E-026

矯正歯科医と大規模言語モデルの協働による効率的な自動診断に向けて

Toward Efficient Automated Diagnosis by Collaboration of Orthodontists and Large Language Models

杉原 壮一郎† Soichiro Sugihara　梶原 智之† Tomoyuki Kajiwara　池田 直樹‡ Naoki Ikeda　谷川 千尋‡ Chihiro Tanikawa　二宮 崇† Takashi Ninomiya

1. はじめに

矯正歯科治療において，適切な診断を実現するために，矯正歯科医には豊富な専門知識と長年の臨床経験が求められる．そこで我々は，自然言語処理を応用した自動診断の併用によって，矯正歯科医のための診断支援 [1-4] に取り組んできた．本研究では，大規模言語モデル (LLM) に基づく自動診断システム [1] を改良した上で，矯正歯科医と LLM の協働による効率的な診断の実現を目指す．自動診断は，経験の浅い矯正歯科医の判断を補助する材料となり，見落としや誤診のリスクを低減する有効な手段となることが期待できる．また，経験豊富な専門医にとっても，自動診断の導入は診断業務の効率化や負担軽減につながり，質の高い医療をより多くの患者に提供するために重要である．

先行研究 [1-3] では，所見文書を対象とするマルチラベル文書分類の問題として矯正歯科治療の診断をタスク設計し，自然言語処理の技術を応用して自動診断に取り組んできた．具体的には，Support Vector Machine (SVM) などの機械学習モデル [2] や Convolutional Neural Network (CNN) などの深層学習モデル [3]，大規模言語モデル (LLM) [1] を用いた研究があり，10 年以上の経験を持つ専門医に匹敵する性能 [4] が確認されている．しかし，LLM の訓練データでも充分にカバーできていないと考えられる専門用語が所見文書には多く含まれている [3] ため，依然として診断性能には改善の余地が残されている．

本研究では，この課題に対処するため，入力の所見文書や出力の病状ラベルに含まれる専門用語に対して，平易な表現への言い換えや英語訳などの前処理を施す．これによって，テキストを LLM にとってより理解しやすい表現に変換し，LLM に基づく自動診断の性能改善を目指す．

さらに本研究では，矯正歯科医と LLM の協働による診断性能の向上にも取り組む．LLM の性能向上に伴い，自動診断は平均的な性能では専門医の能力にも匹敵する [4] とはいえ，禁忌肢の選択 [5] など人間にとって簡単に見つけられる誤りを犯す場合もある．そこで，図 1 に示すように，LLM の診断結果を矯正歯科医が確認し，確実な誤答を訂正したり誤答の可能性がある病状に注目して LLM に再診断させたりする診断支援システムを構築し，LLM に基づく自動診断を改善できる可能性について議論する．

実験の結果，所見文書中の専門用語を平易に言い換える工夫や，選択肢の病状に含まれる専門用語に対して平易な言い換えや英語訳を併記する工夫により，LLM に基づく自動診断の性能を改善でき，本タスクにおける最高性能を達成した．また，LLM が出力した診断結果から病状を 1 件ずつ無作為に抽出して再診断したところ，診断性能の更なる改善を確認できた．この結果は，LLM と矯正歯科医の協働による効率的かつ高品質な診断の可能性を示している．

† 愛媛大学　Ehime University
‡ 大阪大学　Osaka University

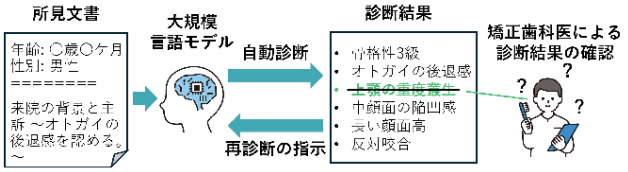

図 1：矯正歯科医と大規模言語モデルの協働

2. 関連研究

2.1 自然言語処理の医療応用

医療現場においては，初診記録や手術記録，退院サマリなど，多様な種類の文書が日々作成されており，これらは患者の情報を記録および共有するための重要なテキストデータとして蓄積されている．近年では電子カルテの普及に伴い，医療情報の電子化が進み，テキストデータの利活用が容易になってきている．こうした背景のもと，医療分野に特化した自然言語処理の研究開発が活発化している．例えば，PubMed[*1] や MIMIC-III [6] などの医療テキストに基づき，PubMedBERT [7] や ClinicalBERT [8] のような医療言語処理に特化した事前訓練モデルが開発されてきた．

