大規模言語モデルを用いた読影レポートからの情報抽出：ChatGPT3.5、ChatGPT4 および Google Bard の比較

Information Extraction from Radiology Reports Using Large Language Models: A Comparison of ChatGPT3.5, ChatGPT4, and Google Bard

土橋 大樹[1, 2]、平田 健司*[2, 3, 4, 5, 6]、渡邊 史郎[2, 3, 6]、竹中 淳規[2, 3]、若林 直人[2, 3]、木村 理奈[1, 2]、坂本 圭太[1, 2, 6]、工藤 與亮[1, 2, 4, 5, 6]

Hiroki Dobashi, Kenji Hirata*, Shiro Watanabe, Junki Takenaka, Naoto Wakabayashi, Rina Kimura, Keita Sakamoto, Kohsuke Kudo

1 北海道大学病院放射線診断科、2 北海道大学大学院医学研究院放射線科学分野画像診断学教室、3 北海道大学病院核医学診療科、4 北海道大学大学院医学研究院医療 AI 教育研究分野、5 北海道大学病院医療 AI 研究開発センター、6 北海道大学グローバル生体医理工学分野

1 Department of Diagnostic and Interventional Radiology, Hokkaido University Hospital, Sapporo, Japan
2 Department of Diagnostic Imaging, Faculty of Medicine, Sapporo, Japan
3 Department of Nuclear Medicine, Hokkaido University Hospital, Sapporo, Japan
4 Division of Medical AI Education and Research, Faculty of Medicine, Hokkaido University, Sapporo, Japan
5 Medical AI Research and Development Center, Hokkaido University Hospital, Sapporo, Japan
6 Global Center for Biomedical Science and Engineering, Faculty of Medicine, Hokkaido University, Sapporo, Japan

2023年12月8日論文受領、修正依頼2024年1月9日、最終受領日2024年1月18日

【要旨】【目的】本研究の目的は、自由文記載形式で作成された読影レポートから特定の情報を抽出し、二次利用可能な形式に変換するタスクを、大規模言語モデル（Large Language Models、LLM）によって遂行可能かどうかを検証することである。
【対象・方法】肺癌CTの読影レポートからなる公開データセット（読影医9名による5症例、45件分の画像診断レポート）を用いた。ChatGPT（ver.4と3.5）とGoogle Bardを使用して、良悪性等の9項目の抽出を試み、放射線科医がその精度を評価した。
【結果】ChatGPT4は高い精度で情報を抽出し、ほぼ全ての項目で優れた結果を示した。ChatGPT3.5も同様に良好な結果を得たが、Google Bardはやや劣る結果であった。全体的にLLMによる情報抽出は有用性が高いことが示された。
【結論】LLMを使用した読影レポートの情報抽出は、多くの場合で良好な精度を示し、二次利用に貢献しうることが明らかになった。今後は他の疾患に対する適用や精度向上のための詳細なプロンプトの作成が必要である。
【責任著者の連絡先】平田 健司　北海道大学大学院 医学研究院 画像診断学教室
〒060-8638　札幌市北区北15条西7丁目東南棟1階 ES1・106-2　TEL：011-706-7779　FAX：011-706-7408
EMAIL：khirata@med.hokudai.ac.jp
【キーワード】Large language model, Natural language processing, ChatGPT, Google Bard, Radiology report
【利益相反】著者の平田は、過去3年以内にGEヘルスケアジャパン株式会社および株式会社バスクリンより研究費を受け入れています。
【グラント】本研究の内容の一部はJSPS科研費JP23K07150の助成を受けたものです。

【Abstract】 Objective: This study aimed to evaluate the feasibility of using large language models (LLMs) to extract specific information from freely written radiology reports on CT of lung nodules, and to transform it into a format suitable for secondary use.
Methods: A publicly available dataset of radiology reports of CT with lung nodules was utilized. ChatGPT4, ChatGPT3.5, and Google Bard were employed to extract nine designated items, and the extraction performance was assessed by a radiologist.
Results: ChatGPT4 demonstrated high accuracy in information extraction, delivering superior results across nearly all items. ChatGPT3.5 also performed well, but Google Bard showed slightly inferior outcomes. Overall, the study indicated that the use of LLMs for information extraction is highly beneficial.
Conclusion: The extraction of information from radiology reports using LLMs was accurate for the currently used dataset. Future researches should focus on expanding the application to other diseases and improving accuracy through the development of sophisticated prompts.

