言語モデルによる心理的構成概念の再構成

藤澤逸平[1] 山田 祐樹[2] 川北 源二[3] 濱田太陽[1]
[1]株式会社アラヤ 研究開発部 [2]九州大学 基幹教育院 [3]Imperial College London
[1]{fujisawa, hamada_h}@araya.org
[2]yamadayuk@gmail.com
[3]g.kawakita22@imperial.ac.uk

概要

言語モデルを活用し，学術的研究を加速させる取り組みが行われている．心理学では，言語モデルが特定の被験者群の回答を模倣できるか，またモデル自身の心理学的バイアスを検証する研究が行われている．これらの出力を言語モデルが可能な理由の一つとして，言語モデルが概念間の関係性を学習している可能性がある． 本研究では、GPT-4 を含む言語モデルによって，質問項目から概念のカテゴリー分類の再構成が可能かどうかを検証した．複数の言語モデルを用いて心理学的質問紙の項目間の類似度を計算し，概念のカテゴリー分類性能を比較した．実験結果は、GPT-4 が最も高い分類性能を示し， 言語モデルが心理学的概念間の関係性を保持している可能性を示唆している．

1 はじめに

言語モデルを活用し，学術的研究を加速させる取り組みが行われている．創薬における実験の自動化[1]や，医療での応用可能性の検討[2]，心理学におけるサンプリングの代替や介入方法としての活用も試みが進んでいる[3,4]．心理学においては，大規模言語モデル(Large Language Model; LLM)が，被験者の性格特性を診断できるか[5]，被験者の回答を代替できるか[6,7]，アニメのキャラクターの回答を模倣できるかといった検証が行われている．これらの研究は，GPT-4 などの言語モデルが膨大なテキストデータによるモデル訓練を経て，ヒトが持つ心理的概念と関連する単語や文章との距離関係を学習している可能性を示唆している．しかし,言語モデルがこの距離関係に基づいて回答しているかどうかは明らかではない．

本研究では，GPT-4 を含む言語モデルが，心理学的質問紙の質問項目から概念のカテゴリー分類を再構成することが可能かどうかを検証した．GPT-3.5 や GPT-4 などの言語モデルを用いて，43 の心理学的質問紙の質問項目から類似度を計算し，類似度に基づいた分類と質問紙に付与されている構成概念のラベルとの一致度を測ることで，モデルの分類性能を比較する．これにより，言語モデルが心理学的概念間の関係性を保持している可能性について検証する．

2 関連研究

本研究で取り扱う心理学的質問紙は，人の行動から観察可能な現象を説明する構成概念に基づいて作成されている．さらに，この構成概念に基づいて質問項目が作成されており，それぞれの質問項目には構成概念のラベルが割り振られている．例えば，好奇心に関する質問紙 CEI-II(Curiosity and Exploration Inventory)は，10 項目で構成されている．“I actively seek as much information as I can in new situations.”という質問項目には“Stretching”というラベルが割り振られている．本研究では，これらの質問項目同士の類似度を複数の言語モデルを用いて計算し，ラベルを正解データとして分類性能を評価する．ここでは，文章同士の類似度を計算する方法と，汎用型言語モデルを利用してヒトの知識や性格特性を再構成する研究について触れる．

2.1 文章間の類似度算出

文章間の類似度を算出することは，文書のカテゴリ分類や関連度の高い文書の検索などに利用される．文章間の類似度を算出する方法として，単語分散表現や文章分散表現の利用が知られている．第一に，文章を構成する単語を全て低次元ベクトルに変換する単語分散表現モデルがある．例えば，大規模なテキストコーパスを用いて、各単語を固定長のベクトルとして表現する word2Vec[8]や，transformer のエン

コーダーを利用して文脈に基づいて単語の意味を動的に調整する BERT (Bidirectional Encoder Representations from Transformers) [9]がある. 第二に, 単語ではなく文自体を直接ベクトル化する文書分散表現もある. sentenceBERT [10]や Universal Sentence Encoder [11]などが知られている. 通常の BERT は, 単語の分散表現から文章レベルの表現を行う一方で, 文章自体のベクトル化に最適化されていない. sentenceBERT は, 文章のベクトル化を直接的に行う. また, OpenAI が提供しているテキスト埋め込みモデル(“text-embedding-ada-002”)[12]も, 文章分散表現を利用していると考えられている. 分散表現で得られたベクトルに対しコサイン類似度を計算することで, 単語間や文章間の類似度とする.

