大規模言語モデルと患者表現辞書を用いた病名診断の精度検証

宇都宮和希 [1]　坂野遼平 [2]
[1] 工学院大学　[2] 一橋大学
em23007@ns.kogakuin.ac.jp r.banno@r.hit-u.ac.jp

概要

自然言語処理技術を活用した医療支援が始まりつつあり，医療分野における AI の需要が上昇している．本研究では大規模言語モデル（LLM）の医療応用に着目し，医療プロセスの中でも AI による診断への応用可能性を探る．具体的には，曖昧な患者表現から病名を予測するタスクについて，複数 GPT モデル間の差異検証を行う．また，予測病名から ICD-10 コードを階層的に推定し，その推定精度の検証を行う．結果，差異検証ではファインチューニングを行うことで，より正解病名に近い予測ができる傾向を確認した．また，予測病名が正解病名と一致しない場合においても ICD-10 コードを正しく推定可能なケースが存在することを確認した．

1　研究背景

近年，自然言語処理技術を活用した医療支援が始まりつつある．例えば，新型コロナウイルス感染症の動向を把握する発生届を自動化する取り組み [1] や，BERT[2] を用いた事故分析支援手法の提案 [3] などがある．少子高齢化による医者や病院の不足から AI の遠隔診療への応用 [4] なども挙げられており，医療分野における AI の需要が上昇している．また，AI の医療応用に関する既存研究として，耳鼻咽喉科専門医試験における GPT の有効性に関する検討 [5] などが行われている．一方で，近年急速な発達を遂げている大規模言語モデルを用いた病名診断の可否や精度については明らかではない．

本研究では，大規模言語モデルによる医療支援の一環として，GPT と患者表現辞書を用いて病名予測を行い，GPT のバージョンおよびファインチューニングの有無による影響分析や量的評価を行う．また，ICD-10 コードを用いた病名予測精度の評価についての検討も行う．

2　関連研究

野田ら [5] は ChatGPT による日本語を用いた医療分野における有効性についての報告は少ないという観点から，耳鼻咽喉科専門医試験の選択肢問題に関して日本語のプロンプトと英語のプロンプト，GPT-3.5，GPT-4 などを組み合わせて，多角的に検証，評価を行い，日本語の耳鼻咽喉科領域においての有効性と AI 活用の課題について検討を行った．結果として英語プロンプトの GPT-4 が最も精度が良かった．また，日本語でも耳鼻咽喉科領域において一定の水準を達成できることが確認された．以上のように特定の分野において AI を用いた医療応用に関する研究が為されている一方，日本語による病名予測や LLM を用いた病名診断などの研究は少なく，その精度や予測傾向は明らかではない．

本研究では，曖昧な患者表現を大規模言語モデルに入力し，予測病名を出力することで，大規模言語モデルによる病名予測の検証を行い，医療プロセスの中でも診断への応用可能性を探る．

3　検証方法

本章では，検証手法について説明する．検証手法の流れとして，複数の GPT モデルから病名予測を行い，差異検証を行う．その後，プロンプトを変更し，量的評価を行うことでモデルごとの差異やプロンプト変更による予測精度の変化を確認する．また，病名予測の精度だけでなく，ICD-10 コードを用いた探索の有効性を確認する．

また，各検証方法では表 1 のように特定の病名と患者表現の対応関係が 6393 件まとめられている患者表現辞書 [6] をデータセットとして利用する．

3.1　GPT モデルごとの差異検証

初期的な調査として，ファインチューニングの有無や複数の大規模言語モデルを用いた差異検証を行う．

表 1　患者表現辞書 (一部抜粋)

出現形（患者表現）	ICD-10	標準病名
できものができている	R229	腫瘤
できものがある	R229	腫瘤
放心	F239	急性一過性精神病性
できもん	R229	腫瘤
できものができる	R229	腫瘤
おでき	L029	せつ

図 1 のように使用データセットである患者表現辞書 [6] の 8 割である 5115 件を学習データとし，大規模言語モデルのファインチューニングを行う．大規模言語モデルへの入力データとして残り 2 割の 1279 件からランダムで 10 件の患者表現と標準病名を抽出し，学習済みモデルへ症状を入力，予測病名の出力を行う．

学習済みモデルだけでなく，学習を行っていないモデルなど，複数の大規模言語モデルにも予測病名を出力させ，標準病名とのコサイン類似度を算出することで予測精度やモデルごとの差異調査を行う．ファインチューニングの際には図 2 のようなフォーマットで大規模言語モデルのファインチューニングを行う．