【背景】

日本で放射線診断医の主たる業務は読影レポートの作成である。読影レポートは、画像を必要十分な形に要約したデータの集合体と考えられ、検査をオーダーした主治医あるいは担当医が画像所見を確認して診断・治療に活かすことを第一の目的として作成される。読影レポートにより、主治医は自らの専門領域以外の画像所見を知ることができ、また気づかなかった画像所見を知ることもできる。こうした意味において読影レポートはきわめて有用である。しかし、読影レポートには画像診断医の「知恵」が結集しており、その価値は当該患者の診察時以外にも活かされるべきである。本論文において、著者らはこれを読影レポートの二次利用と呼ぶ。実際には、本邦において読影レポートが二次利用されている例はあまりない[1]。理由は、読影レポートの多くが自由文記載形式で作成され、情報の抽出が難航するためである。自由文記載形式とは**図１**に挙げられるような形式であり、自然な日本語で記載され、記載項目や順序に明確なルールは定めないものである。自由文記載形式は、多様な症例に対応可能で、必要最小限の文章量で、情報に重要性の強弱をつけて読者に伝えることを可能にしている。

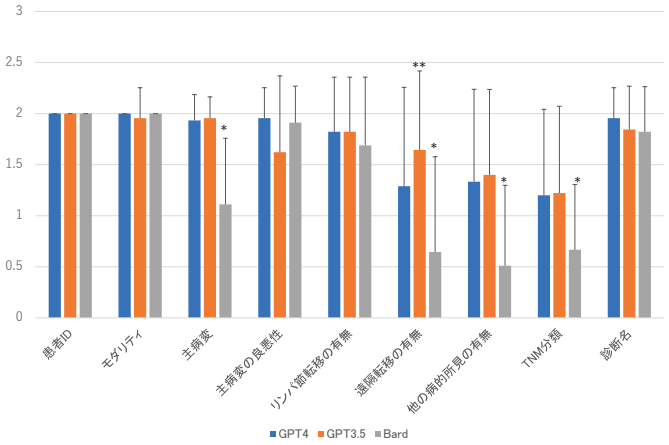

図１　項目ごとのLLM間の性能比較
対応のある t検定による検定結果。
* 他 2 者に比べて有意に低値である。
** 他 2 者に比べて有意に高値である。

一方、読影レポートの記載様式には構造化記載形式が存在する。構造化記載形式では、見出しや記載欄による記載様式の統一や網羅的かつ一貫した記載項目の構成、統一用語の使用が行われる。テンプレートに沿って与えられた選択肢からの選択もしくは短い単語で記載する形式である。記載様式や用語が統一されているため、構造化記載形式は情報抽出、二次利用に向いている。しかし、読影レポートの構造化には肯定的な意見だけではなく、例えば、生産性の低下や変化への抵抗、複雑な症例への不適合性といった否定的な意見[2]もあり、置き換わっていくとしても時間を要するだろう。

読影レポートの二次利用が可能になれば、疾患のビッグデータを作成し、臨床研究に利用し更なる医学の発展に寄与しうる。また、両者をシームレスに自動変換することができれば、日々蓄積されるデータがそのままビッグデータ解析に使用できることになり、現代のAI技術の発達と相まって、その有用性は大きい。これまでも自由文記載形式の読影レポートからの情報抽出は試みられている[3]ものの、特別な訓練や専用のプログラムを作成する必要があり、多大な労力を要するものとなっていた。そのため、より簡便な方法の確立が待たれていた。

大規模言語モデル（Large Language Models、LLM）は、非常に巨大なデータセット（インターネット上に存在するすべての文章が対象）とディープラーニング技術を用いて構築された言語モデルである。「大規模」は従来の自然言語モデルと比べ、「計算量」「データ量」「パラメータ数」を大幅に増やして構築されていることに由来する。大規模言語モデルは、人間に近い流暢な会話が可能であり、自然言語を用いたさまざまな処理を高精度で行うことを可能にした。LLMは、テキスト入力に対して人間のような回答を生成する能力を持つ、いわゆる生成系AIの一種である。従来のAIが決められた行為の自動化が目的であるのに対し、生成系AIはデータのパターンや関係を学習し、新しいコンテンツを生成することを目的としている[4]。ユーザーはプロンプトと呼ばれる指示文を与えることにより、LLMをチューニングすることができる。