近年では，LLaMA [9] や Mistral [10] などの LLM が自然言語処理の多様なタスクにおいて高い性能を発揮している．医療言語処理においても，一般的なドメインで事前訓練された LLM を医療テキストで追加訓練した PMC-LLaMA [11] や BioMistral [12] などのドメイン特化 LLM が開発されている．しかしながら，歯科矯正学に特化した LLM は現時点では存在しておらず，先行研究 [1] においても汎用的な LLM のひとつである Qwen2.5[*2] [13] が採用されている．

2.2 自動診断の関連研究

自然言語処理や画像処理の技術を応用した診断支援は，医療分野の全般にわたって研究が進められている．例えば，胸部 X 線や CT 画像を用いた COVID-19 の感染有無の判定 [14] や医師の診断ロジックを LLM に組み込んだ医療対話システム [15] などの取り組みが報告されている．

矯正治療においては，bag-of-words を用いた所見文書の特徴抽出と SVM などの機械学習モデルによる自動診断 [2] や，CNN などの深層学習モデルによる自動診断 [3]，LLM による自動診断 [1] などが提案されている．LLM に基づく自動診断に関する先行研究 [1] では，LLM に与えるプロンプトを英語化および JSON 形式化することによって，自動診断の性能を大幅に改善できることが示された．しかし，先行研究 [3] で挙げられた専門用語の頻出に関する課題には依然として対処できていない．本研究では，専門用語に前処理を施すことで，LLM の診断性能の向上を目指す．

*1 https://pubmed.ncbi.nlm.nih.gov/download/
*2 https://huggingface.co/Qwen/Qwen2.5-72B-Instruct

表 1：矯正歯科治療の自動診断のためのプロンプト（緑字は本研究で先行研究 [1] に追加した部分）

例

```
{
    "Persona": "You are a dentist.",
    "Instructions": [
        "A report of orthodontic findings and a list of options will be provided below.",
        "Please select from the options the medical condition that can be identified from the findings."
    ],
    "Options": [{“ID”: “1”, “Medical Condition_JP”: “臼歯部開咬”,
                “Medical Condition_EN”: “Posterior open bite”, “Meaning”: “奥歯の咬み合わせが開いている”},…],
    "Findings report": {専門用語を平易化した所見文書},
    “Output”:
}
```

3. 専門用語の前処理による自動診断の改善

先行研究 [1] と同様に，本研究では LLM による自動診断を，所見文書を入力して当該患者に該当する全ての病状ラベルを出力するマルチラベルのテキスト分類タスクとして実現する．先行研究 [1] は LLM に与えるプロンプトの形式を工夫して診断性能を改善したが，本研究では専門用語の取り扱いに着目し，表 1 に示すように，以下の 3 つの工夫を加えたプロンプトを用いて診断性能の改善を目指す．

3.1 所見文書内の専門用語の平易化

LLM は様々なドメインにわたる大規模コーパスを用いて事前訓練されているとはいえ，歯科矯正学に特化したテキストを充分に含んでいるわけではなく，当該ドメインに特有の専門用語を必ずしも適切に理解できるとは限らない．そこで本研究では，専門用語と一般的かつ平易な表現の対からなる言い換え辞書を作成し，所見文書に含まれる専門用語の自動的な語彙平易化を試みた．この言い換え辞書は，255 語の専門用語を対象としており，歯科矯正学の専門知識を持つ著者の 1 人が作成した．