2.2 言語モデルによるヒトの知識や性格特性の再現

GPT-3 や ChatGPT などの LLM の登場以降, LLM を用いてさまざまなドメインでヒトの知識や性格特性を再現できるか研究が行われている[3-6]. 心理学では, LLM で得られた回答がどんな集団の回答を反映しているのか検証が進められている. 例えば, ヒトの集団や, 小説やアニメなどのキャラクターの応答を模倣させる研究がある[14,15]. Argyle ら(2023)の研究では, GPT-3 にヒトの被験者の背景情報を盛り込んだ上で人口動態に基づいた質問を答えさせた[14]. その結果, ヒトの実際のサンプルと GPT-3 で模倣した結果に対応関係が見られた. Li ら (2023)の研究では, アニメのキャラクターのセリフを学習させ, それを元に性格質問紙を回答させている[15]. Culter&Condon (2023)は, 代表的な質問紙である Big Five を用いた被験者の回答と、言語モデルによる質問項目の形容詞の類似度の間に一致を確認している[16]. 他にも, 大規模言語モデルが学習した色同士の対応関係がヒトの色に関する認知的な構造を反映しているのかどうかの検証も行われている[17,18]. このようにヒトの膨大なテキストや画像情報から学習した LLM を心理学的検査方法によって分析することが行われている[19]. ヒトが持つ認知バイアスと LLM の回答の対応関係が明らかになることで, ヒトの回答を取得するために直接実験を行わずとも LLM を通じて同等の情報を取得できる可能性を示している.

3 実験

本実験では, GPT-4 などの言語モデルの分類性能を検証する. まず, 質問項目の分類に関する実験によって, 言語モデル間の性能評価を行う. 次に, プロンプトの違いによって, 性能に差が出るか比較を行う. 最後に, 質問項目を与える順番による分類性能の変化を調べる.

3.1 データセット

心理学的質問紙データベース”Psychological Scales”[i]より, 文章で構成されている質問項目数 30 以下の英文質問紙を 43 個抽出する. それぞれの質問紙は, “好奇心”, “自己効力感”, “不安”などの構成概念より作成されており, 下位のカテゴリーを 2 以上 6 以下包含している.

3.2 実験設定

BERT, OpenAI が提供する埋め込みモデル, GPT-3.5, GPT-4 を含む 6 つの言語モデルの質問項目の分類性能を比較する[表 1]. BERT と埋め込みモデルでは, 全ての質問紙の質問をそれぞれ埋め込むことで、項目間の類似度としてコサイン類似度を算出し, これを類似度行列とする[図 1]. GPT-3.5 と GPT-4 では, 質問紙ごとに質問項目間の意味に応じて[-1, 1]の範囲で回答するようにプロンプトで指示を与える[付録 A, 表 2]. ここで, 1 は意味が同じことを、-1 は意味が逆であることを指す. 得られた回答を, 質問紙ごとの質問項目間の類似度行列とする.

得られた類似度行列に, 階層的クラスタリングを適用する. この際, カテゴリー数は質問紙ごとに決められたカテゴリー数を利用する. 文章分類の指標として, 分類した質問項目と質問項目が属するカテゴリーの正答率と, クラスタリングの一致度の評価に用いられる調整済みランド指数 (Adjusted Rand Index, ARI)[10]を利用する.

