# 大規模言語モデルを用いた電子カルテの SOAP 作成支援システムの開発

斉藤翼 [1]　山中稜斗 [1]　北岡教英 [1]
[1] 豊橋技術科学大学
{saito.tsubasa.xk, yamanaka.rikuto.lo, kitaoka}@tut.jp

## 概要

医療従事者の負担軽減を目的として，大規模言語モデル（LLM）を用いた SOAP ノート生成手法を提案する．提案手法は，SOAP 生成プロセスを SO 抽出タスクと AP 生成タスクの 2 段階に分割する．SO 抽出タスクでは，患者と医療従事者との対話文から主観的情報と客観的情報を抽出する．AP 生成タスクでは，抽出された主観的情報と客観的情報からアセスメントと計画を生成する．さらに，Retrieval-Augmented Generation (RAG) を導入し，過去の SOAP データを参照することで LLM の医学知識を補完する．実験の結果として，論理的一貫性と安全性の配慮において高い評価を得られたが，臨床的妥当性と完全性においては課題が残った．

## 1　はじめに

医療現場における看護記録や電子カルテの作成業務は診療以外の業務負荷において大きな割合を占めており，勤務時間の約 2 割を費やすという調査がある [1]．この現状は，医療従事者の疲弊や医療サービスの質低下に繋がる可能性があり，抜本的な改善が求められている．看護記録において広く採用されている SOAP 形式は，患者の主観的情報（Subjective），客観的情報（Objective），アセスメント（Assessment），計画（Plan）の 4 項目に基づき記録を行う方法である．主観的情報は患者の訴えている内容，客観的情報は検査結果や身体所見のデータ，アセスメントは主観的・客観的情報を基にした診断や評価，計画はアセスメントに基づく治療計画で構成され，以下が SOAP の例である．

S：さっぱりした。いろいろあったでね。
O：モニター：SR BP90 台 胸部症状なし 呼吸苦なし 体重：47.1kg（-0.1kg）下肢浮腫なし ADL：自立 病棟内歩行散歩される SW 見守りにて施行
　日中危険行動：なし
A：VS 著変なく経過。引き続き転倒に注意していく
P：プランに沿う

SOAP 形式は，患者の状態把握と治療計画の明確化や多職種間での情報共有を促進する効果を持つ．しかし，詳細な情報整理と分析を必要とするため，特にアセスメントと計画の作成は医療従事者にとって大きな負担となっている [2]．したがって，看護記録および電子カルテ作成の効率化は医療従事者の働き方改革，ひいては医療サービス全体の質向上に不可欠な要素と言える．

こうした課題解決に向けて，近年では大規模言語モデル（Large Language Model, LLM）を応用した医療支援技術が注目を集めている．LLM は，大量のテキストデータから学習した知識に基づき高精度な自然言語処理を可能にする技術であり，医療分野においても診断支援，治療計画の策定，患者とのコミュニケーション支援など様々な応用が期待されている．特に，SOAP 形式の記録作成支援は LLM の得意とする自然言語処理能力を活かせる分野として，国内外で研究開発が進められている．例えば，Google 社の MedLM[3] では電子カルテや医療データの検索，要約，質問応答などの支援が行われている．また，Yuan ら [4] は LLM を用いて医療会話から SOAP 形式の要約を自動生成する手法を提案している．しかしながら，これらの研究の多くは英語を対象としており，日本語医療現場への応用には課題が残されている．

そこで本研究では，日本語 LLM と国内の医療機関から提供されたデータセットを活用し，日本語の医療現場に特化した SOAP 作成支援システムの開発に取り組む．本システムの導入により医療従事者の文書作成負担を軽減し，より質の高い医療サービスの提供に貢献することが期待される．本稿では特に，アセスメントの作成支援について述べる．