3.2 選択肢内の専門用語の平易化

所見文書と同様に，診断結果の候補となる病状ラベルにも専門用語が含まれる．3.1 節では文脈情報を有効活用するために所見文書中の専門用語を平易な表現に置換したが，文脈を持たない病状ラベルは元の表現を残したまま，それに対応する平易な表現も併用する．3.1 節と同様に，著者の 1 人が 322 種類の病状ラベルに言い換えを付与した．

3.3 選択肢内の専門用語の英語化

先行研究 [1] では，プロンプトに含まれる指示テキストを英語に翻訳することによって，診断性能の向上が報告されている．しかし，所見文書や選択肢などの入力テキストに含まれる専門用語は日本語のままであった．そこで，3.2 節と同様に，日本語の病状ラベルに対する英語訳を併用する．これにより，LLM が大規模な事前訓練を通して得た英語に関する豊富な知識を有効活用することが期待できる．ただし，専門用語を多く含むテキストは機械翻訳の品質も不充分であることを考慮し，コストの問題で所見文書は日本語のまま使用し，病状ラベルのみを人手で英語に翻訳する．3.2 節と同様に，著者の 1 人が 322 種類の病状ラベルに英語訳を付与した．

表 2：再診断用のプロンプト

例（表 1 に続けて，以下のプロンプトを追加）

```
{
    “Instruction”: [
        “If you determine it is necessary to remove the
         following diagnoses by referring to the
         'Findings report and previous 'Problems', output
         {{"Answer": "True"}}; if not,
         output {{"Answer": "False"}}.”,
         "However, if there is any ambiguity, or if the
         documents do not clearly justify removal, output
         {{"Answer": "False"}}."
    ],
    “Problem”: {再診断の対象となる病状ラベル}
}
```

4. 矯正歯科医と大規模言語モデルの協働

本研究では，LLM を単独で自動診断に用いるだけでなく，矯正歯科医との協働により診断性能の更なる向上を目指す．診断の際に矯正歯科医が 3 節の自動診断システムを補助的に使用する状況を想定し，

(1) LLM の診断結果を矯正歯科医が修正する機能

(2) 矯正歯科医の指摘を受けて LLM が再診断する機能

の 2 つを実装する．これにより，矯正歯科医と LLM が相互に補完し合うことができ，高品質な診断を期待できる．

機能 (1) は，矯正歯科医が特定の病状ラベルの有無について自信を持っている状況で，矯正歯科医と LLM の診断に相違がある際に，診断結果に矯正歯科医の判断を反映させる機能である．これは，矯正歯科医の判断に従い，単純に当該ラベルを出力に含めるまたは出力から除外する．

機能 (2) は、矯正歯科医と LLM の診断に相違があるものの，矯正歯科医がその差分について充分な自信を持っていない状況を想定している．本研究では経験の浅い矯正歯科医の診断支援を目的としているため，このような状況が考えられる．この場合，矯正歯科医が指定する特定の病状ラベルについて，LLM に出力の妥当性を再検討させる．具体的には，表 2 に示すプロンプトを用いて，ある病状ラベルを診断結果に含めるか否かを True / False で回答させ，LLM による再診断を実現する．

5. 評価実験

5.1 専門用語の前処理による自動診断の改善

本実験では，先行研究 [1] と同様の設定で，LLM に基づく自動診断における 3 節の提案手法の有効性を評価する．

5.1.1 実験設定

データ

本実験で使用するデータセットは，大阪大学歯学部附属病院に所蔵されている矯正歯科治療に関する文書である．先行研究 [1-3] と同様に，合計 976 件の症例データを対象とした．本データセットには，診察時に担当医が記述した所見文書とともに，患者の顔画像および X 線画像が含まれている．所見文書には，症状の記載に加えて，年齢や性別などの基本的な患者属性情報も含まれる．本研究では，テキストデータの活用に焦点を当て，画像情報の併用については今後の課題とする．

