# 診療テキストからの必要な検査項目の予測

榎原芽美[1] 柴田大作[2] 辻川剛範[2] 宇野裕[2]
北出祐[2] 河添悦昌[1] 大江和彦[1] 久保雅洋[2]
[1] 東京大学大学院 医学系研究科
[2] 日本電気株式会社 バイオメトリクス研究所
m-ebara@m.u-tokyo.ac.jp
{daisaku-shibata, tujikawa, yutaka_uno, t-kitade, masahirokubo}@nec.com

## 概要

Large Language Models（LLM）は，その高い自然言語処理能力からさまざまな領域で注目されており，医学領域においても高い関心が寄せられている．本研究では、実際の医療現場において医療従事者が行うタスクの一つである初診時の診療テキストからの検査項目の予測を行う．LLM の精度と医療従事者の精度を比較した結果，全ての訓練データで Fine-tuning した Bidirectional Encoder Representations from Transformers の精度が最も高いことが確認された．また，Zero-Shot や Few-Shot で学習した LLM の精度は医療従事者による精度と遜色ない値であることが明らかとなり，LLM の有効性が示唆された．

## 1 はじめに

症例報告における症例記述は，経過を日付順に記載することが一般的であり，性別や年齢などの患者基本情報，主訴，現病歴，既往歴や現症などと言った初診時に取得される情報から始まり，続いて実施された検査の結果，治療内容や症状の変化などの臨床経過を記載することが一般的である．基本的にはこの初診時に収集される情報を用いて，医師が必要な検査項目の選択を行うが，これはしばしば医師の経験や知識に依存し，電子カルテやオーダリングシステムを導入している病院では，医師が多岐に渡る検査項目の中から必要な検査をシステム上から選択することになるため，必要な検査を過不足なく選択する業務は，医師の負担 [1] となっている．

そこで本研究では，Large Language Models（LLM）を用いた初診時の問診情報から必要な検査項目を自動で予測するタスクに取り組む（図 1）．日本語の症例報告テキストを初診時問診で得られる情報について記載したと考えることができる前半部分と，その後の臨床経過を記載した後半に分割し，後半部分のテキストに記載されている検査所見に基づいて実施された検査項目のアノテーションを行う．そして必要な検査項目を，前半部分のテキストから Bidirectional Encoder Representations from Transformers (BERT) や Generative Pre-trained Transformer 4 (GPT-4) などの LLM を用いて予測することで，それらの性能を評価する．また医療従事者の予測精度とそれらを比較することで LLM の有効性について調査を行う．

## 2 関連研究

医療領域における臨床タスクに対して LLM を応用した研究がいくつか報告されている．例えば，実際の患者情報と遺伝情報を元にしたテキストから，特定の治療薬に対する反応性があるかを二値分類で予測するタスクにおける LLM の性能評価を行う研究や [2]，患者からの健康相談に関する質問テキストに対して回答文を作成するタスクについて，医師の回答文と LLM の生成文を，情報の質と共感性の側面から比較評価する研究 [3] などが報告されている．医療分野における LLM の適用は研究段階であるが [4]，これを更に推し進めるためには LLM が実際の臨床現場で想定されるタスクに対してどの程度有効であるのかを明らかにする必要がある．

## 3 実験

### 3.1 実験材料

コーパスとして，難病・希少疾患を対象とした日本語の症例報告テキスト 183 件から構成される iCorpus[5] を使用した．このコーパスは，厚労省指定難病名をキーワードに検索して得た症例報告テキストを使用していることから，特定の診療科や疾患


