LLM はメンタルヘルス予測が可能か？ LLM によるメンタルヘルスデータセットの構築と評価

奥村紀之 [1] 赤田太郎 [2]
[1] 武庫川女子大学 社会情報学部 社会情報学科 情報サイエンス専攻
[2] 常葉大学 健康プロデュース学部 スポーツ健康科学科
okumura_noriyuki_x@mukogawa-u.ac.jp, t-akada@hm.tokoha-u.ac.jp

概要

本稿では，被験者の日記に基づく LLM を用いたメンタルヘルスの不調の検出性能について評価している．被験者実験を実施するにはハードルが高いため，LLM によって擬似的な日記を生成し、擬似的に生成された記述に対して各日のメンタルヘルスの状態を予測している．生成する擬似的な日記は 3 ヶ月分とし，各日のメンタルヘルスの状態に関して，日記の記述に対する LLM による予測と臨床心理士による判定を行い評価している．評価実験の結果，LLM はメンタルヘルスの状態を過大評価する傾向にあることが分かった．

1 はじめに

メンタルヘルスは現代社会における極めて重要な課題であり，WHO の調査[1)]によれば，世界で 4 人に 1 人が何らかの精神疾患を抱えているとされている [1, 2]．日本の人口が 1 億 2000 万人とすれば，WHO の統計によるとメンタルヘルスに何らかの不調を来している人はおよそ 3000 万人存在していることになる．メンタルヘルスの不調を早期に発見し，適切な対応をすることは急務であるが，日本の精神科医はおよそ 16000 人[2)]，臨床心理士では 43000 人ほど[3)]であり，およそ 500 人に 1 人の割合でしかカウンセリングを実施できる専門家がいない状況であることから，人間による対応が難しい状況にあると考えられる．

このような状況であるため，2022 年 11 月に公開された ChatGPT[4)]や Gemini[5)]に代表される大規模言語モデル（以下，LLM）によるチャットシステムを用いて精神医学におけるカウンセリングに適用しようとする試みはあるが，人間同士のカウンセリングには及ばないとされている [3]．また，仮に人間のカウンセリングに近い水準のものが実施できるとしても，その根拠となるデータセットは大半が英語で構成されており，文化の違いなどを考慮した上で，英語以外の言語における適切なカウンセリングに適用できる可能性は高くはないと推察される．

本研究では，One-shot プロンプティングにより擬似的な日記を生成し，LLM によるメンタルヘルスの状態予測を行っている．また，生成した日記に対し，臨床心理士によるメンタルヘルス状態の予測も同時に実施している．これにより，LLM と専門家の評価の差異を明確にし，LLM がメンタルヘルスをどのように捉えているかを評価している．評価実験により，LLM は人間よりもメンタルヘルスの状態を過大評価する傾向にあることが分かった．

2 関連研究

LLM によるメンタルヘルスの状態推定に関して，3 つの観点から関連研究を紹介する．1 つ目はデータセット構築における LLM の活用方法に関する研究，2 つ目はデータセットのソースとアノテーションに関する研究，3 つ目はデータセット構築における課題と倫理的配慮に関する研究である．

2.1 データセットの構築

1 節で述べたとおり，メンタルヘルスに関連するデータセットは主として英語で構成されている．LLM によるデータ構築法として，既存の英語のデータセットを多言語に翻訳し，多言語対応のデータセットを構築する試みがある [4]．Skianis らの研究では，リソースが乏しい言語（たとえばトルコ語や

1） https://www.who. int/health-topics/mental-health
2） https://www.mhlw.go.jp/toukei/saikin/hw/ishi/22/dl/R04_kekka-1.pdf
3） http://fjcbcp.or.jp/rinshou/about-2/
4） https://chatgpt.com/
5） https://gemini.google.com/

フィンランド語など）でも，低コストで分析用データを確保することが可能となるとしている．

日本語データセットの研究としては，LLM を用いて X（旧 Twitter）の投稿からメンタルヘルスの不調を示唆する特徴的なキーワードを抽出しリスト化する手法が提案されている [5]．高畠らの研究では，専門家（カウンセラー）の役割を LLM にあたえるプロンプトエンジニアリングにより，文脈に応じた深刻度の高い事例を優先的に特定することが可能とされている．