所見文書は 1 件あたり 312～6,379 トークンで構成され，平均で約 1,886 トークンである．また，それぞれの所見文書には複数の病状ラベルが付与されており，322 種類のラベルのうち，1 人あたり平均 12 件（最少 3 件，最多 28 件）が付与されている．データセットの分割は，訓練用・検証用・評価用にそれぞれ 632 件・159 件・185 件を使用した．

モデル

本研究では，患者の個人情報を含む所見文書を扱うため，情報漏洩のリスクがないローカル LLM を実験に使用した．本実験では，先行研究 [1] において高い性能を示した Qwen2.5[*2] [13] を採用した．Qwen2.5 は，日本語にも長文入力にも対応しているため，本実験に適している．

72B の大規模モデルを効率的にファインチューニングするため，4 ビット量子化されたモデル[*3] に対して QLoRA チューニング [16] を適用した．学習には unsloth[*4] を使用し，最適化手法には AdamW [17] を使用した．最大の学習率を 2×10^{-4}，バッチサイズを 8 とし，検証用データにおける損失が 3 エポック連続で改善しない場合に訓練を停止する early stopping を適用した．

比較手法として，先行研究の SVM [2]，CNN [3] および LLM [1] による手法を用いた．SVM は scikit-learn[*5]，CNN は PyTorch[*6]，LLM は unsloth[*4] を用いて実装した．

プロンプト

LLM に与えるプロンプトは，表 1 に示したように，先行研究 [1] で提案されたプロンプトに対して，専門用語への前処理を加えたものを用いた．プロンプトには，ペルソナ，タスクの説明，選択肢，入力テキストの 4 種類の情報が含まれる．ペルソナは，LLM が矯正歯科医として振る舞うように設定する．タスクの説明としては，所見文書と所与の選択肢に基づき，該当する全ての病状ラベルを回答するように指示を与える．また，選択肢はマルチラベル分類のラベル集合であり，入力テキストは診断対象の所見文書である．なお，LLM の出力は，選択肢と同様，ID とラベル名から構成される出力形式を想定する．

*3 https://huggingface.co/unsloth/Qwen2.5-72B-Instruct-bnb-4bit
*4 https://github.com/unslothai/unsloth
*5 https://scikit-learn.org/stable/
*6 https://pytorch.org/
*7 https://scikitlearn.org/stable/modules/generated/sklearn.metrics.f1_score.html

表 3：大規模言語モデルによる自動診断の性能

モデル	手法	F1
SVM [2]	-	0.425
CNN [3]	-	0.440
LLM [1]	①指示の JSON 形式化+英語化	0.482
LLM [1]	日本語のマークダウン形式の指示	0.267
LLM	②所見文書内の専門用語の平易化	0.445
LLM	③選択肢内の専門用語の平易化	0.308
LLM	④選択肢内の専門用語の英語化	0.496
LLM*	①～④全て	0.503

評価方法

先行研究 [1-3] に従い，マルチラベル分類における F1 スコア[*7] を評価した．各モデルは病状ラベルの ID を出力に含むため，これらの ID を抽出して自動評価した．

5.1.2 実験結果

表 3 に，自動診断の F1 スコアを示す．まず，マークダウン形式のプロンプト [1] をベースラインとし，所見文書内の専門用語を平易化（②）すると，F1 において 0.267 から 0.445 へと著しい性能向上を確認できた．次に，病状ラベルとして専門用語の平易化を併記（③）すると，ベースラインからわずかな性能改善を得た．そして，病状ラベルとして専門用語の英語訳を併記（④）することで，ベースラインからの最も大きな性能改善が見られた．最後に，先行研究 [1] で有効性が確認されたプロンプトの JSON 形式化および英語化（①）に，本研究で提案した 3 手法を組み合わせたところ（LLM*モデル），F1=0.503 を達成し，矯正歯科治療の自動診断における最高性能を更新できた．