**図 1**　電子カルテ入力支援システム概要

## 2　提案手法

音声認識と LLM を用いて SOAP を自動作成することにより，医療従事者の文書作成負担を軽減するシステム [5][6] を提案する．図 1 にシステムの概要を示す．本システムは医療機関内ネットワークを介したデバイス間連携により動作する．具体的には，医療従事者がスマートフォンを用いて音声入力を行うと，音声認識サーバでテキストデータに変換される．その後，テキストデータを LLM サーバに送信して SOAP を生成しクライアントに返送する．最終的には，医療従事者がクライアントであるスマートフォン上で生成された SOAP を確認・修正し，電子カルテデータベースに送信する．一連の操作は Web アプリケーションを介して行うことが可能である．本研究では，LLM サーバ部分での SOAP 生成に llm-jp/llm-jp-3-13b-instruct [7] を採用する．この LLM は，ローカル環境下での限られた計算リソースによる運用と，日本語テキストへの高い適合性という 2 つの要件を満たすことから選定した．SOAP 生成するプロセスとして，主観的情報（S）と客観的情報（O）の抽出，アセスメント（A）と計画（P）の生成という 2 つのタスクに分割する．SOAP の各項目はそれぞれ異なるレベルの情報処理を必要とするため，段階的な生成プロセスを採用することでより高精度な SOAP 生成を可能にする．具体的には，主観的情報と客観的情報は患者の発言や検査結果といった事実情報の抽出，および，主観的・客観的という視点からの分類が中心となる一方，アセスメントはそれらの情報に基づいた考察や解釈を含むため，より高次な情報処理が求められる．また，計画はアセスメントの内容に依存するためアセスメントと同時に生成を行う必要がある．

### 2.1　SO 抽出タスク

SOAP 生成の第一段階として，患者と医療従事者との対話文から主観的情報（S）と客観的情報（O）を抽出するタスクを設定する．ここでは，2 種類のプロンプトを用いて LLM によって主観的情報と客観的情報を抽出する．1 つ目は患者の発言を要約し，主観的な情報を抽出するためのプロンプトである．2 つ目はバイタル情報や身体所見といった客観的な情報を抽出するためのプロンプトである．これらのプロンプトを用いることで，LLM は対話文から主観的情報と客観的情報を効果的に抽出することが可能となる．ただし，本稿ではこの SO 抽出タスクについては述べない．

### 2.2　AP 生成タスク

2.1 節の SO 抽出タスクに続いて，抽出された主観的情報と客観的情報からアセスメント（A）と計画（P）を生成するタスクを設定する．実験で使用する LLM は一般的な日本語テキストデータで学習されており，医学知識は豊富ではない．そこで，Retrieval-Augmented Generation (RAG) [8] を導入し過去の SOAP データを参照することで LLM の医学知識を補完し，アセスメントと計画を作成する．RAG では，過去の SOAP データをベクトルストアとして構築し，入力された主観的情報と客観的情報に類似した過去の SOAP データを検索する．その後，検索された SOAP データを外部知識として LLM に与えることで，より適切なアセスメントと計画を生成する．アセスメント生成時のプロンプトを A 付録の表 3 に示す．なお，ベクトルストアの構築にはマイクロソフト社が開発した多言語対応のテキスト埋め込みモデルである，intfloat/multilingual-e5-large[9] を用いる．

### 2.3　データセット

本研究では，共同研究先の医療機関から提供された匿名化した看護記録データセットを利用する．本データセットには，SOAP データ，患者 ID，記録日時，診療科情報，SOAP に関連しないコメントが含まれている．データは 2012 年から 2024 年までの期間にわたり収集され，約 2 万人の患者に関する約 21 万件の記録で構成されている．本データセットは，RAG におけるベクトルストアの構築とシステムの評価において利用する．

# 3 実験

## 3.1 実験内容

本研究では，提案手法の有効性を検証するため，共同研究先の医療機関から提供された SOAP データセットを用いて実験を行った．ただし，対話音声データは現在収集中であるため，対話文から SO を作成する実験は実施しておらず，SO は作成されているものとして実験を行う．また，計画の生成に関してはデータセットにおける計画の大部分が「プラン継続」や「プラン終了」など定型的な文言で占められており，RAG を用いて参照するにも適切な計画のデータが少ないため，本稿ではアセスメント生成に焦点を絞って実験を行った．

実験の内容として，SOAP データセットの SO データからアセスメントを生成する実験を行った．データセットのうち，約 20 万件を RAG での検索時に関連文書として使用するためにベクトルストアに変換した．残り約 5000 件を実験用として利用し，SOAP データのうち SO を利用してアセスメントを LLM を用いて生成した．

その後，生成したアセスメントに対して ROUGE[10] と BERTScore[11] を用いた自動評価と，医療従事者による主観評価を実施した．

自動評価では，生成した 5000 件のアセスメントに対して ROUGE-1, ROUGE-2, ROUGE-L, BERTScore-F1 を計算した．具体的には，RAG を使用した生成文と RAG を使用していない生成文のそれぞれについて，SOAP データセットのアセスメントとの類似度を算出した．これらの指標を用いることで，RAG の導入がアセスメントの生成精度に与える影響を定量的に分析することが可能となる．

