大規模言語モデルを用いた複数情報源統合による希少疾患診療支援システム

DBCLS,東京大学
吉桑実弥

DBCLS
高月照江

DBCLS
千葉啓和

DBCLS
藤原豊史

DBCLS, 東京大学
五斗進

Abstract

希少疾患の診断は、症例が少なく表現型も複雑なため困難である。診断支援のための従来の手法は、顔写真や表現型など、単一のデータを用いて予測することが主だったが、LLMによって複数のツールによる予測結果に加え、関連する症例情報、文献情報を統合させることで、より高い精度での疾患の予測が可能になることが期待されている[1]。本研究では、
1.GestaltMatcher[2]による顔写真解析
2.PubCaseFinder[3]による表現型に基づいての疾患予測
3.疾患名や表現型をクエリとしたWeb検索による情報収集
4. LLMによるZeroshot推論
を使用し、LLMによってweb上の情報と共にこれらの結果を統合することで、根拠と共に優先順位を付けた疑わしい疾患のリストを生成し,
各ツールの出力結果と比較することでその性能を評価する。

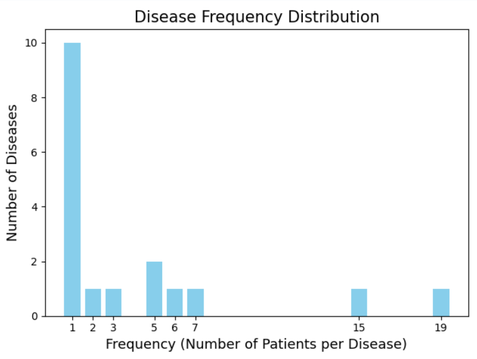
Data

この研究では、PhenopacketStore[4]から取得した症例情報の一部に引用論文の顔写真を紐づけ、それをテストデータとしてを使用した。
また、GestaltMathcerの訓練データとして使用された患者のデータはテストデータから排除した。
これらのデータから、患者の表現型、患者に見られなかったことが明記されている表現型、性別、顔写真のみを取得し、モデルへの入力として使用した。
PhenopacketStoreに含まれているOMIM ID及び、疾患名を用いて正解の判定を行った。

用いたデータの詳細

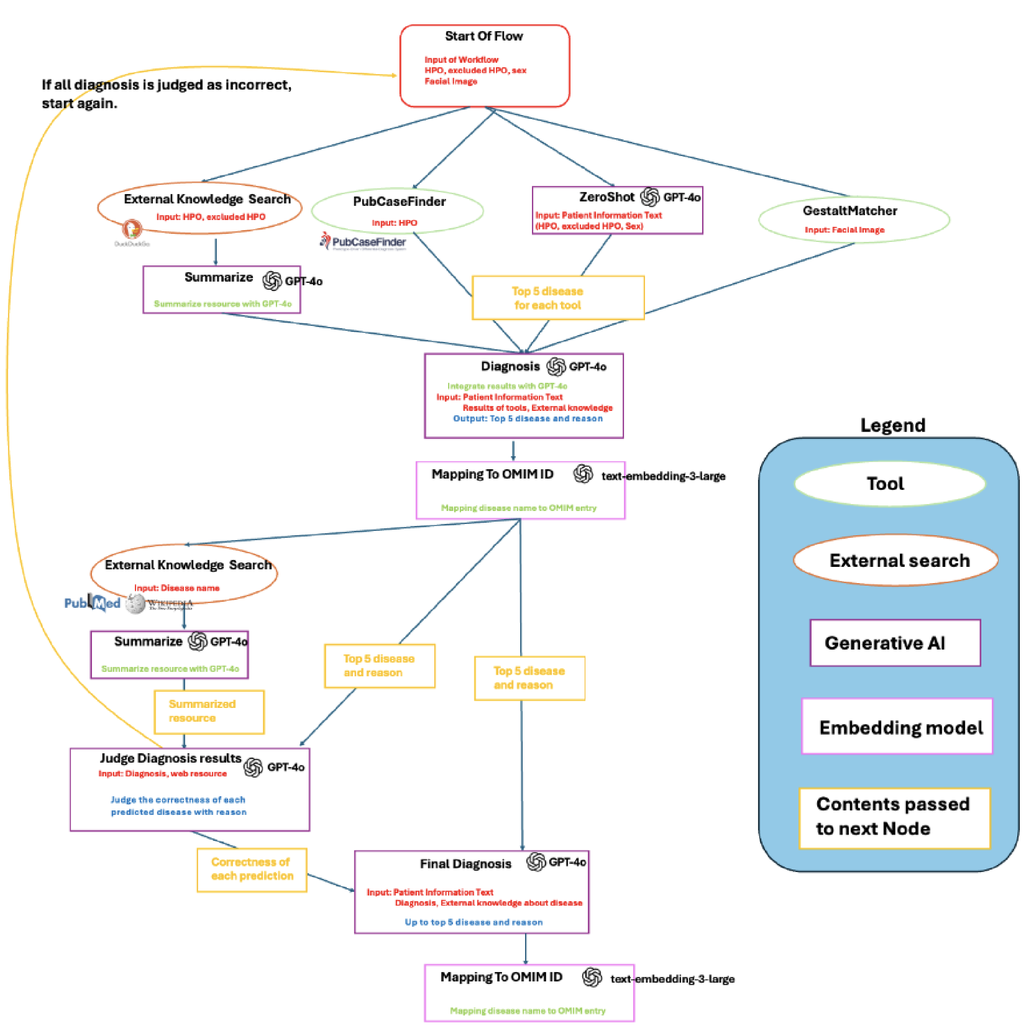
PhenopacketStore 0.1.25から取得したデータであり、その一部に引用元論文から取得した写真を紐付けた。
症例数: 72
疾患数: 18

