大規模言語モデルと検索拡張生成を用いた医療対話からのカルテ経過記録生成

斉藤翼　山中稜斗　若林佑幸　北岡教英
豊橋技術科学大学
saito.tsubasa.xk@tut.jp

概要

経過記録の作成は勤務時間の約 2 割を占める重い業務負荷であり，効率化が急務である．本研究では，医療分野で広く利用されている SOAP 形式の経過記録を回診時の臨床対話を起点として生成する手法を提案し，その実用性と自動評価の妥当性を検証する．生成には GPT-5 を用い，Zero-Shot，過去記録の付与，RAG，Reranking RAG の 4 手法による比較検証を実施した．看護師 7 名による主観評価（5 点満点）の結果，対象患者の直近記録を Few-Shot 事例として活用する過去記録の付与手法が，平均スコア 3.6，標準偏差 0.73 と最も安定した品質を示した．一方，GPT-5.2 による自動評価（LLM-as-a-Judge）の検証では，看護師と比較して事実の正確性を平均 0.9 点厳しく評価する一方，構造の適切性を平均 0.4 点甘く判定するバイアスが確認された．本知見は，臨床現場における実用的な LLM 活用の指針となるとともに，自動評価における専門家的視点の介在の重要性を示唆している．

1　はじめに

臨床現場における経過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 [1] がある．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．

特に，入院患者に対する日々の回診業務では，病状の変遷や提供したケアの詳細を記載する経過記録の作成が必要である．この経過記録作成において中核をなすのが，SOAP 形式に代表される構造化された記録手法である．SOAP 形式は，患者の主観的情報（Subjective），客観的情報（Objective），評価（Assessment），計画（Plan）の 4 項目で構成され，単なる記録フォーマットにとどまらず臨床上の問題解決を促す思考のフレームワークとして機能する [2]．特に看護の現場においては個々の患者に対して立案された看護計画に沿ったケアを評価するフレームワークとして機能する [3]．

しかし，主観的情報（S）と客観的情報（O）を統合して看護計画に則した評価（A）を下し，評価に基づいた計画（P）を継続・修正する作業は高度な認知能力を要するため看護師の負担が大きい．これにより，患者ケア時間の不足や慢性的な超過勤務を招くだけでなく，増大する認知負荷によるバーンアウトや臨床情報の見落としといった医療安全に関わる深刻な問題を引き起こす可能性がある．

この問題を解決する技術として，近年，大規模言語モデル（LLM: Large Language Model）の臨床分野への応用が注目を集めている．LLM は高度な文脈理解や情報要約，自然な文章生成能力を有しており，複雑な臨床判断を伴う SOAP 記録作成の自動化において可能性を秘めている．

しかし，臨床記録を生成する既存研究の多くは既存記録の要約が中心であり，より実務的な業務である対話から直接臨床記録を生成する試みは発展途上にある．そこで本研究では，LLM を用いて国内の看護師と患者の臨床対話を起点とした経過記録を生成し，その正確性や実用性を検証する．

また，専門家による生成文の手動評価は正確性や安全性を担保する上で不可欠である一方，多大な時間とコストを要するため臨床実装のボトルネックとなっている．そのため，LLM-as-a-Judge を用いた評価と看護師による評価との差異を分析することで，専門家による高コストな評価プロセスの代替可能性についても検証する．

2　関連研究

2.1　臨床記録生成の動向

LLM を用いた臨床記録生成の研究として，Asgari ら [4] は臨床対話文から診療記録を生成して品質と安全性を評価する包括的なフレームワークを提案している．彼らは，生成文に含まれるハルシネーションや情報の欠落を詳細に分類し，プロンプトの調整によって重大なエラー率を人間が作成する診療記録以下の水準に抑制できる可能性を示している．

また，生成精度のさらなる向上を目的として RAG（Retrieval-Augmented Generation）技術の適用が進んでいる．RAG [5] は，外部のデータベースから関連情報を動的に検索しプロンプトに統合する手法であり，LLM にとって未知の情報や医療施設特有の記述ルールなどの文脈内学習が可能となる．