医療従事者による主観評価では，LLM による医療生成文書の評価に関する研究 [12][13] を参考に，生成したアセスメント 10 件に対して臨床的妥当性，論理的一貫性，完全性，安全性の配慮の 4 項目で評価した．各項目を 1 から 5 点の尺度で評価しその平均点を算出した．

**臨床的妥当性** 重症度が適切に評価され優先順位付けがされているか，アセスメントとして適切であるか．

**論理的一貫性** 主観的情報および客観的情報を適切に反映しており矛盾がないか．

**完全性** 幻覚や情報の欠落の有無を含め，出力が提供された情報と一致しているか．

**安全性の配慮** 患者の安全に関わる重要なリスクが適切に評価されているか．

**表 1** RAG 使用有無による ROUGE および BERTScore の比較

| 評価指標 | RAG 未使用 | RAG 使用 |
|---|---|---|
| ROUGE-1 | 0.181 | 0.275 |
| ROUGE-2 | 0.042 | 0.080 |
| ROUGE-L | 0.121 | 0.200 |
| BERTScore | 0.653 | 0.691 |

**表 2** 医療従事者による主観評価結果

| 評価項目 | 平均点数 |
|---|---|
| 臨床的妥当性 | 3.0 |
| 論理的一貫性 | 4.2 |
| 完全性 | 3.3 |
| 安全性の配慮 | 4.2 |

## 3.2 実験結果

表 1 に自動評価指標を用いた結果を示す．ROUGE-1, ROUGE-2, ROUGE-L, BERTScore のいずれにおいても RAG を使用した方が高い値を示しており，SOAP データセットのアセスメントに近い文を生成できていることがわかる．

また，表 2 に医療従事者による主観評価の結果を示す．臨床的妥当性については，3.0 点と低い評価となった．これは，LLM が生成したアセスメントにおいて，患者の状態を適切に反映できていない点が散見されたためである．例として，異常値が出ているにも関わらず正常と判断するケースや，DOE（労作時呼吸困難）や AAA（腹部大動脈瘤）といった略語に対して適切な処理が行えていないケースが見受けられた．

一方，論理的一貫性と安全性の配慮についてはそれぞれ 4.2 点と高い評価を得た．これは，LLM が生成したアセスメントが主観的情報と客観的情報に基づいて論理的に構成されており，患者の安全に配慮した内容であったことを示している．完全性については 3.3 点とやや低い評価となった．これは，LLM が生成したアセスメントにデータセットに存在しない臨床情報が含まれるなど，情報の正確性に課題が見られたためである．

## 4 考察

本研究では，対話文から SOAP を生成する LLM の開発に取り組んだ．自動評価指標を用いた評価では，RAG を用いて関連 SOAP データを LLM に与えた場合に ROUGE と BERTScore で優れた結果を得られた．これは，RAG が関連する SOAP データを参照することで LLM がより適切な単語やフレーズを選択し，SOAP データセットのアセスメントに近い文を生成できたことを示唆している．

また，医療従事者による主観評価の結果として，論理的一貫性や安全性の配慮は比較的高い評価となった一方で臨床的妥当性と完全性において低い評価となった．臨床的妥当性の低さについては以下の 2 点が要因として考えられる．一つ目は，客観的情報の数値を適切に評価できていない点が挙げられる．以下は，入力の客観的情報の数値を適切に評価できていない例である．

入力
S: 健康のため。奥さんには内緒にして欲しい。
前回はお酒の席で吸ってしまい禁煙に失敗。
O: たばこ 10〜20 本/日 20 年間喫煙
ブリンクマン指数 400 TDS5 点 CO 濃度 16ppm
BW65.2kg 禁煙外来 2 回目 同居人に喫煙者なし
チャンピックス処方あり。
出力
A: 喫煙本数や年数からすると比較的軽度の喫煙者
禁煙に対する気持ちは前向き。禁煙外来 2 回目
の受診であるため前回の禁煙外来での失敗を聞
いて禁煙に対するサポートを考える。