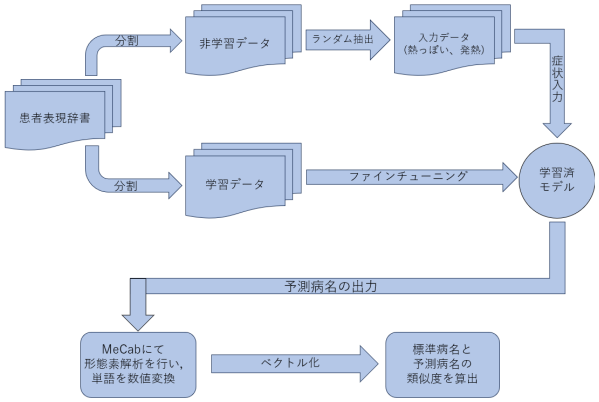

図 1　GPT モデルごとの差異検証

```
{"messages": [{"role": "system", "content": "Answer the disease that matches
the user entered symptoms."}, {"role": "user", "content": "できものができている"},
{"role": "assistant", "content": "腫瘤"}]}

{"messages": [{"role": "system", "content": "Answer the disease that matches
the user entered symptoms."}, {"role": "user", "content": "できものがある"},
{"role": "assistant", "content": "腫瘤"}]}

{"messages": [{"role": "system", "content": "Answer the disease that matches
the user entered symptoms."}, {"role": "user", "content": "放心"},
{"role": "assistant", "content": "急性一過性精神病性障害"}]}
```

図 2　学習用フォーマット

3.2　プロンプト変化による量的検証

次に事前に分割したテストデータからランダムで 100 件の患者表現と標準病名を抽出し，以下 2 つのプロンプトを用いてモデルの量的検証を行う．

- "Answer the disease that matches the user entered symptoms."
- "Answer the **japanese** disease that matches the user entered symptoms."

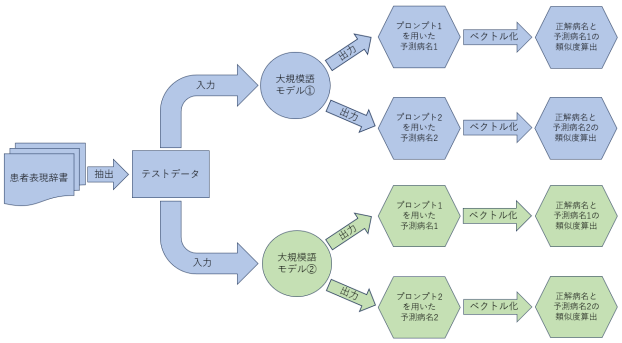

図 3　プロンプト変化による予測結果の量的検証

プロンプトの変更が予測精度にどの程度影響するのかを図 3 のように 2 つの大規模言語モデルを利用して確認する．

3.3　ICD-10 コードによる正誤検証

大規模言語モデルごとの予測病名を用いた類似度の検証だけでなく，ICD-10 コード[1)]による正誤検証も行う．すなわち，予測病名と対応する ICD-10 コードを推定し，正解病名の ICD-10 コードと合致するか否かを評価する．例えば，肺出血や肺胞出血など，病名が異なる場合でも ICD-10 コードが共通の場合や糖尿病と 2 型糖尿病のような階層関係のある病名の判断にも ICD-10 コードの正誤検証は有効だと考えられる．

具体的な手順としては，まず厚生労働省の統計分類 [7] に記載している基本分類表を参考に，中間分類と 3 桁分類を人手で作成する．その後，図 4 のように，中間分類と 3 桁分類のそれぞれの階層において，予測病名と各 ICD-10 コードの病名との類似度を算出し、最も高い類似度となるものを予測病名と対応する推定 ICD-10 コードとする．この推定 ICD-10 コードを，中間分類と 3 桁分類の各階層にて，正解病名の ICD-10 コードと照合し，細粒度・粗粒度での診断の妥当性を評価する．

4　評価

実験環境には Google Colaboratory[8]，大規模言語モデルには GPT モデル [9] を利用した．なお表 2，表 3，表 4 ではファインチューニングを行っていない GPT モデルを N，ファインチューニングを行わ

1)　国際的に統一した基準で定められた死因・疾病の分類

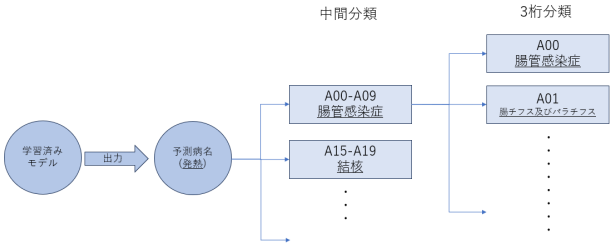

図 4　ICD-10 コードによる正誤検証

ず，初期設定に「You are a doctor.」という性格を付与した GPT モデルを C，ファインチューニングを行った GPT モデルを F として「GPT-3.5(N)」のように表記している．また，実験において明確な病名が出力されなかったケースがあり，表中では「N/A」と記載している．例えば「あなたが言及している病名はありません」などがある．