2.2 データセットのソースとアノテーション

Reddit や X（旧 Twitter）などの匿名プラットフォームは，ユーザが自身のメンタルヘルスの状態を自己報告する場となっており，大量のテキストを収集する主要なソースとなっている [1]．Ji らの提案している MentalBERT においてもこれらのテキストデータはソースとして活用されている．

また，病院でのインタビュー記録や，専門家が診断した EEG（脳波）データと対話データを組み合わせたマルチモーダルなデータセットも構築されている [6]．Cai らの研究は，LLM が生成したラベルの信頼性を検証するためのゴールド標準（ある診断や評価が正しいかどうかを判断するための最も信頼できる基準）となっている．

2.3 データセット構築における課題と倫理的配慮

LLM を用いたデータセット作成にはいくつかの課題がある．SNS 由来のデータは必ずしも臨床的な診断と一致するわけではなく，自己報告に基づくものであるため，LLM による自動ラベル付けでは護身のリスクが伴う [4]．また，学習データに含まれる人種や性別の偏りがデータセットに継承される健康データ貧困の問題も懸念されている [7]．Arora らは，自由記述テキストの生成品質を評価するための統一された指標が不足しており，現状では BLEU や ROUGE といった従来の機械翻訳等で利用される評価指標や人間による評価を組み合わせて使用しているのが実態であるとしている．さらに，データの匿名化や収集プロセスにおけるインフォームドコンセントの確保，データ収集方法の透明性（サンプリング手法の明示など）の維持などのために標準化が必要であることが述べられている．

このように，LLM を用いたメンタルヘルスデータセットの構築においては，多言語化，コスト削減，細かな特徴抽出を可能とする一方で，臨床的な妥当性と倫理的な安全性を確保するためには依然として専門家による検証と厳格なデータ管理の基準が不可欠となっている．

3 LLM による擬似的な日記データの作成とアノテーション

関連研究でも述べたように，被験者への倫理的配慮などの問題から，日記等のテキストデータを実際に被験者から取得することは研究を実施する上でもハードルが高い．一方で，LLM のように無数のテキストデータから学習された言語モデルであれば，不特定多数の日記のような記述についても学習していると考えられる．そこで本稿では，LLM による日記データの自動構築を行い，被験者実験に依存しない実験を行う．

本節以降の実験において使用する LLM は GPT5 であり，API による生成を行っている．temperature は規定値の 1.0 である．

3.1 One-shot プロンプティングによる擬似的な日記データの作成

事例を提示することなく，無作為に擬似的な日記の生成を行うと，日記として違和感のあるものが生成される可能性がある．そのため本稿では，奥村が検証のために利用した自身の日記データとメンタルヘルスの状態を事例 [8] として付与した One-shot プロンプティングにより生成を行う．

擬似的な日記データの作成においては，以下の条件を疑似被験者のペルソナとしてランダムに与えるものとする．

年齢　10 代後半, 20 代前半, 20 代後半, 30 代前半, 30 代後半, 40 代前半, 40 代後半, 50 代前半, 50 代後半, 60 代, 70 代

性別　男性, 女性

職業　高校生, 大学生, 大学院生, 会社員 (営業), 会社員 (事務), 会社員 (エンジニア), 会社員 (企画), 会社員 (人事), 公務員 (市役所), 公務員 (教員), 公務員 (警察), 医療従事者 (医師), 医療従事者 (看護師), 医療従事者 (薬剤師), 介護士, 自営業 (飲食), 自営業 (IT フリーランス), 経営者, 主婦/主夫, パート/アルバイト, 無職, 求職中, 退職者

性格　真面目, 楽観的, 心配性, 社交的, 内向的, 完璧主義, マイペース, 短気, 繊細

趣味　ゲーム, 読書, スポーツ, 料理, 旅行, 映画鑑賞, 音楽鑑賞, 散歩, 特になし

ストレッサー　仕事のプレッシャー, 人間関係, 将来

の不安, 健康問題, 金銭的な悩み, 家庭の問題, 孤独, 特になし

擬似的な日記の生成に用いたプロンプト

あなたは熟練した心理療法士です。様々な背景を持つ人々のメンタルヘルス状態をシミュレーションすることに長けています。指示された詳細なペルソナに基づいて、その人物が書くであろう日記と、その時のメンタルヘルス状態（心理療法士による客観的評価）を生成してください。