Liu ら [6] による分析では，臨床的な意思決定や質疑応答タスクにおいて RAG を適用したシステムが，LLM 単体の性能に比べオッズ比で 1.35 倍向上することが示されている．この結果は，RAG が臨床分野においてもハルシネーションを抑制し，情報の正確性と信頼性を確保するための手法として有効であることを示唆している．

2.2　評価指標の限界

生成された臨床記録の評価において，従来の ROUGE [7] や BERTScore [8] などの自動評価指標は表面的なテキストの類似性に依存するため，臨床的な文脈や専門知識，事実の正確性を十分に捉えきれないという課題が指摘されている [4]．そのため，医師や看護師など専門家による評価が適切だが，これには膨大な時間とコストを要する．これらの課題に対し，推論能力の高い LLM を評価者として活用する LLM-as-a-Judge [9] の検討が進んでいる．

Croxford ら [10, 11] は，LLM が生成した臨床要約の評価尺度「PDSQI-9」を開発し，LLM による評価が専門医による評価と高い相関を示すことを実証した．これにより，臨床記録の評価プロセスにおける高速化および効率化が期待される．

2.3　国内における医療情報処理研究

国内における医療文書を対象とした研究は，診療記録からの固有表現抽出 [12] や退院時要約から病名の自動分類 [13] といった既存のテキストからの情報抽出に主眼が置かれていた．近年の LLM の普及にともない，蓄積された診療記録から退院時要約を生成する試み [14] や，その基盤としての日本語医療 LLM 開発 [1] が発展している．しかし，既存研究の多くは診療情報を対象としており，看護業務の起点である回診現場での対話音声から直接経過記録を生成するアプローチは未だ発展の途上にある．

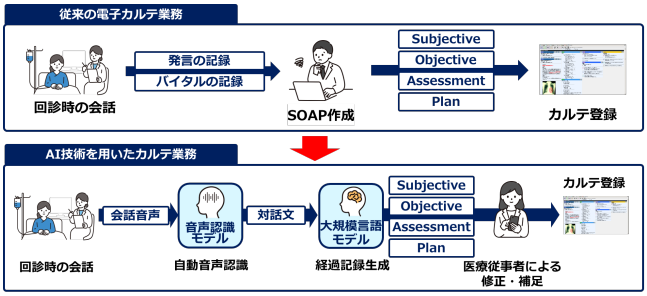

図 1　従来の経過記録生成手順と LLM を用いた生成手法システムの全体像

3　提案手法

3.1　システムの全体像

図 1 に示す本研究で提案するシステム [15] は，回診時における看護師と患者の対話音声を起点として経過記録を生成し，修正を経て電子カルテシステムへと登録する一連の枠組みである．システムのフローとして，まず回診時に収録した対話音声を音声認識技術によってテキスト化し LLM への入力とする．LLM によって生成された SOAP 案は，医療従事者による確認および修正後に電子カルテシステムへと登録される流れである．

本論文では，LLM を用いて対話記録から経過記録の SOAP 案を生成する手法について述べる．

3.2　経過記録の生成手法

対話記録から経過記録を生成するにあたって，図 2 に示す 4 種の生成戦略を比較検証した．すべての手法において，役割定義やスタイル指示を含むシステム指示文とプロンプト構造は共通である．また，LLM は OpenAI 社の GPT-5 [2]，ベクトルストア構築には text-embedding-3-large[3] を用いた．

1. Zero-Shot

追加の参照情報なしに，看護師と患者との対話文と患者の看護診断に基づいて立案した看護計画を

1）　https://llmc.nii.ac.jp/topics/sip-jmed-llm-2/
2）　https://platform.openai.com/docs/models/gpt-5
3）　https://platform.openai.com/docs/models/text-embedding-3-large