大規模言語モデルの中で最も有名なものはOpen AIのChatGPTである。2022年11月に発表されたChatGPTは大規模言語モデル（LLM）の1つである。人間の会話に近い文章を高精度に生成できることから、Nature誌にも取り上げられ、今後議論を要する部分がありつつも科学への応用が期待されている[5, 6]。同様に、Google社もGoogle Bardを発表し、クラウドサービスとして無料で提供している。

本研究の目的は、LLMが長文から重要情報を抽出することも可能であるとされていることを踏まえ、有料・無料のChatGPTおよびGoogle Bardの3つのLLMを利用し、自由文記載形式で作成された読影レポートから複数の項目を抽出し、二次利用可能な形式に変換できるかどうかを検討することである。

【対象と方法】

対象データ

現在のChatGPTおよびGoogle Bardの仕様では、インターネット上のクラウド・サーバーに入力データをアップロードする必要がある。実臨床の読影レポートは個人情報保護法における要配慮個人情報を完全に除外することが困難であり、

解析のためにクラウド・サーバーにアップロードすることは法的・倫理的に認められないと考えた。そこで代替策として、インターネット上に公開されている肺癌の読影レポートデータセット[7]を使用した。この肺癌の読影レポートデータセットはRadiopaediaより引用された5症例の肺癌CT画像に対して、9人の日本人放射線科医が日本語の自由文記載形式で作成した全45本の読影レポートである。レポートの形式は所見のみが記述されているが、中にはimpressionが追記されているものもあった。**表1**に使用したデータセットの読影レポートの1例を示す。

表 1　自由文記載形式の一例

所見	ID：3-8 左舌区S4に長径28mm大の境界明瞭な不整形陰影が認められます。辺縁にはspiculaや胸膜陥入像を伴っており、T1bの肺癌として矛盾しません。強く胸膜浸潤を疑う所見は指摘できません。 右肺下葉には不整形結節が認められます。炎症性変化かもしれませんが、あわせて経過観察をお願いします。 肺尖部には陳旧性炎症性変化と思われる胸膜肥厚を認めます。 その他肺野に粗大な腫瘤や浸潤影は指摘できません。 有意な縦隔リンパ節腫大や胸水認めません。

情報抽出

3種のLLM（ChatGPT4、ChatGPT3.5、Google Bard）を使用した。ChatGPT4は有料、ChatGPT3.5とGoogle Bardは無料である。

ChatGPT4とChatGPT3.5では性能面で大きな差がみられる。Open AIのテクニカルレポート[8]によれば、米国司法試験においてChatGPT4は上位10%のスコアを獲得した一方で、ChatGPT3.5では下位10%であった。また、ベンチマークテストにおいても、ChatGPT4は、ChatGPT3.5を含む従来の大規模言語モデルを上回る結果を残すなど、大きな向上がみられた。さらに誤った情報の生成がChatGPT3.5に比べて減少しており、多くの点でChatGPT4はChatGPT3.5を上回っていると考えられる。

各LLMに前述の公開読影レポートとプロンプトを与え、情報を抽出させた。抽出結果を、後述する評価基準に則り、放射線科医1人（医師3年目）が評価した。

読影レポートからの抽出項目は二次利用のしやすさを考慮し、「ID」、「モダリティ」、「主病変（の位置や画像所見）」、「主病変の良悪性」、「リンパ節転移の有無」、「遠隔転移の有無」、「他の病的所見の有無」、「TNM分類」、「診断名」の9項目とした。

実際に使用したプロンプトを**表2**に、ChatGPTとGoogle Bardの抽出結果の一例をそれぞれ**表3**、**表4**に示す。今回は初期検討ということもあり、プロンプトに詳細な指示は含めず、シンプルな構成でどこまで抽出可能かを確認した。

解析は2023年6月5日から6月18日にかけて行った。

表 2　使用したプロンプト

下記のフォーマットに沿って，単語で要約してください． ---フォーマット--- ID： モダリティ： 主病変： 主病変の良悪性： リンパ節転移の有無： 遠隔転移の有無： 他の病的所見の有無： TNM分類： 診断名： （以下にレポート本文をペーストした）