出力のアセスメントでは，ブリンクマン指数 400（たばこを 1 日当たり 10〜20 本，20 年間喫煙）を比較的軽度の喫煙者と評価している．しかし，一般的にブリンクマン指数 400 の喫煙者は重度であると判断されることが多く，このアセスメントの臨床的妥当性は低いと評価された．LLM がこのような判断を下した原因として，A 付録の表 4 に示す検索された関連 SOAP の客観的情報に記述されているブリンクマン指数が 400 を超えており，入力のブリンクマン指数は比較的軽度だと判断された可能性が考えられる．

二つ目に，DOE（労作時呼吸困難）や AAA（腹部大動脈瘤）などの略語を，LLM が理解できていなかったことも評価の低下に繋がったと考えられる．LLM は学習データに含まれていない医学用語や略語を正確に解釈することが難しいため，医学的な専門知識を必要とするアセスメントの生成においてはさらなる改善が必要であるといえる．

完全性の低さについては，RAG で取得した関連 SOAP のバイタル情報が生成文に含まれてしまうことにより，臨床情報の過多が発生していたことが要因として考えられる．LLM は入力された情報に基づいて文を生成するため，関連 SOAP の情報が過度に含まれるとアセスメントとして必要な情報が不足したり，逆に不要な情報が含まれたりする可能性がある．

しかしながら，実際の医療従事者によるアセスメントよりも高い評価を得た生成文も存在した．これは，SOAP の記述には個人差があり必ずしも全ての医療従事者が同じ基準でアセスメントを記述するわけではないことを示唆している．

これらの結果を踏まえ，今後の研究では LLM の医学知識を補足するためのバイタルの基準値や略語データの拡充，臨床情報の過不足を調整するためのプロンプト制御などが課題として挙げられる．

## 5 結論

本研究では，LLM を用いた SOAP 生成手法を提案し，医療従事者の負担軽減を目的とした電子カルテ入力支援システムの開発に取り組んだ．提案手法では，SOAP 生成プロセスを SO 抽出タスクと AP 生成タスクの 2 段階に分割し SOAP の生成を目指した．

アセスメント作成実験の結果，RAG を用いることで SOAP データセットのアセスメントに近い文を生成できた．また，論理的一貫性と安全性の配慮において高い評価を得られた一方で，臨床的妥当性と完全性において課題が残る結果となった．これは，LLM が医学用語や略語の理解，患者の状態を適切に反映したアセスメント生成，および臨床情報の過不足の調整に課題を抱えていることを示唆している．

今後の展望としては，LLM の医学知識を補足するためのバイタルの基準値や略語データの拡充，臨床情報の過不足を調整するためのプロンプト制御などが挙げられる．さらに，対話音声データを用いた SO の抽出から AP の生成までの一連のタスクの評価を行うことで，より実用的な電子カルテ入力支援システムの構築を目指す．

## 謝辞

本件の一部に, 愛知県が公益財団法人科学技術交流財団に委託し実施している「知の拠点あいち重点研究プロジェクト第 IV 期（第 4 次産業革命をもたらすデジタル・トランスメーション（DX）の加速）」の研究成果が使われている.

## 参考文献

[1] 小川晃司, 竹内朋子. 勤務帯別にみた看護記録時間の関連要因. 日本看護管理学会誌, Vol. 25, No. 1, pp. 245–252, 2021.

[2] 豊福佳代, 川本利恵子. 電子カルテを使用している看護師の看護記録に関する認識. 日本職業・災害医学会会誌, Vol. 66, No. 3, pp. 201–209, 2018.

[3] Google Cloud. Introducing medlm for the healthcare industry. https://cloud.google.com/blog/topics/healthcare-life-sciences/introducing-medlm-for-the-healthcare-industry, (2024-12 閲覧).

[4] Dong Yuan, Eti Rastogi, Gautam Naik, Sree Prasanna Rajagopal, Sagar Goyal, Fen Zhao, Bharath Chintagunta, and Jeffrey Ward. A continued pretrained llm approach for automatic medical note generation. In **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)**, pp. 565–571, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[5] Ryo Maejima and Norihide Kitaoka. Speech recognition interface for updating electronic medical records with automatic itemization. In **Proceedings of the International Conference on Artificial Intelligence and Computer Technology Applications (ICAICTA)**, 2023.

[6] 山中稜斗, 北岡教英. 音声認識と複数の大規模言語モデルを活用した電子カルテ自動入力インタフェース. 情報処理, 2024.