生成された日記の一部を表 1 に示す．この例で選択されているペルソナは 10 代後半, 男性, 会社員 (企画), 繊細, 料理, 金銭的な悩みである．LLM には，各日の日記について，メンタルヘルスの不調を来している可能性について，none（可能性なし），low（可能性低）, mid（可能性中）, high（可能性大）として判定させている．

表 1　生成した擬似的な日記の例

日付	日記	LLM によるメンタルヘルスの判定
2021/01/20	財布の中身を数えてため息が出た。今日は節約でうどん、ネギと卵だけでも意外と満足。	low (none)
2021/01/21	資料の締め切りが重なり、肩と首が張ってつらい。帰宅後に生姜たっぷりの豚汁で温まった。	mid (none)
2021/01/22	上司からフィードバックをもらい改善点が明確になった。夜は焼きそばで手早く済ませて早めに寝た。	none (none)

3.2　臨床心理士による評価

表 1 に示した擬似的に生成した日記とメンタルヘルスの状態を予測を付与したデータに対し，臨床心理士によるアノテーションを行う．本稿の第 2 著者である赤田は臨床心理士の資格を有している[6)]ため，本稿においては赤田の基準でのアノテーションとしている．表 1 における LLM によるメンタルヘルスの判定に関して，括弧書きで挿入されているものは赤田による評価結果である．

6）　https://www.youtube.com/aktr_ch

4　評価

本節では，LLM によるメンタルヘルスの状態の推定結果と臨床心理士による判定の差について検討する．評価の観点としては，LLM による推定結果が臨床心理士による判定結果に比べて過大評価されているのか，過小評価されているのか，等価であるのかという点である．

評価対象として，3.1 節で生成した 26 名分の擬似的な日記（90 日分），2340 文の日記について臨床心理士による目視評価を行う．none と評価されたものを 0，low と評価されたものを 1，mid と評価されたものを 2，high と評価されたものを 3 として，LLM によって予測したメンタルヘルスの状態と臨床心理士による評価の差を求め集計したものを表 2 に示す．

表 2　LLM による予測と臨床心理士による予測の差

負	ゼロ	正	計
694	1447	199	2340

5　考察

表 2 では，臨床心理士の評価よりも LLM による予測の方が大きな値を付けることが示唆されている．たとえば，臨床心理士が low と判断した日記に対して，LLM は mid や high を予測しているということになる．したがって，LLM の方がメンタルヘルスに不調を来していると判断する可能性が高いということになる．たとえば，「近所を軽く散歩してから、昔の名作映画を一本見た。途中で胸のあたりが重く感じて不安になり、深呼吸で落ち着かせた。」という日記に対して，LLM は mid と予測しているが，臨床心理士は low と評価している．これは，「胸の辺りが重く感じて不安になり」という表現に LLM は強く反応しているが，臨床心理士は映画のイメージと自身の環境がリンクして不安が高く出ているが、それに対し深呼吸で対応できたところは対処として成功していると見なせるので，低度のメンタルヘルスへの影響が出ていると評価している．過大評価されてしまう要因として，文全体の中でメンタルヘルスの不調と考えられる表現に強く影響されることがあげられる．他にも「罪悪感」というキーワードを持つ日記に対しても同様の傾向が見られる．また，「慰め」という語を含む日記に対しても過大評価されており，メンタルヘルスに不調を来している

際に慰めを求めているという状況が多く見られた可能性が考えられる．

一方で，LLM によって生成された被験者ごとの評価を見ると，26 件の被験者に対して，22 件が過大評価されているが，4 件については過小評価されているという結果になっている．たとえば，「健診の結果が届いて、重大な異常はなしとのこと。ほっとした瞬間、力が抜けて涙が出そうになった。まだ不調はあるが、少し安心した。」という日記に対して，LLM は none と予測し，臨床心理士は mid と評価している．文全体を見ると，異常なし，ほっとした，安心したといったポジティブな印象を受ける表現が多く，この文の本質である「まだ不調はあるが」という点が見落とされてしまっている．また，別の被験者に対しては，「少し頭痛がありペースを落とした。無理をせず読書は短めにした。」という日記に対して，LLM は low と予測し，臨床心理士は high と評価している．これは，頭痛がストレスによる心身反応の 1 つである可能性を考え，身体反応については臨床心理士としてリスクが高いと判定する傾向があるためである．頭痛については医学的な知見が必要となる判定のため，スクリーニングという意味合いでも可能性を高めと評価している．仮にこの頭痛が緊張性頭痛であるならば対処行動も可能であるが，そういった対処法を被験者に伝えていないことを想定し高めの評価としている．