実験 1 本実験では, GPT-3.5 と比べ GPT-4 が質問紙の分類性能が良いと仮説を立てる. 43 の質問紙よりそれぞれ類似度行列を作成し, 質問紙のカテゴリーのラベルと一致率を算出し, GPT-3.5 と GPT-4 を含む 6 つの言語モデルで統計的検定(対応ありスチューデントの t 検定)を用いた比較を行う.

[i] https://scales.arabpsychology.com/

表 1 本研究で利用する 6 種の言語モデル

ID	言語モデル	モデル情報
1	bert-base-uncased (以下, BERT)[ii]	パラメーター数: 110M
2	Open AI Embedding (以下, 埋め込みモデル)	OpenAI が提供する 埋め込みモデル “text-embedding-ada-002”
3	gpt-3.5-turbo-0613 (以下, GPT-3.5(0613))	OpenAI が提供する API. パラメーター数: 未公開, コンテキストウィンドウ: 4096 トークン,学習データ: 2021 年 9 月まで
4	gpt-3.5-turbo-1106 (以下, GPT-3.5(1106))	パラメーター数: 未公開, コンテキストウィンドウ: 16,385 トークン, 学習データ: 2021 年 9 月まで
5	gpt-4-turbo-0613 (以下, GPT-4(0613))	パラメーター数: 未公開, コンテキストウィンドウ: 8,192 トークン, 学習データ: 2021 年 9 月まで
6	gpt-4-turbo-1106-preview (以下, GPT-4(1106))	パラメーター数: 未公開, コンテキストウィンドウ: 128,000 トークン, 学習データ: 2023 年 4 月まで

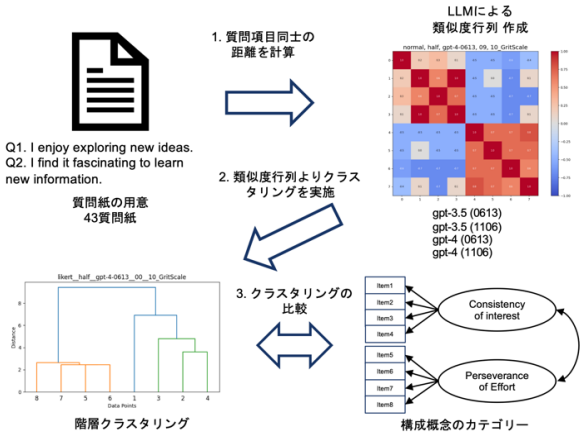

図 2 実験の手順

実験 2 プロンプトの与え方により GPT-3.5/4 の変動が出るか検証を行う. プロンプトには, 多段階の選択肢を用いて回答させるリッカート尺度(9 段階)を用いたものと, 通常の連続値による出力を行うものを利用する. この際, 10 の質問紙に同じプロンプトを用いて, 質問項目の分類正答率と ARI に違いが出るのか, 統計的検定(対応ありスチューデントの t 検定)を用いた比較を行う.

実験 3 GPT-4 による質問項目間の類似度の出力がどの程度安定的なのか 8 つの質問紙を用いて検証する. 同じプロンプトによる出力を 10 回行い平均値と標準偏差の算出を行う. また, GPT-4(0613)に与える順序で分類正答率に違いが出るのか, 対応ありスチューデントの t 検定を用いた比較を行う.