表 3　ChatGPTの出力例と評価例

ChatGPT4の出力例とその評価例	ID：3-8　2点 モダリティ：CT　2点 主病変：左舌区S4、径28mm、不整形陰影　2点 主病変の良悪性：悪性（肺癌）2点 リンパ節転移の有無：無し　2点 遠隔転移の有無：無し　0点 他の病的所見の有無：右肺下葉の不整形結節、左肺尖部の胸膜肥厚　2点 TNM分類：T1b　1点 診断名：肺癌　2点
ChatGPT3.5の出力例とその評価例	ID：3-8　2点 モダリティ：CT　2点 主病変：左舌区S4 の境界明瞭な不整形陰影　2点 主病変の良悪性：T1b の肺癌として矛盾しません　1点 リンパ節転移の有無：有意な縦隔リンパ節腫大や胸水は認めません　1点 遠隔転移の有無：不明　2点 他の病的所見の有無：右肺下葉に不整形結節あり。経過観察が必要です。胸膜浸潤を疑う所見は指摘できません。2点 TNM分類：T1b、N不明、M不明　2点 診断名：左舌区のT1bの肺癌　2点

表4　Google Bardの出力例

ID	モダリティ	主病変	主病変の良悪性	リンパ節転移の有無	遠隔転移の有無	他の病的所見の有無	TNM分類	診断名
id：3-8	CT	※	悪性	なし	なし	なし	T1b	肺癌
2点	2点	0点	2点	2点	0点	0点	1点	2点

※主病変のテキスト
左舌区S4に長径28mm大の境界明瞭な不整形陰影が認められます。辺縁にはspiculaや胸膜陥入像を伴っており、T1bの肺癌として矛盾しません。強く胸膜浸潤を疑う所見は指摘できません。
Google Bardの出力は一括の表形式で出力されたため、一部を抜粋した。

評価基準

各項目の抽出結果を0-2点で採点した。抽出された情報が誤っていた場合には0点とした。抽出された情報が不十分であった場合は1点とした。例えば、読影レポートの原文が「右下葉S6 に12mm 大の比較的境界明瞭なpure GGN」であった場合、主病変に関する抽出結果が「GGN」のみであった場合は、抽出が不十分として1点とした。

各項目の具体的な評価基準は以下の通りである。「ID」、「モダリティ」、「主病変」、「主病変の良悪性」は抽出の正確性に応じて0-2点を与えた。正解の場合は2点、一部不正解の場合は1点、不正解の場合は0点を与えた。「リンパ節転移の有無」、「遠隔転移の有無」は、レポートに転移の有無が明示されていない場合は抽出結果に不明とあれば満点としたが、転移なしとした場合には0点とした。一部不正解であれば1点を与えた。「TNM分類」はTNM分類（日本肺癌学会編：肺癌取扱い規約第8版）[9]に沿って採点し、基準に沿っていない場合は0点とした。一部のみ不正解であれば１点を与えた。「診断名」は、レポートに診断名やimpressionの記載がある場合は、それを正確に抽出出来ているかで判定した。元々のレポートに記載が無い場合は，肺癌と抽出出来ていれば2点とした。**表3**はChatGPT4を評価した一例である。

統計解析

3種のLLMが取得した点数の差に対して対応のあるt検定、Holm補正（p値を大きいものから順位付けし、p値に順位を乗じたものと有意水準を比較した）を用いて検討した。p値が0.05未満である場合に統計学的に有意とした。統計解析ソフトとして、Microsoft Excel for Macを使用した。

【結果】

3種のLLMを用いて全45レポートに対して情報抽出を実施した。項目毎の点数をLLM毎にまとめて比較した結果が**図1**である。p値は**表5**に、平均と標準偏差は**表6**に記載した。**表5**においてt検定では値が完全に一致した場合、ゼロによる除算が生じるため、計算不能な箇所はN/Aと表記した。

「ID」については全てのLLMが満点を獲得した。「モダリ

表5　項目ごとのLLM間性能比較

	ID	モダリティ	主病変	主病変の良悪性	リンパ節転移の有無	遠隔転移の有無	他の病的所見の有無	TNM 分類	診断名
GPT4-GPT3.5	N/A	0.323	0.323	0.029	1.000	0.031	0.705	0.860	0.071
GPT3.5-Bard	N/A	0.323	P<0.001	0.062	0.057	P<0.001	P<0.001	0.004	0.800
Bard-GPT4	N/A	N/A	P<0.001	0.160	0.135	P<0.001	P<0.001	0.005	0.114