表 2 によると，約 6 割の日記に対しては，LLM による予測と臨床心理士による評価が一致している．しかし，平賀ら [3] が述べているように，6 割程度の一致率では臨床心理士と同等の評価ができるとは判断できない．したがって，過大評価される傾向にある LLM によるメンタルヘルスの状態を予測するモデルを正しく判断させられるようにするための仕組みが必要となる．

6　まとめ

本稿では，LLM を用いてメンタルヘルスに不調を来している可能性のある被験者を擬似的に生成し，その被験者が書いたと考えられる擬似的な日記を生成し，データセットの構築を行った．26 名の擬似的な被験者の日記に対し，LLM によるメンタルヘルスの状況の予測を行った結果，臨床心理士による評価と比較し，LLM による予測の方が過大評価される傾向にあることが分かった．

今後の課題として，過大評価される傾向にある LLM によるメンタルヘルスの状況の予測に関して補正できるようなモデルの構築が挙げられる．よく多くの擬似的なデータに対し，臨床心理士の評価と合わせて検討していく必要があるだろう．

謝辞

本研究は JSPS 科研費 23K02877 の助成を受けたものである．

参考文献

[1] Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. MentalBERT: Publicly available pretrained language models for mental healthcare. **arXiv [cs.CL]**, pp. 7184–7190, October 2021.

[2] M M Syeed, Ashifur Rahman, Laila Akter, Kaniz Fatema, Razib Hayat Khan, Md Rajaul Karim, Md Shakhawat Hossain, and Mohammad Faisal Uddin. A comprehensive standardized dataset on mental health problems of university students, 2024.

[3] 平賀裕貴, 藤崎弘士, 靖史, 吉川栄省. 生成 ai の精神医学またはカウンセリングへの適用について. 日本医科大学基礎科学紀要 = The Bulletin of liberal arts sciences, Nippon Medical School / 日本医科大学基礎科学紀要編集委員会 編, No. 53, pp. 49–81, 2024.

[4] Konstantinos Skianis, John Pavlopoulos, and A Seza Doğruöz. Building multilingual datasets for predicting mental health severity through LLMs: Prospects and challenges. **arXiv [cs.CL]**, September 2024.

[5] 高畠暖佳, 須藤克仁, 狩野芳伸. 大規模言語モデルを用いたメンタルヘルス不調群ツイートの特徴抽出. JSAI 大会論文集, Vol. JSAI2025, No. 0, p. 1Win434, July 2025.

[6] Hanshu Cai, Zhenqin Yuan, Yiwen Gao, Shuting Sun, Na Li, Fuze Tian, Han Xiao, Jianxiu Li, Zhengwu Yang, Xiaowei Li, Qinglin Zhao, Zhenyu Liu, Zhijun Yao, Minqiang Yang, Hong Peng, Jing Zhu, Xiaowei Zhang, Guoping Gao, Fang Zheng, Rui Li, Zhihua Guo, Rong Ma, Jing Yang, Lan Zhang, Xiping Hu, Yumin Li, and Bin Hu. A multi-modal open dataset for mental-disorder analysis. **Scientific Data**, Vol. 9, No. 1, p. 178, apr 2022.

[7] Anmol Arora, Joseph E. Alderman, Joanne Palmer, Shaswath Ganapathi, Elinor Laws, Melissa D. McCradden, Lauren Oakden-Rayner, Stephen R. Pfohl, Marzyeh Ghassemi, Francis McKay, Darren Treanor, Negar Rostamzadeh, Bilal Mateen, Jacqui Gath, Adewole O. Adebajo, Stephanie Kuku, Rubeta Matin, Katherine Heller, Elizabeth Sapey, Neil J. Sebire, Heather Cole-Lewis, Melanie Calvert, Alastair Denniston, and Xiaoxuan Liu. The value of standards for health datasets in artificial intelligence-based applications. **Nature Medicine**, Vol. 29, No. 11, pp. 2929–2938, nov 2023.

[8] 奥村紀之. リモートワーク時代におけるオンライン言語資源に基づくメンタルヘルス予測. システム制御情報学会論文誌, Vol. 67, No. 7, pp. 269–274, 7 2023.