.org/

indiaan2 のみを対象とした．

3.2 構築環境

開発環境は Google Colaboratory[2] を使用した．コーパスから必要な自然言語文のみを抽出するために PyLangAcq [18] を使用した．

3.3 QA データセット

JGLUE [19, 20, 21] のうち，JCommonsenseQA を参考に 5 択形式の QA を作成した．QA は 6 コマのイラストである indiaan2 と「ASD の小児が話したストーリーはどれですか？」という意味のオランダ語の質問，選択肢，解答からなる．画像は GPT-4o Nov 20 2024 version で使用可能な.png 形式に変換した．質問文の違いにより結果が変動する可能性があるため，同じ意味で表現の異なる質問を用意し，全部で 3 種類の QA データセットを作成した．質問文の翻訳・生成は，付録 A に示す．これにより，以下 3 つの質問文を得た．

1. Welke verhaal werd verteld door een kind met ASS?
2. Welk verhaal is door een kind met ASS verteld?
3. Door welk kind met ASS werd een verhaal verteld?

各データセットの選択肢は，コーパスから無作為・重複なしで抽出した 1 つの ASD の小児のストーリーと 4 つの定型発達の小児のストーリーからなる．よって，QA データセットの QA 数は 9 個である．また，多肢選択問題における選択肢の並び順は LLM の性能に影響する [22] ため，解答の順番がランダムになるよう選択肢を配置した．

4 実験

本節では LLM に ASD の小児と定型発達の小児のストーリーを識別する能力があるかを調べる実験手法を述べる．3 章で構築した QA データセットを用いて，LLM の識別精度を評価した．

4.1 実験環境

開発環境は Google Colaboratory を使用した．評価対象のモデルは OpenAI API[3] の GPT-4o Nov 20 2024 version とした．

4.2 LLM の識別精度の評価

6 つのイラストと QA データセットの選択肢までを入力として，GPT-4o Nov 20 2024 version に回答させ，正答率を算出した．*temperature* = 1.0(既定値) とした．また，temperature により毎回出力結果が変わる [23] ため，各 QA につき 10 回ずつ試行し，最多となった回答をその QA の回答として採用した．実験にて使用したコードのプロンプト部分は付録 B に示す．各質問文における正答率と平均正答率を算出した．正答率は式 (1)，平均正答率は式 (2) により算出した．両者を 5 択 QA のチャンスレートである 0.20 と比較して識別精度を評価した．

$$\text{各質問文における正答率} = \frac{\text{正答数}}{QA\ \text{数}} \tag{1}$$

$$\text{平均正答率} = \frac{\sum(\ \text{各質問文の正答率}\ )}{\text{質問数}} \tag{2}$$

4.3 結果

各質問文における正答率を表 1 に示す．各質問文の場合の対象モデルの識別精度はそれぞれ，0.11，0.33，0.22 であることが分かった．また，各質問文の正答率は，質問文 2> 質問文 3> 質問文 1 であった．これより，質問文によって正答率に違いが生じることが分かった．平均正答率は 0.22 であった．

5 考察

5.1 LLM の識別精度の評価

各質問文における正答率 (表 1) と平均正答率 (平均正答率 = 0.22) を 5 択 QA のチャンスレートである 0.20 と比較して識別精度を評価した．チャンスレートと比べて質問文 1 では 9%低く，質問文 2 と質問文 3 ではそれぞれ 2%，13%高かった．平均正答率では 2%高かった．以上より，対象モデルが ASD の知識を持っている可能性がある．しかし，その識別精度は 0.22 程度であり，少なくとも対象モデルの持つ ASD の知識と QA タスクで与えた {6 コマの画像，選択肢のストーリー } の対を紐づけて識別することが難しいといえる．実際，臨床現場においても言語学的特徴だけでなく，質問表などのツールを活用 [24, 25, 26] しながら多面的に評価・診断が行われているため，現状の対象モデルの知識と実験で与えた情報だけでは識別が難しいと示唆される．

2) https://colab.research.google.com/?hl=ja
3) https://openai.com/api/

表 1　各質問文における正答率.