有意水準を 0.05 に統一するために p値に順位を乗じて補正を実施した（Holm 補正）。

表6　各LLMの項目ごとの平均値と標準偏差

	ID	モダリティ	主病変	主病変の良悪性	リンパ節転移の有無	遠隔転移の有無	他の病的所見の有無	TNM分類	診断名
GPT4	2.000 ± 0.000	2.000 ± 0.000	1.933 ± 0.252	1.956 ± 0.298	1.822 ± 0.535	1.289 ± 0.968	1.333 ± 0.905	1.200 ± 0.842	1.956 ± 0.298
GPT3.5	2.000 ± 0.000	1.956 ± 0.298	1.956 ± 0.208	1.622 ± 0.747	1.822 ± 0.535	1.644 ± 0.773	1.400 ± 0.837	1.222 ± 0.850	1.844 ± 0.424
Bard	2.000 ± 0.000	2.000 ± 0.000	1.111 ± 0.647	1.911 ± 0.358	1.689 ± 0.668	0.644 ± 0.933	0.511 ± 0.787	0.667 ± 0.640	1.822 ± 0.442

ティ」はほぼ全ての読影レポートに対してCTと正答したが、少数例ながら未抽出や胸部レントゲンと抽出したものもあった。「主病変」は、ChatGPT4はほぼ正確に抽出結果が得られたが、複数の病変があり、どちらか一方しか答えられず減点したものもあった。「遠隔転移の有無」「TNM分類」ではChatGPT3.5の得点がChatGPT4を上回っていた。Google Bardは独自のTNM分類評価や情報抽出が単なる複数行の抜き出しだったものもみられ、全体的に点数は低めであった。

各LLMの平均点は、ChatGPT4が1.721点、ChatGPT3.5が1.719点、Google Bardが1.373点であった。

各項目の取得した点数の差に対して対応のあるt検定、Holm補正を実施した。Google Bardは「主病変」「遠隔転移の有無」「他の病的所見の有無」「TNM分類」の項目で他2者に比べて有意に低値であった。

【考察】

今回の結果から、LLMを用いた自由文記載形式の読影レポートからの情報抽出は有用性が高いことが示唆された。項目ごとに見ると、「ID」と「モダリティ」はほぼ正確に抽出できたが、これらの情報は文中に明示されていることが多いため、LLMにとって容易なタスクであったと考えられる。「主病変」「主病変の良悪性」「リンパ節転移」「診断名」は、若干の減点があったものの概ね正確に抽出できており、これらも文中に比較的明確に記載されていたためと考えられる。ただし、症例によっては良悪性を明示することが難しいことも少なくなく、そのような場合にはレポートにも読影者間の大きな表記ゆれが生じることが予想される。その場合にLLMが正しく情報抽出できるかどうかは今後の検討課題である。「遠隔転移の有無」「TNM分類」では、精度の低下が目立ち、興味深いことにChatGPT3.5の得点がChatGPT4を上回った。その理由は複数の症例で遠隔転移の有無に関する記載がなかった際に、ChatGPT4は大半の場合で「なし」と評価した一方で、ChatGPT3.5は「不明」と評価したことに起因すると考えられる。

3種類のLLMの比較では、ChatGPT4とChatGPT3.5には大きな差がなく、Google Bardがやや劣る結果になった。ChatGPT4はChatGPT3.5に比べて多くのパラメーターを持つとされるが、パラメーター数は公開されていない。先行してリリースされたChatGPT系がより細かいファインチューニングを受けている可能性があるが、詳細は不明である。ただし、いずれのモデルも短期間でアップデートされており、今回の実験を実施した2023年6月時点から現在までに性能が変化している可能性がある。