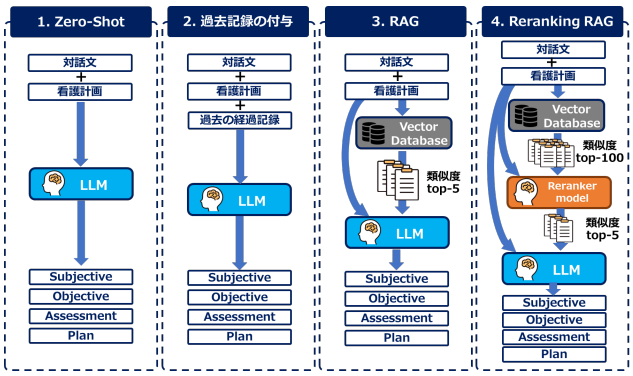

図 2　本研究における 4 種の経過記録生成手法フローチャート

入力として経過記録を生成する．臨床リスクの高い捏造や誤った推論などのハルシネーションを抑制するために，生成理由と根拠の引用を内部的に出力させる戦略を適用した．

2. 過去記録の付与

患者のもつ直近の経過記録を Few-Shot 事例としてプロンプトに統合し，病状の推移や施設特有の記述スタイルを LLM に付与する．

3. RAG

共同研究先の病院に蓄積された約 10 万件の経過記録を対象に類似症例をベクトル検索で取得する．対話文と看護計画を検索クエリとして検索し，類似度の高い上位 5 件の経過記録を取得してプロンプトに統合する．施設特有の記述スタイルや医療知識を LLM に付与することで正確性と実用性の向上を図る．

4. Reranking RAG

単純なベクトル検索による関連性の低い症例の混入を抑えるため，二段階検索プロセスを導入した．第一段階で看護計画をクエリとして類似する上位 100 件の看護記録とその経過記録を抽出し，第二段階で看護計画と対話文を組み合わせたクエリを用いて Cross-Encoding モデルの ruri-v3-reranker-310m [4] でリランキングを行う．リランキングによって特定した対話内容と整合性が高い 5 件の経過記録を活用する．

3.3　評価方法

生成した記録の品質を検証するため，看護師による主観評価と LLM を評価者として活用する LLM-as-a-Judge の枠組みを構築した．

主観評価実験には看護師 7 名が参加し，患者 3 名分の対話記録から 3.2 節で示した各 4 手法で生成した計 12 件の経過記録に対し手法名を伏せた状態で評価を行った．

評価指標は，LLM による臨床要約の評価尺度として開発された PDSQI-9 [10] を基盤として策定した．この指標に対し，共同研究先の看護師へのヒアリング結果や経過記録作成に関する専門書 [16] の定義を反映させ，経過記録の評価用に調整した指標を採用した．評価実験にあたっては，以下に示す 5 項目の指標を用いた 5 段階リッカート尺度でスコアリングを実施した．さらに，評価者間の一致度を検証するために級内相関係数（ICC(2,1)）を算出した．

正確性: 事実の捏造がなく，患者の発言やバイタル，実施されたケアの内容が元の対話に忠実であるか
適切性: SOAP の区分が適切であり，時系列が臨床経過として理解しやすい構成であるか
網羅性: 看護計画に関連する重要事項や症状の変化，介入，評価などの当日の経過に漏れがないか
有用性: 後続の看護師が状況を素早く把握し適切なケアを提供できるか
簡潔性: 冗長性を排し，カルテとして適切な文体で簡潔にまとめられているか

また，専門家による評価の代替可能性を検証するため，同一の指標を用いて LLM によるスコアリングを実施した．評価用 LLM には GPT-5.2 [5] を採用し，正確かつ慎重な判断を行わせるために reasoning effort を medium に設定した．評価プロンプトは看護師による評価結果やフィードバックを反映した，臨床現場の判断基準により近い視点を組み込んだ指示文とした．

4　結果と考察

4.1　看護師による主観評価結果

表 1 に看護師による主観評価の結果を示す．級内相関係数（ICC）は 0.286 となり，Koo ら [17] の基準では評価者間の信頼性は不十分な結果に留まった．この要因として，評価に参加した看護師の実務年数や記録頻度の多様性が評価基準のばらつきに影響した可能性が考えられる．