	質問文	正答率
質問文 1	Welke verhaal werd verteld door een kind met ASS?	0.11
質問文 2	Welk verhaal is door een kind met ASS verteld?	0.33
質問文 3	Door welk kind met ASS werd een verhaal verteld?	0.22

実験を通じて，質問文によって識別精度が異なることも明らかになった．特に質問文 1 と質問文 2 の識別精度の間には 22%の差があった．このことから，QA タスクを設計する際には同じ意味の別表現の質問文も用意することが，適正な評価のために重要であるといえる．

5.2　今後の展望

5 択問題を作成するには，定型発達の小児のサンプル数が不足していたため，QA データセットの QA 数は 9 個と少なかった．Dubois らの研究 [27] から人間のフィードバックと LLM が生成したフィードバックはほぼ等価なものとみなせる．よって，今後の展望として定型発達の小児のダミーデータを対象モデルとは別の LLM に対するプロンプトエンジニアリングで生成し，QA 数を増やすことがある．

対象モデルの temperature について，$temperature = 1.0$(既定値) のみで実験を行った．しかし，temperature の値によって性能に違いがある [23] ため，今後は複数の temperature で実験を行う必要があるといえる．これは，先述のダミーデータを生成する際にも考慮すべきことである．

本稿では，LLM のみを対象に識別精度を評価しており，人間の識別精度は評価していない．元々人間でどの程度識別可能な問題設定であるのかを評価することは，今後の研究方向性を考える上でも重要である．今後，オランダ語を母語とする人間を対象に ASD の小児が作ったストーリーを回答させるタスクを課し，識別精度を評価する．

今後，LLM に対して ASD 小児のペルソナを与えることで，LLM の ASD の小児と定型発達の小児のストーリーの識別精度がどの程度変化するかを評価する．

実験では GPT-4o Nov 20 2024 version のみを対象とした．しかし，実際には多くのモデルがある[4)] ことから対象モデルの種類を増やして比較することも必要と考える．

4)　https://tatsu-lab.github.io/alpaca_eval/

6　おわりに

本稿では，5 択 QA タスクにより LLM に ASD の小児と定型発達の小児のストーリーを識別する能力があるかを調査した．実験により，対象モデルでは識別精度が 22%であることが明らかになった．今後は，QA 数を増やし，複数の temperature で実験を行うとともに，LLM にペルソナを付与することで識別精度がどの程度変化するかを評価する．さらに，ペルソナを与えた LLM にストーリー生成タスクを課す．QA データセットの ASD の小児の選択肢を生成したストーリーで置換して，自身の考えに近い選択肢を回答するタスクを解かせる．これにより，ペルソナを与えた LLM が ASD の小児のように振る舞うかを明らかにする．

謝辞

本研究は JST 次世代研究者挑戦的研究プログラム JPMJBS2423 の支援を受けたものです．

参考文献

[1] Jinan Zeidan, Eric Fombonne, Julie Scorah, Alaa Ibrahim, Maureen S Durkin, Shekhar Saxena, Afiqah Yusuf, Andy Shih, and Mayada Elsabbagh. Global prevalence of autism: A systematic review update. **Autism research**, Vol. 15, No. 5, pp. 778–790, 2022.

[2] Catherine Lord, Mayada Elsabbagh, Gillian Baird, and Jeremy Veenstra-Vanderweele. Autism spectrum disorder. **The lancet**, Vol. 392, No. 10146, pp. 508–520, 2018.

[3] Naila Z Khan, Lilia Albores Gallo, Aurora Arghir, Bogdan Budisteanu, Magdalena Budisteanu, Iuliana Dobrescu, Kirsty Donald, Samia El-Tabari, Michelle Hoogenhout, Fidelie Kalambayi, et al. Autism and the grand challenges in global mental health. **Autism Research: Official Journal of the International Society for Autism Research**, Vol. 5, No. 3, pp. 156–159, 2012.