4.1　GPT モデルごとの差異検証評価

GPT モデルごとの差異検証を行った結果を表 2 に記載する．表 2 より，GPT-3.5(N)，GPT-3.5(C)，GPT-4o-mini(N)，GPT-4o-mini(C) では「白内障」や「胸膜炎」などに対して病名を出力できていない一方，ファインチューニングを行った GPT-3.5(F) では正確に出力できており，他の病名についても比較的高いコサイン類似度が得られていることを確認した．

また，GPT-4o-mini では「あせだぐ」のような崩れた症状に対しても，GPT-3.5(F) と同様に高い精度で予測を行えている事を確認した．一方，「白内障」や「胸膜炎」ともにファインチューニングを行っていないモデルでの予測を行えなかった理由として，漢字のみの入力だったからだと考えられる．

4.2　プロンプト変化による量的検証評価

GPT-3.5 と GPT-4o-mini のプロンプトを変更し，量的評価を行った結果を表 3 に記載する．モデルの比較として「ささやき」や「げろっぽ」などの抽象的，または崩れた症状に対して GPT-3.5 は予測できなかったが，GPT-4o-mini は高い類似度の病名を予測している事を確認した．そして，プロンプトを変更したことで，GPT-3.5，GPT-4o-mini 共に N/A の数を減らし，平均コサイン類似度も上昇した．理由として，4.1 でも記載しているように漢字のみの症状に対しても，プロンプト変更後ではしっかりと日本語での病名を予測できていたからだと考えられる．

また，GPT-3.5 と GPT-4o-mini のコサイン類似度で正解病名との類似度が「1.0」になった予測病名，予測病名と正解病名が一致した病名数をカウントした所，GPT-4o-mini では変更前では 4 つ，変更後で 5 つとなっていた．GPT-3.5 では変更前では 6 つ，変更後では 8 つとなっていた．コサイン類似度に関しては GPT-4o-mini の方が高かったが，正解病名を予測できている数は GPT-3.5 の方が多かった．理由として，GPT-3.5 のカットオフ期間が 2021 年であり，利用している患者表現辞書の更新日が 2021 年であったこと，または GPT-4o-mini のカットオフ期間が 2023 年であり，病名の変更など患者表現辞書と扱っているデータが異なっていたこと，またはその両方が原因だと考えられる．

4.3　ICD-10 コードによる正誤検証評価

差異検証では平均コサイン類似度の精度が一番高い GPT-3.5(F) でも「多汗症」を正しく予測できなかった．しかし，3 桁分類では「発汗異常」という患者表現辞書，厚生労働省の分類表にはなかった予測病名に対して正しい ICD-10 コードを探索できている事を確認した．また，「夜間頻尿症」など，正しい ICD-10 コードの探索を行えなかった理由として，使用した形態素解析器に問題があると考えられるため，より医療に適した辞書を使用することで正誤検証の探索精度が向上する可能性がある．

5　終わりに

本研究では大規模言語モデル（LLM）を用いた医療応用研究に着目し，医療プロセスの中でも AI による診断への応用可能性を探るため，複数の GPT モデルを用いた病名予測による差異検証や量的検証，ICD-10 コードを用いた類似度探索による正誤検証を行った．

結果，差異検証ではファインチューニングを行うことで，より正解病名に近い予測ができる傾向を確認し，量的検証ではプロンプトを変更することで N/A の数が減少し，平均コサイン類似度が上昇することを確認できた．正誤検証では「発汗異常」など予測病名と正解病名が一致しない場合においても ICD-10 コードを正しく推定可能なケースが存在することを確認した．

今後の展望としては以下の点が挙げられる．

- 複数入出力による予測病名精度の改善検証や ICD-10 コードの探索
- 医療用辞書を用いた ICD-10 コードの探索

参考文献

[1] 福本拓也, 坂根亜美, 村松俊平, 五十嵐正尚, 狩野芳伸, 荒牧英治, 堀口裕正, 奥村貴史. 新型コロナ感染症発生届の分析-記載における非効率と自然言語処理による解決への課題と展望-. 2023 年度人工知能学会全国大会（第 37 回）, 2023.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert:pre-training of deep bidirectional transformers for language understanding. **arXiv**, No. 1810.04805, 2019.