3.3 実験結果と考察

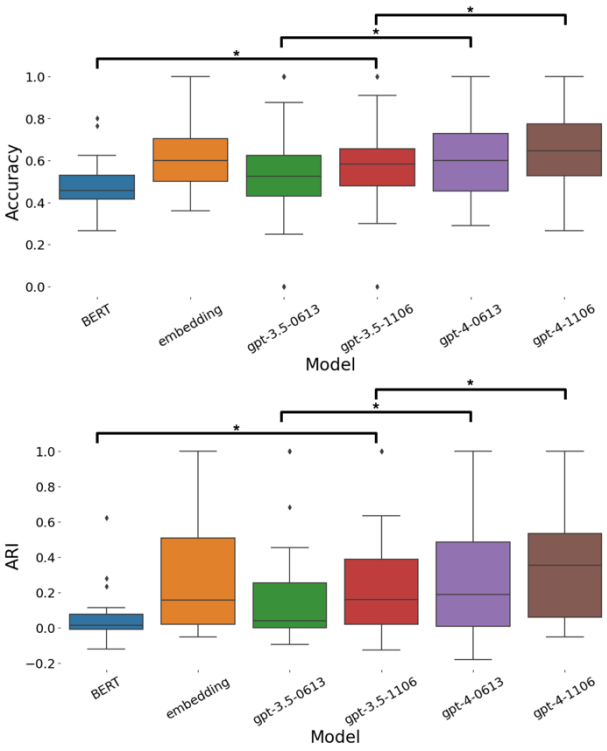

図 2 モデル別の正答率と ARI
(*は, p<0.05 を示す, 多重比較補正済)

3.3.1 実験 1: 分類性能の評価

言語モデルの正答率と ARI の結果を図 2 に示す. 平均の正答率と平均の ARI ともに, GPT-4(1106)が最も高い精度 66.0%を示した. また, BERT は最も低い精度 47.8%を示した.

GPT-3.5(0613)と GPT-3.5(1106)の正答率の平均と GPT-4(0613)と GPT-4(1106)の正答率の平均の間に, 統計的有意差が認められた (p<0.001, 付録 A, 図 5). これは, 平均的に GPT-4 が GPT-3.5 より質問項目の分類性能が高いことを示している. その原因として,

[ii] https://huggingface.co/bert-base-uncased

モデルのアーキテクチャー，モデルのパラメーター数，GPT-4(1106)が 2023 年までのデータを利用している訓練時のデータ量の違いによる影響を受けた可能性がある．

埋め込みモデル，GPT-4(0613)，GPT-4(1106)の間で正答率と ARI で統計的有意差は示されなかった(図 2)．埋め込みモデルは，一度文章間の距離をベクトル化すれば，コサイン類似度を利用して類似度を計算することができ，計算量が小さく出力が一定である．また，GPT-3.5/4 は，transformer モデルのデコーダーの部分を利用しており，コサイン類似度ではなく，プロンプトに基づいて，文章の類似度を生成している．これにより，出力にランダム性があるものの，高い平均正答率と平均ARIが得られた可能性がある．

3. 3. 2 実験 2：プロンプトの比較

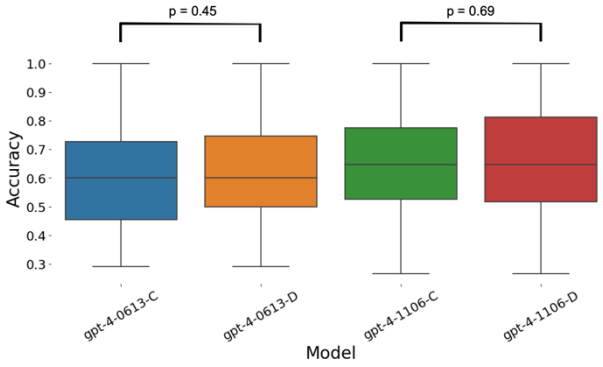

図 3 プロンプトの比較
(-C は連続値, -D は離散値を出力するプロンプトの結果を示す.)

各言語モデルにおける離散値（リッカート尺度）と連続値による正答率の比較の結果を図 3 に示す．GPT-3.5 では，余分な文字列を含む出力をしたため，GPT-4 の結果のみ示す．連続値と離散値のプロンプトの出力は，正答率も ARI も同様に統計的有意差を示さなかった．これは，類似度の出力は，連続値と離散値による差分が小さいことを示唆している．

3. 3. 2 実験 3：質問項目の一貫性

図 4 に実験結果を示す．8 つ質問紙の 10 回の試行の平均正答率は正順 64.3%，逆順:62.7%，標準偏差は正順 4.4%，逆順:7.0%となった．また，8 つの質問紙のうち 2 つの質問紙で，プロンプトに与える項目の順序を逆転したことによる分類正答率に統計的優位差が認められた($p<0.0001$,多重比較補正済)．これは，プロンプト文に与える質問文の順序により類似度の出力，分類性能に影響を与えることを示唆している．