This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

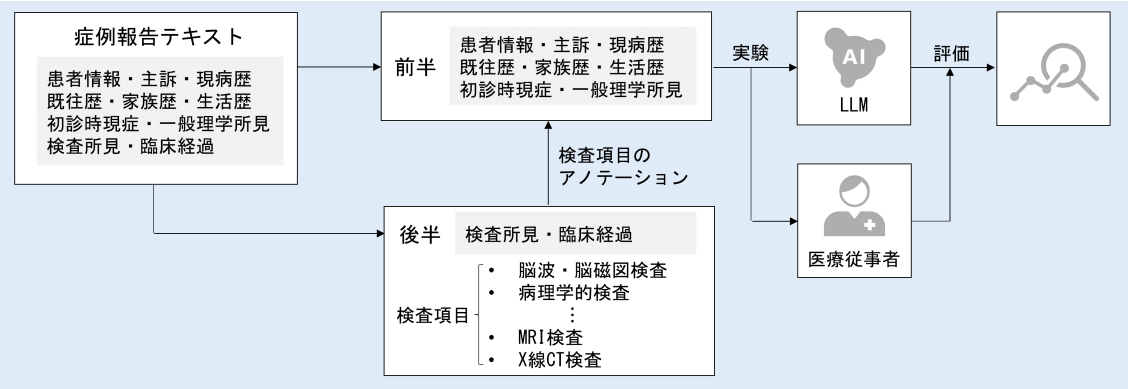


**図 1**　本研究の概要

領域に限定されない幅広い症状や所見が記載されている．（図 2）

症例は専門領域ごとに実施されている検査が異なると考えられるため，コーパスを掲載雑誌ごとに層別化した上で，訓練データとテストデータを 8:2 の割合で分割した．そして各コーパスを初診時問診までの段階で得られる情報のみを含むと考えられる前半部分と，検査結果と臨床経過を含む後半部分に分割し，後半部分に記載された検査結果を参考に臨床検査マスター [6] と画像検査マスター [7] を用いて正解ラベルのアノテーションを実施した．ここで 183 件のうち，初診時問診で得られる情報についての記載がないものや 1 ファイルにつき複数症例について記載されていたもの，前半部分と後半部分への分割が困難であった 11 ファイルを実験材料から除外し，最終的に 172 ファイルを使用した．訓練データとテストデータにおけるラベルとその出現頻度を表 1 に示す．なお，正解ラベルのアノテーションは第一著者が担当したが，アノテーションの前処理として必要なテストデータの前半部分と後半部分への分割は第二著者により実施された．最初に第一著者が訓練データ全体を参考に正解ラベルを決定し，アノテーションを実施した．続いてテストデータの前半部分から必要な検査項目を予測（医療従事者の予測）し，その後テストデータの後半部分を参考に正解ラベルのアノテーションを行った．

## 3.2　モデル

LLM として，OpenAI から提供されている GPT-4，ELYZA から公開されている elyza/ELYZA-japanese-Llama-2-7b-fast-instruct（ELYZA），東北大学から公開されている bertbase-japanese-whole-word-masking（以下 東北大 BERT）と NII から公開されている

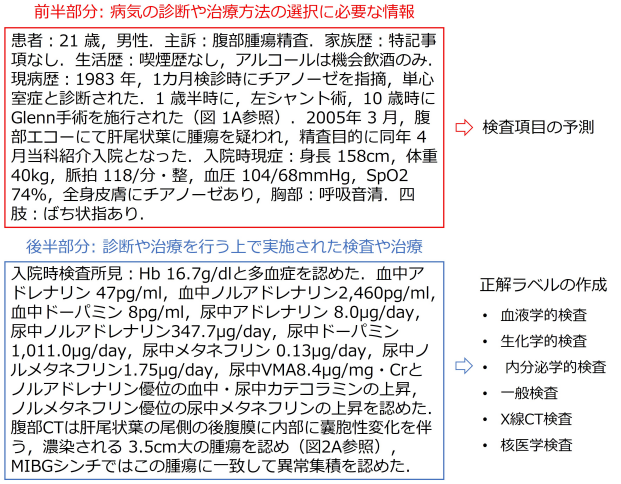


**図 2**　実験材料の詳細：前半部分を用いて必要な検査項目を予測し，後半部分を用いて正解ラベルのアノテーションを行う．

alabnii/jmedroberta-base-sentencepiece（JMedRoBERTa）を使用した．各 LLM の詳細は A 付録（表 4）に示す．