手法別の分析では，Zero-Shot 手法の平均スコアが 3.7 と最高値を記録したが，標準偏差は 1.17 と全手法で最大であり生成品質の不安定さが確認された．これに対し，過去記録の付与手法は平均 3.6，標準

4）　https://huggingface.co/cl-nagoya/ruri-v3-reranker-310m

5）　https://platform.openai.com/docs/models/gpt-5.2

表 1　看護師による主観評価結果［平均（標準偏差）］

手法	正確性	適切性	網羅性	有用性	簡潔性	平均
Zero-Shot	**4.0** (1.15)	3.3 (1.25)	**3.6** (1.13)	**3.6** (1.13)	**4.0** (1.15)	**3.7** (1.17)
過去記録の付与	3.7 (0.95)	**3.7** (**0.49**)	3.4 (**0.53**)	3.4 (**0.79**)	3.9 (0.90)	3.6 (0.73)
RAG	3.6 (**0.55**)	3.4 (0.55)	3.4 (0.89)	3.4 (0.89)	4.0 (**0.71**)	3.6 (**0.72**)
Reranking RAG	3.4 (**0.55**)	2.8 (0.84)	3.2 (0.84)	2.6 (0.89)	3.8 (0.84)	3.2 (0.79)

表 2　看護師評価と LLM 評価のスコア差分（看護師スコア – LLM スコア）［平均値の差（標準偏差の差）］

手法	正確性	適切性	網羅性	有用性	簡潔性	平均
Zero-Shot	1.0 (0.46)	−0.4 (0.80)	0.4 (0.36)	**0.3** (0.59)	−0.2 (0.76)	0.2 (0.82)
過去記録の付与	**0.8** (**0.30**)	**0.0** (−0.08)	**0.3** (**−0.08**)	0.4 (0.30)	**−0.1** (0.59)	0.3 (0.21)
RAG	1.0 (−0.35)	−0.3 (**0.00**)	0.8 (0.10)	0.6 (**0.03**)	**0.1** (**0.28**)	0.4 (**0.01**)
Reranking RAG	1.0 (−0.30)	−0.9 (0.27)	0.6 (−0.12)	**−0.3** (**0.03**)	**−0.1** (0.54)	**0.1** (0.08)

偏差 0.73 と出力が安定しており，特に適切性において標準偏差 0.49 と低いばらつきを示した．これは，施設特有の記述スタイルを Few-Shot 事例として参照させることが臨床的に整った記録生成に寄与することを示唆している．

一方，RAG 手法は過去記録の付与手法と同等の平均値を示したが，網羅性や有用性の標準偏差が 0.89 と高く検索された情報の不安定さが確認された．この要因として，検索された過去記録自体の品質に課題があった可能性が考えられる．検索された経過記録には，主観的情報（S）に患者発言以外の情報を記載しているものや，客観的情報（O）と評価（A）が混在している不適切な記録が含まれていた．LLM がこれらの記法を模倣した結果，適切性や網羅性，有用性が低下したと推察する．

さらに，Reranking RAG 手法の平均スコアが 3.2 と評価が最も低い結果となった．これは，看護計画の類似性のみに基づいて検索・リランキングを行ったことで，記述スタイルや情報の整理が不十分な記録が RAG 手法よりも優先的に検索され精度の悪化を招いた可能性がある．

以上の結果から，経過記録支援においては対象患者の過去記録をプロンプトに活用するアプローチが，実用上のリスクを抑制しつつ高い精度を維持する手法であることが明らかとなった．

4.2　LLM による自動評価結果

表 2 に，看護師による評価結果と LLM による評価結果のスコア差分（看護師スコア – LLM スコア）結果の平均値と標準偏差を示す．

分析の結果，LLM は看護師と比較して事実の正確性を平均 0.9 ポイント厳しく評価する一方で，SOAP 構造の適切性については平均 0.4 ポイント甘く判定するバイアスが確認された．特に，Reranking RAG 手法における適切性の差分は –0.9 と全手法中で最も顕著であり，看護師が臨床的な論理の不整合やノイズの混入を厳しく判定したのに対して LLM は形式的なセクション構成のみを評価してしまう可能性が示唆された．