疾患ごとの症例数の分布は右図の通り。
(横軸: 度数、縦軸:疾患数)



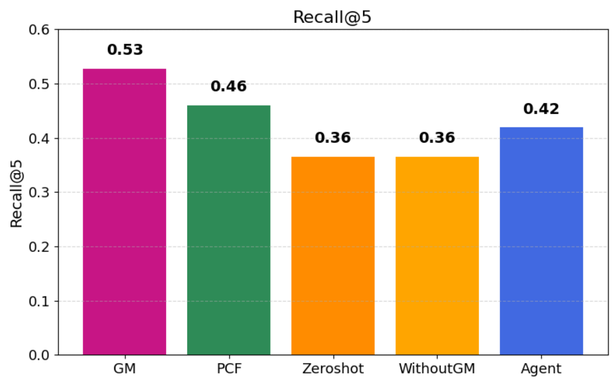
Method

エージェントのフローは下図に示す。このエージェントでは、GPT-4oが複数の情報源を統合し出力を行う。 まず、以下の4つのツールを実行し、それぞれ上位5つの疾患を予測する。
1. Web検索によって患者の表現型情報から関連する情報を検索する
2. PubCaseFinder によって患者の表現型情報から可能性の高い疾患を予測する
3. LLMによるゼロショット推論によって患者の表現型情報から可能性の高い疾患を予測する
4. GestaltMatcher DBによって患者の顔画像から可能性の高い疾患を予測する
次に、GPT-4oが患者情報とこれらの情報を統合し上位5疾患を含んだ一次予測を作成する。さらに、予測された疾患名を用いてPubMed等から情報を収集し、それに基づき予測を自己検証してその妥当性を判断する。その後これらの情報をもとに最終的な予測を生成する。 出力された疾患はOMIM IDにマッピングした。
このワークフローは、PythonのLangGraphで実装した。



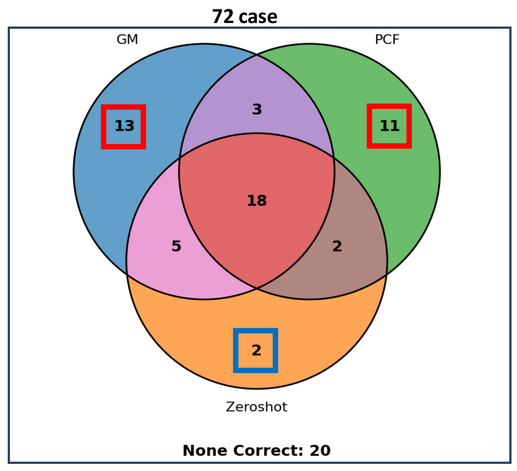
Result

本研究の評価では希少疾患のサブタイプを許容するため、OMIM IDが完全一致したものを”Match”、text-embedding-largeモデルでのcos類似度が0.7以上を”Similar”とし、それ以外は”Miss“とした。MatchかSimilarがモデルやツールによる予測上位5疾患に含まれている場合に正しく予測が行われているとし、各ツールについて正解したデータ数及びRecall@5を計算した。