## 3.3　実験設定

LLM ごとの学習方法を表 2 に示す．GPT-4 は Zero-shot，One-shot と Few-shot Learning の 3 種類の方法で実験を行い，ELYZA については GPT-4 と同様の実験に加えて全訓練データで Fine-tuning する実験を追加し，合計で 4 種類の方法で実験を行った．ELYZA と GPT-4 に使用したプロンプトの詳細については A 付録（表 5）に記載した．東北大 BERT と JMedRoBERTa は Fine-tuning を行った際の精度，One-shot と Few-shot Learning の 3 種類の方法で実験を行った．加えて，LLM の精度と比較するため，日本の医療国家資格を有する著者の一人による分類実験も実施した．なお評価指標には，Micro-Precision,


This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

**表 1**　データセットにおける各ラベルの出現頻度

| ラベル | 訓練データ | テストデータ |
|---|---|---|
| X 線単純撮影 | 56 | 13 |
| X 線透視・造影 | 17 | 3 |
| X 線血管撮影 | 25 | 3 |
| X 線 CT 検査 | 72 | 16 |
| MRI 検査 | 49 | 16 |
| 核医学検査 | 37 | 5 |
| 超音波検査 | 61 | 14 |
| 臨床一般検査 | 42 | 13 |
| 血液学的検査 | 131 | 33 |
| 生化学的検査 | 134 | 33 |
| 内分泌学的検査 | 45 | 9 |
| 免疫学的検査 | 123 | 31 |
| 微生物学的検査 | 14 | 2 |
| 病理学的検査 | 70 | 15 |
| 脳波・脳磁図検査 | 13 | 2 |
| 心電図検査 | 42 | 10 |

**表 2**　LLM ごとの学習方法

| モデル | Fine-tuning | Zero-Shot | One-Shot | Few-Shot |
|---|---|---|---|---|
| 東北大 BERT | ✓ | | ✓ | ✓ |
| JMedRoBERTa | ✓ | | ✓ | ✓ |
| ELYZA | ✓ | ✓ | ✓ | ✓ |
| GPT-4 | | ✓ | ✓ | ✓ |

Micro-Recall，Micro-F1（以下, 単に Precision，Recall，F1）を使用した.

#### 3.3.1　Fine-tuning による実験設定

学習時のバッチサイズは 1, Epoch 数は 20, optimizer には AdamW を使用し，学習率は 3e-5, 最大系列長は 512 トークン，損失関数にはロジット付きバイナリ交差エントロピーを用いた．シード値を変更し、5 回評価を行ったそれぞれの評価値の平均を最終的な評価とした.

#### 3.3.2　Zero-Shot Learning による実験設定

予測すべき検査項目の候補に加えて，入力文を連結したものをプロンプトとし，それに続く文字列として予測検査を出力した．使用したプロンプトの例を A 付録（表 5）に示す．表中の input の部分に入力文を挿入する．出力されたテキストから，予測すべきラベル候補のうちの各ラベルの文字列と完全一致したものを予測ラベルとし，評価を行った.

#### 3.3.3　One-Shot Learning による実験設定

BERT では，各ラベルが少なくとも 1 例となるよう訓練データから 5 件抽出したデータで学習し予測を行った. 本来であれば，各ラベルが 1 例ずつとなるように訓練データを調整することが理想であるが，本研究で使用した不均衡データを扱う多ラベル分類においてはこれを行うことは不可能であったため，含まれるラベルが最低 1 例ずつとなるような 5 件を抽出する措置を行った.

GPT-4 と ELYZA では，訓練データのテキストから 1 例を事例として，Zero-Shot Learning で示したプロンプト文と同様の予測すべき検査項目の候補と入力文の間に追加したものとし，その続きをモデルに生成させ，それに続く文字列として予測検査を出力し，Zero-Shot Leaning と同様の手順で評価を行った.

#### 3.3.4　Few-Shot Learning による実験設定

BERT では，One-Shot Learning と同様の手順で，各ラベルが少なくとも 4 例となるよう訓練データから 13 件抽出したデータで学習し予測を行った.