[7] 国立情報学研究所大規模言語モデル研究開発センター. Llm-jp-3 1.8b・3.7b・13b の公開. https://llmc.nii.ac.jp/topics/post-707/, (2024-12 閲覧).

[8] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In **Proceedings of the 34th International Conference on Neural Information Processing Systems**, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[9] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual e5 text embeddings: A technical report. **arXiv preprint arXiv:2402.05672**, 2024.

[10] Chin-Yew Lin and Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. pp. 150–157, 2003.

[11] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. 2020.

[12] Junhyuk Seo, Dasol Choi, Taerim Kim, Won Chul Cha, Minha Kim, Haanju Yoo, Namkee Oh, YongJin Yi, Kye Hwa Lee, and Edward Choi. Evaluation framework of large language models in medical documentation: Development and usability study. **J Med Internet Res**, Vol. 26, p. e58329, Nov 2024.

[13] Yining Huang. A comprehensive survey on evaluating large language model applications in the medical industry. **arXiv preprint arXiv:2404.15777**, May 2024.

# A 付録

**表 3** アセスメント生成時のプロンプト

あなたは診察記録に対する assessment を作成する専門家です。以下の指示に従ってください。
- 提供された患者の発言と診察記録のみに基づいて assessment を作成してください。
- assessment は客観的な分析と解釈に焦点を当て、患者の状態を総合的に評価してください。
- 類似の記録は参考情報としてのみ使用し、そこに含まれる具体的なバイタルサインや数値を引用しないでください。
- 患者の発言と診察記録に明示されていない情報や推測は含めないでください。
- assessment 以外の関係のない文書は出力しないでください。
- 文体は常体（である調）を用いてください。

類似の記録:
{**context**}
患者の発言:
{**subjective**}
診察記録:
{**objective**}
assessment:

**表 4** 取得した関連 SOAP データ例

検索テキスト
subjective: 健康のため。奥さんには (禁煙外来) 内緒にして欲しい。前回はお酒の席で吸ってしまい失敗。
objective: たばこ 10〜20 本/日 20 年間喫煙 ブリンクマン指数 400 TDS5 点 CO 濃度 16ppmBW65.2kg
禁煙外来 2 回目 同居人に喫煙者いない チャンピックス処方あり 。

検索された関連文 SOAP データ
1. subjective: 先生にすすめられた。会社でたばこ吸ってる人は 1 割しかいないし吸える場所がほとんど無いです。
objective: たばこ 20 本/日 32 年間喫煙 ブリンクマン指数 640 TDS8 点 CO 濃度 18ppmBW76.7kg H30/11/6 アブ施行
同居人 (妻、娘) 喫煙者なし 自宅では外、車で喫煙している チャンピックス内服開始 本日 1 人で来院。
assessment: 妻の強いすすめもあり禁煙に対して前向き。たばこを減らし始めてきており努力しようとする様子が見られている。
チャンピックスに対して期待の発言聞かれているので効果が感じられなかった際の精神的サポート必要。
2. subjective: ぼちぼち、たばこ終わっても良いかな。妻と猫に副流煙が気になる
objective: 喫煙 20 本/日 34 年間 ブリンクマン指数 680 TDS4 点 AP、DM、HL あり CO 濃度 7ppm BW61.6kg 妻、息子 2 人の 4 人家族
息子 (長男、次男) も喫煙者 会社社長で社員も 1 割ほど喫煙者 Dr、Ns の話しを真剣に聞いている 10/2 チャンピックス内服開始。
assessment: 禁煙に対し真摯に取り組もうとしている。同居人で 2 人も喫煙者がおり環境的に辛いこと予測される。
3. subjective: 14 日は 5-6 本吸ったが、それからは吸っていない。今回はタバコを止めたいと思っている。
objective: 禁煙外来 1 回目。本数 20 本/日、56 年。ブリンクマン指数 1120。TDS 8 点。CO 濃度 0ppm、BW60.0kg。
assessment: 退院後は吸っても美味しくないため禁煙に対する前向きな姿勢がうかがえる。

生成したアセスメント
喫煙本数や年数からすると比較的軽度の喫煙者。禁煙に対する気持ちは前向き。
禁煙外来 2 回目の受診であるため前回の禁煙外来での失敗を聞いて禁煙に対するサポートを考える。

実際のアセスメント
前回禁煙失敗されており今回 2 回目のチャレンジ。前回再喫煙した要因が分かっているので同じ事を繰り返さないようにしていく必要あり。