5.2 矯正歯科医と大規模言語モデルの協働

本実験では，矯正歯科医と LLM が協働して診断に取り組むというシナリオを想定し，LLM による再診断によってどの程度の診断性能の向上が期待できるかを評価する．

5.2.1 実験設定

データとモデル

本実験では，5.1 節と同じモデルおよびデータセットを使用する．評価用データセットに対する LLM*モデルの診断結果について，以下に該当する 4 種類の病状ラベルを無作為に 1 つずつ選択し，これらを矯正歯科医が指摘したと想定して再診断（4 節の機能 (2)）の正解率を評価する．

- TP: 出力すべきで，正しく診断結果に含まれる
- FP: 出力すべきでなく，誤って診断結果に含まれる
- FN: 出力すべきで，誤って診断結果に含まれない
- TN: 出力すべきでなく，正しく診断結果に含まれない

ただし，評価用データセットのうち 2 件の所見文書では，FN に該当するラベルが存在しなかった．

プロンプトと評価方法

LLM に与えるプロンプトは，表 2 に示したものを用いた．TP および FP の場合は，診断結果に含まれる特定の病状ラベルに対して，それを削除するか否かを再診断した．同様に，TN および FN の場合は，診断結果に含まれない特定の病状ラベルに対して，それを追加するか否かを再診断した．再診断の出力は True / False の 2 値であり，正解率によって自動評価した．つまり，TP および TN の場合は False が正解であり，FP および FN の場合は True が正解である．

表 4：大規模言語モデルによる再診断の性能

データ	出力		正解率
	True	False	
TP	46	139	0.751
FP	78	107	0.422
FN	134	49	0.732
TN	18	167	0.903

5.2.2 実験結果

表 4 に，再診断における正解率を示す．True Negative（TN）の病状ラベルに対しては 9 割，True Positive（TP）および False Negative（FN）の病状ラベルに対しては 7 割強という高い正解率を達成できた．この結果から，LLM による診断結果に対し矯正歯科医が指摘を加えることで，LLM がより良い診断を実現できる可能性を確認できた．

一方で，False Positive（FP）の病状ラベルに対しては再診断の正解率が 4 割程度にとどまっており，他と比較して著しく低い値を示す結果となった．これは，LLM が誤って出力した病状ラベルについては，矯正歯科医からの指摘を受けてもそれを診断結果から適切に除外することが難しいことを表している．したがって，自動診断の技術の活用においては，このような LLM による誤診の可能性に充分な注意が必要であり，今後の更なる改良が求められる．

6. おわりに

本研究では，矯正歯科治療における自動診断の性能改善を目的として，大規模言語モデル (LLM) に基づく自動診断に取り組んだ．具体的には，所見文書および病状ラベルに含まれる専門用語に対して，平易化や英語化といった前処理を施すことで，LLM による診断性能の向上を図った．実験の結果，所見文書内の専門用語の平易化処理によって診断性能が向上することが確認でき，さらに病状ラベル内の専門用語についても，平易化および英語化の前処理を施すことで診断性能が改善することが確認できた．これら全ての工夫が診断性能の改善に貢献し，それらを組み合わせることによって既存手法を上回る診断性能を達成できた．

さらに，自動診断の結果に対する再評価のプロセスを新たに導入し，矯正歯科医と LLM の協働による診断支援の可能性についても検証した．その結果，False Positive ラベルに対する再診断の正解率は 4 割程度と低かったものの，True Positive・False Negative・True Negative の各ラベルに対しては 7 割以上の正解率を達成し，矯正歯科医と LLM の協働による再診断の有効性が期待できる結果を確認できた．

今後の課題として，本研究で対象とした所見文書に加えて，患者の顔画像および X 線画像などの画像情報を統合的に活用するマルチモーダル型の自動診断への拡張が挙げられる．また，LLM の歯科矯正学分野へのドメイン特化などに取り組むことで，更なる診断性能の向上が期待できる．