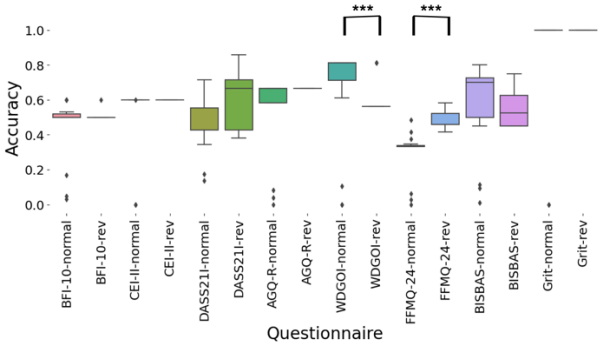

図 4 質問項目間の一貫性

4 おわりに

本研究では，言語モデルを利用して，構成概念に基づいた心理学的質問紙の項目からカテゴリー分類の再構成が可能かを検証した．実験 1 では，学習データ量，モデルのパラメーター数やアーキテクチャーの違いから GPT-4 が GPT3.5 より分類性能が良いという仮説を立て，43 の質問紙を用いて分類評価を行った．その結果，GPT-4 は 63.7%の平均正答率を示し，GPT-3.5 の平均正答率 56.7%を上回った．一方で，埋め込みモデルで，60.5%の平均正答率を出し，GPT-4 の出力と同等の性能を示すことが明らかになった．また，LLM におけるプロンプトにおける連続値か離散値による出力は分類性能に影響を与えるという仮説を立て，実験を実施した．結果，連続値と離散値による出力は分類性能に影響を与えないことが示唆された．さらに，質問項目の順番によって大規模言語モデルの出力に影響があることが知られており，この順番によって分類性能に影響があるという仮説を立て，実験を実施した．結果，質問項目の順番は，分類性能に影響を与える可能性が示唆され，先行研究と同等の結果が確認された[21].

以上の結果より，言語モデルが，質問項目から心理学的構成概念のカテゴリーを再構成できることを示した．言語モデルが，心理学的構成概念と関連する単語や文章に関する情報を保持していることが示唆された．今後は，複数の質問紙に対するヒトの回答と言語モデルによる複数の質問紙間の類似度との対応関係や多言語での対応関係について調査する．これにより，大規模言語モデルが，心理学的構成概念間の距離に関する情報を保持しているのかを明らかにする．

謝辞

本研究は, JST ムーンショット型研究開発事業 JPMJMS2295 の助成を受けた.