GPT-4 と ELYZA では，訓練データのテキストから 2 例を事例として，Zero-Shot Learning で示したプロンプト文と同様の予測すべき検査項目の候補と入力文の間に追加したものとし，その続きをモデルに生成させ，それに続く文字列として予測検査を出力し，Zero-Shot Leaning と同様の手順で評価を行った.

#### 3.3.5　医療従事者の予測

医療従事者である著者の 1 人が，テストデータの各事例に対して，予測ラベルを表 1 から選択した.

### 3.4　結果

実験結果を表 3 に示す．BERT と ELYZA の Fine-tuning を行った場合，Precision と F1 の値においては，医療従事者による予測性能を上回ることが確認された．一方で，ELYZA と GPT-4 の Zero-Shot, One-Shot と Few-Shot Learning における性能は医療従事者の予測性能より低いことが確認された．同様に，One-Shot, Few-Shot における BERT の性能も医療従事者の予測性能より低いことがわかった.

## 4　考察

### 4.1　Fine-tuning

全てのモデルにおいて，Precision と F1 の値が医療従事者の予測結果を上回ることが確認されたが，Recall の値においては，どのモデルも医療従事者の予測結果を下回る結果となった．この要因として，医療従事者は検査を実施しないことによる疾患の見落としリスクを念頭においていること，モデルの予


This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

**表 3** 実験結果: 医療従事者の結果は Fine-tuning 行に記載する．

| | 東北大 BERT | | | JMedRoBERTa | | | ELYZA | | | GPT-4 | | | 医療従事者 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Pre | Rec | F1 | Pre | Rec | F1 | Pre | Rec | F1 | Pre | Rec | F1 | Pre | Rec | F1 |
| Fine-tuning | 0.801 | 0.706 | 0.750 | 0.826 | 0.699 | 0.757 | 0.797 | 0.694 | 0.742 | - | - | - | 0.614 | 0.787 | 0.690 |
| Zero-Shot | - | - | - | - | - | - | 0.418 | 0.606 | 0.495 | 0.613 | 0.746 | 0.673 | - | - | - |
| One-Shot | 0.666 | 0.563 | 0.602 | 0.672 | 0.572 | 0.610 | 0.552 | 0.498 | 0.516 | 0.654 | 0.642 | 0.648 | - | - | - |
| Few-Shot | 0.800 | 0.550 | 0.651 | 0.808 | 0.543 | 0.648 | 0.446 | 0.719 | 0.551 | 0.708 | 0.660 | 0.683 | - | - | - |

測においては訓練データ内の出現頻度が低いラベルの予測が困難であったためと考えられる．ラベルごとの予測結果を分析したところ，ラベル数が 30 件より少ない場合は予測精度が著しく低いことが調査で明らかとなった．これより，本研究では，有病率の低い希少疾患を対象としたコーパスを対象としたため，出現頻度が低いラベル，つまり特定の疾患のみで行われるような実施頻度の低い検査項目の予測精度が低かったと言える．そのため，臨床現場での応用を考える際には，どのような疾患を扱うかなどのタスクの要件定義の検討と，疾患ごとの有病率を考慮したデータの不均衡性に対する対策が必要であると考えられる．

### 4.2 Zero/One/Few-Shot Learning

４つのモデルのうち，３つのモデルにおいては事例数つまり学習データ数が多いほど，高い性能を示すことが明らかとなった．GPT-4 のみ One-shot よりも Zero-shot の方が，性能が高く医療従事者の予測に近い結果となることが確認された．これは GPT-4 ではコストの関係から他のモデルとは異なり，1 回しか実験ができておらず，入力となった事例による影響が大きいと考えられ，今後の更なる検討が必要である．Few-Shot では，Fine-tuning を行った BERT の性能に及ばないまでも，Few-Shot を行ったモデルの中では GPT-4 が最も高い F1 を示し，医療従事者の精度と遜色ない値であることがわかった．本研究で設定したタスクにおいては，Fine-tuning を行った BERT の性能が最も高い結果となったが，与える事例の内容などプロンプトを工夫することで精度の改善につながる可能性がある．