謝辞

本研究は，JSPS 科研費（基盤研究 C，課題番号：JP24K13195）の助成を受けたものです．

参考文献

[1] 杉原 壮一郎, 梶原 智之, 池田 直樹, 谷川 千尋, 二宮 崇: 大規模言語モデルによる矯正歯科治療の自動診断に向けて, 人工知能学会第 39 回全国大会, 2Win5-35 (2025)

[2] 西原 大貴, 梶原 智之, 谷川 千尋, 清水 優仁, 長原 一: 矯正歯科治療における所見文書からの自動診断に向けて, 情報処理学会第 83 回全国大会, pp. 591-592 (2021)

[3] Ohtsuka, T., Kajiwara, T., Tanikawa, C., Shimizu, Y., Nagahara, H., and Ninomiya, T.: Automated Orthodontic Diagnosis from a Summary of Medical Findings, in *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pp. 156–160 (2023)

[4] Shimizu, Y., Tanikawa, C., Kajiwara, T., Nagahara, H., and Yamashiro, T.: The Validation of Orthodontic Artificial Intelligence Systems That Perform Orthodontic Diagnoses and Treatment Planning, *European Journal of Orthodontics*, Vol. 44, No. 4, pp. 436–444 (2022)

[5] Kasai, J., Kasai, Y., Sakaguchi, K., Yamada, Y., and Radev, D.: Evaluating GPT-4 and ChatGPT on Japanese Medical Licensing Examinations, *arXiv*:2303.18027 (2023)

[6] Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G.: MIMIC-III, A Freely Accessible Critical Care Database, *Scientific Data*, Vol. 3, No. 160035 (2016)

[7] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H.: Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing, *ACM Transactions on Computing for Healthcare*, Vol. 3, No. 1, pp. 1–23 (2021)

[8] Huang, K., Altosaar, J., and Ranganath, R.: ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission, *arXiv*:1904.05342 (2019)

[9] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozi`ere, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G.: LLaMA: Open and Efficient Foundation Language Models, *arXiv*:2302.13971 (2023)

[10] Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, de las D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E.: Mistral 7B, *arXiv*:2310.06825 (2023)

[11] Wu, C., Lin, W., Zhang, X., Zhang, Y., Xie, W., and Wang, Y.: PMC-LLaMA: Toward Building Open-source Language Models for Medicine, *Journal of the American Medical Informatics Association*, Vol. 31, No. 9, pp. 1833–1843 (2024)

[12] Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.A., Rouvier, M., and Dufour, R.: BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains, in *Findings of the Association for Computational Linguistics*, pp. 5848–5864 (2024)

[13] Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., and Qiu, Z.: Qwen2.5 Technical Report, *arXiv*:2412.15115 (2024)

[14] Kassania, S. H., Kassanib, P. H., Wesolowskic, M. J., Schneidera, K. A., and Detersa, R.: Automatic Detection of Coronavirus Disease (COVID-19) in X-ray and CT Images: A Machine Learning Based Approach, *Biocybernetics and Biomedical Engineering*, Vol. 41, No. 3, pp. 867– 879 (2021)

[15] Dou, C., Zhang, Y., Jin, Z., Jiao, W., Zhao, H., Zhao, Y., and Tao, Z.: Integrating Physician Diagnostic Logic into Large Language Models: Preference Learning from Process Feedback, in *Findings of the Association for Computational Linguistics*, pp. 2453–2473 (2024)

[16] Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L.: QLoRA: Efficient Finetuning of Quantized LLMs, in *Proceedings of the 37th Conference on Neural Information Processing Systems*, pp. 10088–10115 (2023)

[17] Loshchilov, I. and Hutter, F.: Decoupled Weight Decay Regularization, in *Proceedings of the Seventh International Conference on Learning Representations* (2019)