参考文献

1. Boiko, D., MacKnight, R., Kline, B., and Gomes, G., Autonomous chemical research with large language models. Nature, Vol. 624, pp. 570-8, 2023.
2. Singhal, K., Azizi, S., Tu, T., et al., Large language models encode clinical knowledge. Nature, Vol. 620, pp. 172-80, 2023.
3. Demszky, D., Yang, D., Yeager, D., et al., Using large language models in psychology. Nat Rev Psychol, Vol. 2, pp. 688–701, 2023.
4. Sartori, G., and Orru, G. Language models and psychological sciences. Front. Psychol. Vol. 14, 1279317, 2023.
5. Yang, T., Shi, T., Wan, F. et al., PsyCoT: Psychological Questionnaire as Powerful Chain-of-Thought for Personality Detection. arXiv, 2023.
6. Dilion, D., Tandon, N., Gu, Y., et al., Can AI language models replace human participants? Trends in Cognitive Sciences, Vol. 27 (7), pp. 597-600, 2023.
7. Santurkar, S., Durmus, E., Ldhak, F., et al., Whose Opinions Do Language Models Reflect? arXiv, 2023.
8. Mikolov, T., Chen, K., Corrado, G., et al., Efficient Estimation of Word Representations in Vector Space. arXiv, 2013.
9. Devlin, J., Chang, M., Lee, K. et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv, 2018.
10. Reimers, N., and Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv, 2019.
11. Cer, D., Yang, Y., Kong, S., et al. Universal Sentence Encoder. arXiv, 2018.
12. Embeddings. (引用日: 2024 年 01 月 08 日.) https://platform.openai.com/docs/guides/embeddings/.
13. Kong, A., Zhao, S., Chen, H., et al., Better Zero-Shot Reasoning with Role-Play Prompting. arXiv, 2023.
14. Li, C., Leng, Z., Yan, C., et al., ChatHaruhi: Reviving Anime Character in Reality via Large Language Model. arXiv, 2023.
15. Argyle, L., Busby, E., Fulda, N., et al., Out of One, Many: Using Language Models to Simulate Human Samples. Political Analysis, Vol. 31(3), pp. 337-351, 2023.
16. Culter, A., & Condon, D.M. Deep lexical hypothesis: Identifying personality structure in natural language. Journal of Personality and Social Psychology, Vol. 125(1), pp. 173-97, 2023.
17. Kawakita, G., Zeleznikow-Johnston, A., Tsuchiya, N., et al., Comparing Color Similarity Structures between Humans and LLMs via Unsupervised Alignment. arXiv, 2023.
18. Loyola, P. Marrese-Taylor, E. and Hoyos-Idobro, A. Perceptual Structure in the Absence of Grounding for LLMs: The Impact of Abstractedness and Subjectivity in Color Language. arXiv, 2023.
19. Brinkmann, L., Baumann, F., Bonnefon, J., et al., Machine Culture. Nat Hum Behav, vol. 7, pp. 1855-68, 2023.
20. Wanger, S., and Wanger, D. Comparing clusterings: Karlsruhe: Universität Karlsruhe, Fakultät für Informatik, 2007.
21. Peeshkpour, P. and Hruschka, E., Large Language Models Sensitivity to The Order of Options in Multiple-Choice Questions. arXiv, 2023.Cer, D., Yang, Y., Kong, S., et al. Universal Sentence Encoder. arXiv, 2018.

A 付録

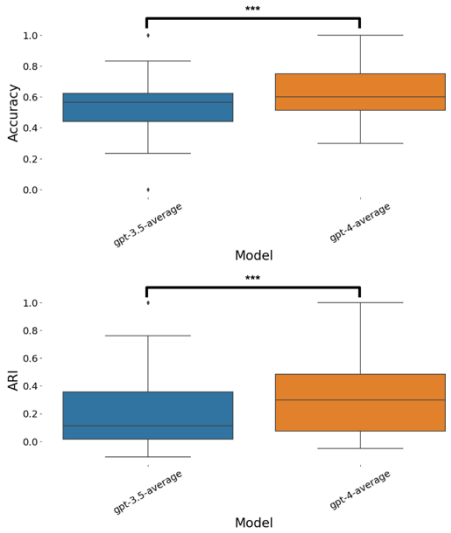

図 5 GPT-3.5 vs. GPT-4 の比較
(***は, p<0.001 を示す.)

表 2 プロンプト文.
連続値と離散値に基づいたプロンプト文

プロンプトカテゴリー	プロンプト文
連続値	You are an expert of natural language. You will get two questions. You should answer the similarity score of two questions between -1 and 1 as a real number. A score close to 1 indicates high similarity, while a score closes to -1 suggests two questions have opposite meanings. A score around 0 implies two questions are neither similar nor opposite. If the similar You must not answer anything else neither add an explanation.
離散値(リッカート尺度)	You are an expert of natural language. You will get two questions. You should answer the similarity score of two questions from 1, 2, 3, 4, 5, 6, 7, 8, and 9. A score close to 9 indicates very similar, while a score closes to 1 suggests two questions are not very similar. A score around 5 implies two questions are neither similar nor opposite. If the similar You must not answer anything else neither add an explanation.