LLMによるデータベース化は、過去の膨大な読影レポートをビッグデータ解析することを可能とし、様々な研究の進展に寄与し得るだろう。しかし、前述した通り、実際の読影レポートは個人情報を含むため、現行のChatGPTやGoogle Bardにアップロードすることはできない。どのような解決法が考えられるだろうか。1つはローカルLLMと呼ばれる方法である。すなわち、ChatGPTと同等の処理を施設内のワークステーションに構築すれば、データをインターネット上にアップロードせず処理が可能になる。Meta社が公開するLLaMAはオープンソースのLLMであり、無料でローカルのコンピューターにインストールすることができる（https://ai.meta.com/llama/）。しかし、インストール後のファインチューニングやメンテナンスが必ずしも容易とはいえない。もう1つの方法は、マイクロソフトのAzure OpenAI Service（AOS）を介してセキュアにChatGPTを利用する方法である。AOSは厚生労働省、経済産業省、総務省の3省が提示する医療情報関連ガイドラインに合致した運用を可能とし、薬歴入力支援システムに応用されている[10]。AOSであれば最新にアップデートされたChatGPTが使用可能である。OpenAIの開発速度は目を見張るものがあり、今後はこうしたセキュアなクラウドシステムの利用が主流になってくる可能性がある。

今回のLLMの抽出結果をより高めるためには、LLM自体の向上のほかに、読影レポートの表現方法を統一するというアイデアもある。しかし、これは一部構造化記載形式の手法を取り入れ、現在の自由文記載形式と異なる点には注意が必要であるが、構造化記載形式に完全に変更するよりは幾分現実的な変更と言えるかもしれない。

本研究においてはいくつかの限界がある。使用した公開読影レポートの記載に一部誤った内容が含まれており、レポート間の精度が不均一であったために評価内容に差が出た可能性が挙げられる。また、同一のLLMに同時期に同様の内容を入力したとしても、異なる出力結果が得られた可能性があり、結果の安定性・再現性については保証が出来ない。また、本実験では極めて単純なプロンプトを用いたが、より詳細なプロンプトを与えることでさらに正確な結果が得られた可能性がある。

結論

LLMに単純なプロンプトを用いて、自由文記載形式の読影レポートから複数項目の抽出を試みたところ、概ね良好な精度で抽出できることが明らかになった。今後は肺癌以外の疾患についても検討し、同様の情報抽出が適用可能か、さらに詳細なプロンプトを用意し精度向上が可能か模索していくことが必要である。

謝辞

本研究を実施するにあたり、研究方法、学会発表、論文作成

の各段階でご意見をいただいた画像診断学教室の先生方に深謝いたします。また読影レポートを公開いただいた東京大学、奈良先端科学技術大学院大学の諸先生方に深謝いたします。

引用文献

1. 武田 理宏, 真鍋 史朗, 松村 泰志, 電子カルテデータ二次利用の現状と課題, 生体医工学, 2017：55巻, 4号 : 151-158.
2. Rocha DM, Brasil LM, Lamas JM, Luz GVS, Bacelar SS, Evidence of the benefits, advantages and potentialities of the structured radiological report: An integrative review : Artif Intell Med, 2020 : 102 : 101770.
3. Sugimoto K, Takeda T, Oh JH, Wada S, Konishi S, Yamahata A, Manabe S, Tomiyama N, Matsunaga T, Nakanishi K, Matsumura Y. Extracting clinical terms from radiology reports with deep learning : J Biomed Inform, 2021 : 116 : 103729.
4. https://www.nri.com/jp/knowledge/glossary/lst/sa/generative_ai　11月6日　22時閲覧
5. Stokel-Walker C, Van Noorden R, What ChatGPT and generative AI mean for science : Nature, 2023 : 614(7947) : 214-216.
6. van Dis EAM, Bollen J, Zuidema W, van Rooij R, Bockting CL, ChatGPT: five priorities for research : Nature : 2023 : 614(7947) : 224-226.
7. Nakamura Y, Hanaoka S, Nomura Y, Hayashi N, Abe O, Yada S, Wakamiya S, Aramaki E, Clinical Comparable Corpus Describing the Same Subjects with Different Expressions: Stud Health Technol Inform : 2022 : 290 : 253-257.
8. Open AI, GPT-4 Technical Report : 2023, https://arxiv.org/abs/2303.08774,　2024年1月10日閲覧
9. James D Brierley, Mary K Gospodarowicz, Christian Wittekind, TNM悪性腫瘍の分類　第8版　日本語版：金原出版：2017：105-111.
10. https://prtimes.jp/main/html/rd/p/000000035.000107062.html、2023年11月28日閲覧