[4] Julia Parish-Morris, Christopher Cieri, Mark Liberman, Leila Bateman, Emily Ferguson, and Robert T Schultz. Building language resources for exploring autism spectrum disorders. In **LREC... International Conference on Language Resources & Evaluation:[proceedings]. International Conference on Language Resources and Evaluation**, Vol. 2016, p. 2100. NIH Public Access, 2016.

[5] Duanchen Liu, Zoey Liu, Qingyun Yang, Yujing Huang,

EmilyPrud'Hommeaux. Evaluating the performance of transformer-based language models for neuroatypical language. In **Proceedings of COLING. International Conference on Computational Linguistics**, 第 2022 巻, p. 3412. NIH Public Access, 2022.

[6] Emanuel DiCicco-Bloom, Catherine Lord, Lonnie Zwaigenbaum, Eric Courchesne, Stephen R Dager, Christoph Schmitz, Robert T Schultz, Jacqueline Crawley, and Larry J Young. The developmental neurobiology of autism spectrum disorder. **Journal of Neuroscience**, Vol. 26, No. 26, pp. 6897–6906, 2006.

[7] Carlos A Pardo and Charles G Eberhart. The neurobiology of autism. **Brain pathology**, Vol. 17, No. 4, pp. 434–447, 2007.

[8] Andrea C Samson, Whitney M Wells, Jennifer M Phillips, Antonio Y Hardan, and James J Gross. Emotion regulation in autism spectrum disorder: evidence from parent interviews and children's daily diaries. **Journal of Child Psychology and Psychiatry**, Vol. 56, No. 8, pp. 903–913, 2015.

[9] Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, et al. From persona to personalization: A survey on role-playing language agents. **arXiv preprint arXiv:2404.18231**, 2024.

[10] 直江大河, 原田宥都, 前田ありさ, 森田早織, 中村啓信, 大関洋平, 沖村宰. 大規模言語モデルへの刈り込みによる精神疾患の思考障害シミュレーション. 言語処理学会第 30 回年次大会, pp. 243–248, 2024.

[11] Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Yu-Ching Hsu, Jia-Yin Foo, Chao-Wei Huang, and Yun-Nung Chen. Two tales of persona in llms: A survey of role-playing and personalization. **arXiv preprint arXiv:2406.01171**, 2024.

[12] Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. PersonaLLM: Investigating the ability of large language models to express personality traits. In **Findings of the Association for Computational Linguistics: NAACL 2024**, pp. 3605–3627, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[13] Quan Tu, Chuanqi Chen, Jinpeng Li, Yanran Li, Shuo Shang, Dongyan Zhao, Ran Wang, and Rui Yan. Characterchat: Learning towards conversational ai with personalized social support. **arXiv preprint arXiv:2308.10278**, 2023.

[14] Carrie L Fortenberry, Cathy L Grist, and David M McCord. Personality trait differences between typically developing children and those diagnosed with autism spectrum disorder. **Individual Differences Research**, Vol. 9, No. 2, pp. 73–83, 2011.

[15] Aleksandra Sorokovikova, Sharwin Rezagholi, Natalia Fedorova, and Ivan P. Yamshchikov. LLMs simulate big5 personality traits: Further evidence. In **Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)**, pp. 83–87, St. Julians, Malta, March 2024. Association for Computational Linguistics.

[16] Petra Hendriks, Charlotte Koster, and John CJ Hoeks. Referential choice across the lifespan: Why children and elderly adults produce ambiguous pronouns. **Language, cognition and neuroscience**, Vol. 29, No. 4, pp. 391–407, 2014.

[17] Sanne JM Kuijper, Catharina A Hartman, and Petra Hendriks. Who is he? children with asd and adhd take the listener into account in their production of ambiguous pronouns. **PloS one**, Vol. 10, No. 7, p. e0132408, 2015.

[18] Jackson L. Lee, Ross Burkholder, Gallagher B. Flinn, and Emily R. Coppess. Working with chat transcripts in python. Technical Report TR-2016-02, Department of Computer Science, University of Chicago, 2016.