## 5 おわりに

症例報告テキストを用いた検査項目の予測において，Fine-tuning を行った LLM の精度が最も高く医療従事者よりも高いことが明らかとなった．データ量が十分にあるのであれば，LLM の Fine-tuning を行うことが有効であるが，一方で，Zero-shot Learning における GPT-4 の精度は医療従事者の精度と遜色ない値となることが確認されたため，今後の応用が期待される．

本研究の課題として不確実性への対応がある．不確実性としては，正解ラベルのアノテーションと医療従事者の予測は著者の一人によって実施されたものであるため正確性に課題が残る．これについては今後，人数を増やし多数決を取るなどすることを考えている．また GPT-4 は 1 回の試行のみの結果であることから，他のモデルと同様に複数回の平均値を算出することがより正確な議論のために必要である．特殊性としては，本研究で使用した iCorpus は，難病・希少疾患に着目した症例報告テキストコーパスであり，このコーパスでは扱われない疾患についての予測については再現性が担保されていないことから，今後，実際のリアルワールドデータを用いた追加の検討が必要であると考えられる．

## 謝辞

本研究は日本電気株式会社の 2023 年度研究インターンシップの一部として実施されたものである．

## 参考文献

[1] 瀬戸僚馬, 蓮岡英明, 三谷嘉章, 山下小百合, 若林進, 渡辺明良, 石神久美子, 武藤正樹, 開原成允. 医師事務作業補助者の業務と電子カルテ等への代行入力の現状. 医療情報学, Vol. 29, No. 6, pp. 265–272, 2009.

[2] Zekai Chen, Mariann Micsinai Balan, and Kevin Brown. Boosting transformers and language models for clinical prediction in immunotherapy. **arXiv preprint arXiv:2302.12692**, 2023.

[3] John W Ayers, Adam Poliak, Mark Dredze, Eric C Leas, Zechariah Zhu, Jessica B Kelley, Dennis J Faix, Aaron M Goodman, Christopher A Longhurst, Michael Hogarth, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. **JAMA internal medicine**, 2023.

[4] Hutson M. Could ai help you to write your next paper? **Nature**, Vol. 611, No. 7934, pp. 192–193, 2022.

[5] Emiko Shinohara, Daisaku Shibata, and Yoshimasa Kawazoe. Development of comprehensive annotation criteria for patients' states from clinical texts. **Journal of Biomedical Informatics**, Vol. 134, p. 104200, 2022.



This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

[6] 大江 和彦山上 浩志. 標準臨床検査マスターのコード充足性に関する定量的評価. 医療情報学, Vol. 33, No. 3, pp. 139–150, 2013.
[7] HIS, RIS, PACS, モダリティ間予約, 会計, 照射録情報連携 指針バージョン 3.4（2022）＜ JJ1017 指針 Ver 3.4（2022）＞.（公社）日本放射線技術学会（JSRT）, 2022 年 7 月 1 日.


This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

# A 付録

表 4 使用モデル

| | LLM | | BERT | |
|---|---|---|---|---|
| モデル | ELYZA | GPT-4 | 東北大 BERT | JMedRoBERTa |
| 開発者 | 株式会社 ELYZA | OpenAI | 東北大学 | 国立情報学研究所 |
| 対応言語 | 日本語 | 複数言語 | 日本語 | 日本語 |
| 事前学習データ | OSCAR や Wikipedia 等 | インターネット上の大規模データ | 日本語 CC-100 や 日本語 Wikipedia 等 | 日本語の医学論文 |
| トークナイザー | * | N-gram | MeCab(unidic)+WordPiece | MeCab(IPAdic+万病辞書) |

* Llama2 を改良した独自のトークナイザー

表 5 プロンプトの詳細

Zero-Shot Learning Prompt.