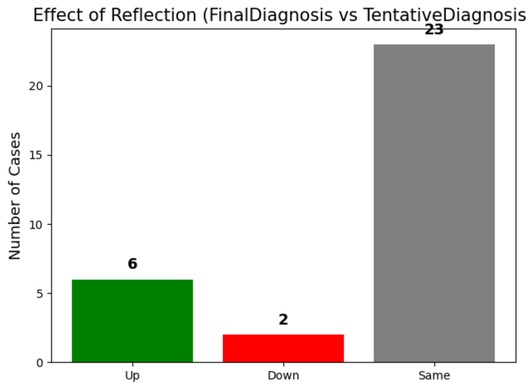
現在のモデルの性能はLLMによるZeroshot推論の性能よりも高い一方でGestaltMather(GM)、PubCaseFinder(PCF)の性能よりも低いという結果が得られた。
その原因として、それぞれのツールのみで正しく予測された疾患や複数のツールで上位3~5位に予測された正解の疾患があっても、LLMがTop 5の疾患をリストアップした時にそれが排除されてしまっているという問題が見られた。

Venn diagram of patients correctly predicted by each tool



左図はそれぞれの症例において、正解の疾患がどのツールによって上位5疾患の中に正しく予測されたかをベン図で表したものである。
GestaltMathcer(GM)とPubCaseFinder(PCF)のみが正しく予測できた症例はそれぞれ13症例と11症例存在している。
また、全72症例のうち、どのツールでも正解の疾患が正しく予測されなかった症例は20症例だった。

Effect of Reflection (FinalDiagnosis vs TentativeDiagnosis



左図は自己検証前後での出力の比較を行なったものである。予測の精度は
Match > Similar >Missとし、
これらが自己検証前後で変化しない場合には、正解の疾患の順位が高いほど精度が高いと判断する。
少なくともどちらかがMissでない31症例のうち、精度が向上したのは6症例、悪化したのは2症例だった。

Discussion & Future Work

各ツール及びモデルの評価をより正確に行うため、より多数のテストデータを収集、使用することが必要である。
また、現状のエージェントは性能が低く、ツールや情報をうまく統合できていないと考えられる。その原因として、
1. 各ツールで3~5位で予測された疾患や、単一のツールのみで正しく予測された疾患がLLMがはじめに統合を行う際に、上位5疾患のリストに含まれていない
2. 自己検証の際に正しい疾患がLLMによって誤っていると判断される、もしくは正解が含まれていないのに、LLMがそれらの疾患のどれかを正しいと判断してしまう

という問題が挙げられる。そのため、その解決策として、
1. 各ツールで出力される病名を統一し、表現の揺れを補正することでLLMが疾患を正しく認識できるようにする
2. 患者の表現型が多数存在するため、HPOベースでのweb検索で正しい情報を取得できていない。そのため、ベクトル検索やLLMにクエリ文を生成させるなどのアプローチを行う
3. 症例情報や医療文献情報などより多くの情報源を参照する
4. 使用するツールを増やし、現状のツールで予測できない疾患にも対応できるようにする
5. 医師や専門家に出力及びLLMが示した根拠についての意見を頂き、プロンプトやLLMの出力を調整する

などの改善を行う必要がある

Reference

1. Weike Zhao et al.「An Agentic System for Rare Disease Diagnosis with Traceable Reasoning」, arXiv:2506.20430, 2025年. https://arxiv.org/abs/2506.20430
2. Lesmann H et al. GestaltMatcher Database - A global reference for facial phenotypic variability in rare human diseases. medRxiv. 2024年10月8日:2023.06.06.23290887. doi:10.1101/2023.06.06.23290887. PMID: 37503210.
3. Fujiwara T et al. PubCaseFinder: A Case-Report-Based, Phenotype-Driven Differential-Diagnosis System for Rare Diseases. Am J Hum Genet. 2018 Sep 6;103(3):389-399. doi:10.1016/j.ajhg.2018.08.003. PMID: 30173820.
4. Danis D et al. A corpus of GA4GH Phenopackets: case-level phenotyping for genomic diagnostics and discovery. medRxiv [Preprint]. 2024 May 29:2024.05.29.24308104. doi: 10.1101/2024.05.29.24308104. PMID: 38854034.