[19] 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. 言語処理学会第 28 回年次大会, 2022.

[20] Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. JGLUE: Japanese general language understanding evaluation. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 2957–2966, Marseille, France, June 2022. European Language Resources Association.

[21] 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.

[22] Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. **arXiv preprint arXiv:2308.11483**, 2023.

[23] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. **arXiv preprint arXiv:2203.11171**, 2022.

[24] Susie Chandler, Tony Charman, Gillian Baird, Emily Simonoff, TOM Loucas, David Meldrum, Mimi Scott, and Andrew Pickles. Validation of the social communication questionnaire in a population cohort of children with autism spectrum disorders. **Journal of the American Academy of Child & Adolescent Psychiatry**, Vol. 46, No. 10, pp. 1324–1332, 2007.

[25] Diana L Robins, Karís Casagrande, Marianne Barton, Chi-Ming A Chen, Thyde Dumont-Mathieu, and Deborah Fein. Validation of the modified checklist for autism in toddlers, revised with follow-up (m-chat-r/f). **Pediatrics**, Vol. 133, No. 1, pp. 37–45, 2014.

[26] Helen McConachie, Jeremy R Parr, Magdalena Glod, Jennifer Hanratty, Nuala Livingstone, Inalegwu P Oono, Shannon Robalino, Gillian Baird, Bryony Beresford, Tony Charman, et al. Systematic review of tools to measure outcomes for young children with autism spectrum disorder. 2015.

[27] Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. **Advances in Neural Information Processing Systems**, Vol. 36, , 2024.

A　質問文の翻訳・生成

A.1　質問文の翻訳

ChatGPT により “Which story was told by a child with ASD?” をオランダ語に翻訳し，“Welke verhaal werd verteld door een kind met ASS?” を得た．

A.2　質問文の生成

ChatGPT に同じ意味の別の表現を作成するよう指示し，“Welke verhaal werd verteld door een kind met ASS?” の他に，

- “Welk verhaal is door een kind met ASS verteld?”
- “Door welk kind met ASS werd een verhaal verteld?”

を得た．

A.3　プロンプトと回答

プロンプト

指示
1. 英語からオランダ語に意味を変えることなく翻訳してください。
2. オランダ語の文章と同じ意味の別の文章をオランダ語で単語の正確さ、流暢さの観点に考慮して追加で 2 文示してください。
文章
Which story was told by a child with ASD?

回答

翻訳:
Welke verhaal werd verteld door een kind met ASS?
同じ意味の別の表現:
1. Welk verhaal is door een kind met ASS verteld?
2. Door welk kind met ASS werd een verhaal verteld?

コードのプロンプト部分

```
{
  'type': 'image_url',
  'image_url': {
      'url': f'{
          base64_image_lis[0]}'
  }
},
{
  'type': 'image_url',
  'image_url': {
      'url': f'{
          base64_image_lis[1]}'
  }
},
{
  'type': 'image_url',
  'image_url': {
      'url': f'{
          base64_image_lis[2]}'
  }
},
{
  'type': 'image_url',
  'image_url': {
      'url': f'{
          base64_image_lis[3]}'
  }
},
{
  'type': 'image_url',
  'image_url': {
      'url': f'{
          base64_image_lis[4]}'
  }
},
{
  'type': 'image_url',
  'image_url': {
      'url': f'{
          base64_image_lis[5]}'
  }
},
{
  'type': 'text',
  'text': f"Question:␣{qa5_dataset['
      question'][i]}\n0:␣{qa5_dataset['
      choice0'][i]}\n1:␣{qa5_dataset['
      choice1'][i]}\n2:␣{qa5_dataset['
      choice2'][i]}\n3:␣{qa5_dataset['
      choice3'][i]}\n4:␣{qa5_dataset['
      choice4'][i]}\nAnswer:"
}
```

B　実験で使用したプロンプト

GPT-4o Nov 20 2024 version を用いた識別性能の確認実験で使用したコードのプロンプト部分を示す．