[3] 小野真子, 中西美和. 医療事故・ヒヤリハット報告に対する bert を用いた事故分析支援手法の提案. 日本人間工学会第 64 回大会, 59 巻, , 2023.

[4] 中尾睦宏. Ai，ict，vr を活用する未来に向けて. バイオフィードバック研究, 48 巻, 第 1 号, pp. 11–15, 2021.

[5] 野田昌生, 上野貴雄, 甲州亮太, 島田 Dias 茉莉, 伊藤真人, 矢本成恒, 吉崎智一, 野村章洋. 耳鼻咽喉科専門医試験における generative pretrained transformer の有効性に関する検討. 日本耳鼻咽喉科頭頸部外科学会会報, 126 巻, 11 号, pp. 1217–1223, 2023.

[6] 西谷実紘, 矢田竣太郎, 若宮翔子, 荒牧英治. 生成アプローチによる患者表現の標準化. 人工知能学会, 2021 巻, AIMED-011 号, pp. 5–01–5–07, 2021.

[7] 厚生労働省. 疾病、傷害及び死因の統計分類. https://www.mhlw.go.jp/toukei/sippei/ (accessed Dec.16,2024).

[8] Google colab, (2024 年 12 月 19 日閲覧). https://colab.research.google.com/?hl=ja.

[9] OPEN AI documentation Models, (2024 年 12 月 19 日閲覧). https://platform.openai.com/docs/models.

表 2　GPT モデルごとの差異検証の評価結果

入力テキスト	正解病名	GPT-3.5			GPT-4o-mini	
	標準病名	N	C	F	N	C
左上のお腹が痛い	上腹部痛	胃潰瘍	胃潰瘍	上腹痛	胃炎	胃炎
胸が痛い	胸痛	心臓疾患	N/A	前胸部痛	心筋梗塞	狭心症
たんがでる	喀痰	風邪	風邪	**喀痰**	風邪	風邪
あせだぐ	多汗症	N/A	N/A	発汗異常	熱中症	**多汗症**
トイレが近くて何度も夜に目が覚める	夜間頻尿症	尿路感染症	頻尿	**夜間頻尿症**	夜間頻尿	夜間頻尿
右白内障	白内障	N/A	N/A	**白内障**	N/A	N/A
下肢の膨張	下肢浮腫	静脈瘤	静脈瘤	四肢浮腫	心不全	静脈瘤
胸膜炎症	胸膜炎	N/A	N/A	**胸膜炎**	N/A	N/A
心音に通常の音以外がある	心雑音	心臓弁膜症	**心雑音**	**心雑音**	**心雑音**	**心雑音**
膵臓のできもの	膵腫大	膵臓にできる腫瘤	膵臓がん	**膵腫大**	膵臓にできる腫瘤	膵臓腫瘤
平均コサイン類似度		0.541	0490	0.951	0.637	0.689

表 3　プロンプト変化による予測結果の量的検証評価

	GPT-3.5(N)		GPT-4o-mini(N)	
	プロンプト変更なし	プロンプト変更あり	プロンプト変更なし	プロンプト変更あり
N/A	40/100	32/100	31/100	15/100
平均コサイン類似度	0.458	0.532	0.522	0.642

表 4　ICD-10 コードによる正誤検証の評価結果

入力テキスト	正解病名/ICD-10 コード			予測病名/ICD-10 コード		
	標準病名	中間分類	3 桁分類	予測病名	中間分類	3 桁分類
左上のお腹が痛い	上腹部痛	R10-R19	R10	上腹痛	S10-S19	**R10**
胸が痛い	胸痛	R00-R09	R07	前胸部痛	N40-N51	**R07**
たんがでる	喀痰	R00-R09	R09	**喀痰**	L10-L14	R05
あせだぐ	多汗症	R50-R69	R61	発汗異常	H60-H62	**R61**
トイレが近くて何度も夜に目が覚める	夜間頻尿症	R30-R39	R35	**夜間頻尿症**	N20-N23	R34
右白内障	白内障	H25-H28	H26	**白内障**	H40-H42	**H26**
下肢の膨張	下肢浮腫	R50-R69	R60	四肢浮腫	M60-M79	R59
胸膜炎症	胸膜炎	R00-R09	R09	**胸膜炎**	K65-K67	R07
心音に通常の音以外がある	心雑音	R00-R09	R01	**心雑音**	H80-H83	**R01**
膵臓のできもの	膵腫大	R90-R94	R93	**膵腫大**	B65-B83	R90