あなたは誠実で優秀な日本人医師です。本文の患者に必要な検査項目の組み合わせを、
検査のリスト：['X 線単純撮影', 'X 線透視・造影', 'X 線血管撮影', 'X 線 CT 検査', 'MRI 検査', '核医学検査', '超音波検査', '一般検査', '血液学的検査', '生化学的検査', '内分泌学的検査', '免疫学的検査', '微生物学的検査', '病理学的検査', '脳波・脳磁図検査', '心電図検査']から選択し、リスト形式で回答してください。
本文：{**input**}
検査項目：

One-Shot Learning Prompt.

あなたは誠実で優秀な日本人医師です。本文の患者に必要な検査項目の組み合わせを、
検査のリスト：['X 線単純撮影', 'X 線透視・造影', 'X 線血管撮影', 'X 線 CT 検査', 'MRI 検査', '核医学検査', '超音波検査', '一般検査', '血液学的検査', '生化学的検査', '内分泌学的検査', '免疫学的検査', '微生物学的検査', '病理学的検査', '脳波・脳磁図検査', '心電図検査']から選択し、リスト形式で回答してください。
例：
本文：患者：44 歳，女性.主訴：低身長.既往歴：特になし.家族歴：母と弟に低身長，母に脊柱管内靭帯骨化症.現病歴：幼少時より低身長，35 歳頃より下肢痛，腰痛を自覚.近医にて後縦靭帯骨化症(OPLL)，黄色靭帯骨化症(OYL)を疑われ，1999 年 7 月，当院整形外科にて頸椎，胸椎椎弓形成術を施行，その後，低身長精査のため当科転科となった.入院時現症：身長 131.7cm，体重 41kg，体温 36.5°C，血圧 120/70mmHg，下肢 0 脚.
検査項目：['生化学的検査', '内分泌学的検査', 'X 線単純撮影', '一般検査', 'X 線 CT 検査'],
本文：{**input**}
検査項目：

Few-Shot Learning Prompt.

あなたは誠実で優秀な日本人医師です。本文の患者に必要な検査項目の組み合わせを、
検査のリスト：['X 線単純撮影', 'X 線透視・造影', 'X 線血管撮影', 'X 線 CT 検査', 'MRI 検査', '核医学検査', '超音波検査', '一般検査', '血液学的検査', '生化学的検査', '内分泌学的検査', '免疫学的検査', '微生物学的検査', '病理学的検査', '脳波・脳磁図検査', '心電図検査']から選択し、リスト形式で回答してください。
例 1：
本文：患者：44 歳，女性.主訴：低身長.既往歴：特になし.家族歴：母と弟に低身長，母に脊柱管内靭帯骨化症.現病歴：幼少時より低身長，35 歳頃より下肢痛，腰痛を自覚.近医にて後縦靭帯骨化症(OPLL)，黄色靭帯骨化症(OYL)を疑われ，1999 年 7 月，当院整形外科にて頸椎，胸椎椎弓形成術を施行，その後，低身長精査のため当科転科となった.入院時現症：身長 131.7cm，体重 41kg，体温 36.5°C，血圧 120/70mmHg，下肢 0 脚.
検査項目：['生化学的検査', '内分泌学的検査', 'X 線単純撮影', '一般検査', 'X 線 CT 検査'],
例 2：
本文：患者：75 歳，男性.主訴：胸やけ.既往歴：前立腺癌(平成 21 年 9 月に前立腺全摘術を施行後，ホルモン療法を継続中)，高血圧，糖尿病，大腸ポリープ.家族歴：特記事項なし.現病歴：CCS の診断にて平成 15 年より近医通院中であった.平成 22 年 4 月頃より胸やけ症状が出現し，上部消化管内視鏡検査を施行したところ胃癌を指摘され，加療目的で当科へ紹介受診した.なお，CCS 症状はプレドニン 7.5mg/日にてコントロール良好であった.入院時現症：身長 162cm，体重 63kg，下腹部正中に手術創を認めた.腹部に明らかな腫瘤は触知されなかった.
検査項目：['生化学的検査', '免疫学的検査', '病理学的検査', 'X 線 CT 検査', '血液学的検査'],
本文：{**input**}
検査項目：


This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).