また，個別のエラー項目では LLM は対話文中に実在する数値を捏造とみなしたり，カルテ記載が不要な情報を欠落として指摘したりする過検出の傾向が確認された．これは，看護師が臨床現場の文脈から記録すべき情報の優先順位を暗黙的に判断しているのに対し，LLM は情報源との逐次的な整合性のみに基づいて評価したことから生じた差異であると考えられる．

以上の結果から，LLM-as-a-Judge は評価の再現性確保には一定の有用性を持つものの，臨床的高度な判断においては依然として専門家による最終確認が不可欠であるといえる．

5　結論

本研究では看護業務の負担軽減をめざし，回診時の臨床対話を起点とした経過記録の生成手法を提案した．4 種類の生成手法を比較検証した結果，対象患者の過去記録を Few-Shot 事例としてプロンプトに活用する手法が安定した品質の記録生成に寄与することが明らかとなった．また，LLM-as-a-Judge による自動評価の検証においては正確性を厳しく判定する一方，適切性を甘く評価するバイアスや臨床的に重要度の低い情報の欠落を過検出する傾向が確認された．

今後の課題として，RAG のベクトルストアに用いる経過記録の改善や症例データの拡充による検証が挙げられる．また，自動評価のプロンプト改善も進め，専門家による評価の代替手段としての確立と臨床現場への安全な AI 実装の両立をめざす．

謝辞

本研究は，愛知県が公益財団法人科学技術交流財団に委託して実施した「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）」の助成を受けたものである．

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] 桑野タイ子. 看護記録検討の歩み. 日本看護研究学会雑誌, Vol. 15, No. 1, 1992.

[4] Elham Asgari, Nina Montaña-Brown, Magda Dubois, Saleh Khalil, Jasmine Balloch, Joshua Au Yeung, and Dominic Pimenta. A framework to assess clinical safety and hallucination rates of llms for medical text summarisation. **npj Digital Medicine**, Vol. 8, No. 1, p. 274, 2025.

[5] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. **Advances in neural information processing systems**, Vol. 33, pp. 9459–9474, 2020.

[6] Siru Liu, Allison B McCoy, and Adam Wright. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. **Journal of the American Medical Informatics Association**, Vol. 32, No. 4, pp. 605–615, 2025.

[7] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In **Text summarization branches out**, pp. 74–81, 2004.

[8] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In **International Conference on Learning Representations (ICLR)**, 2020.

[9] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. **Advances in neural information processing systems**, Vol. 36, pp. 46595–46623, 2023.

[10] Emma Croxford, Yanjun Gao, Nicholas Pellegrino, Karen Wong, Graham Wills, Elliot First, Miranda Schnier, Kyle Burton, Cris Ebby, Jillian Gorski, et al. Development and validation of the provider documentation summarization quality instrument for large language models. **Journal of the American Medical Informatics Association**, Vol. 32, No. 6, pp. 1050–1060, 2025.

[11] Emma Croxford, Yanjun Gao, Elliot First, Nicholas Pellegrino, Miranda Schnier, John Caskey, Madeline Oguss, Graham Wills, Guanhua Chen, Dmitriy Dligach, Matthew M. Churpek, Anoop Mayampurath, Frank Liao, Cherodeep Goswami, Karen K. Wong, Brian W. Patterson, and Majid Afshar. Evaluating clinical ai summaries with large language models as judges. **npj Digital Medicine**, Vol. 8, p. 640, 2025.

[12] 石田真捺, 谷中瞳, 戸次大介. 日本語症例テキストの複合語解析・推論システム medc2l. 自然言語処理, Vol. 30, No. 3, pp. 935–958, 2023.

[13] 津本周作, 平野章二, 岩田春子, 木村知広. 退院時要約の自動分類器の構築. 人工知能学会全国大会論文集第 31 回. 一般社団法人 人工知能学会, 2017.

[14] 石川開, 宇野裕, 石井亮, 定政邦彦, 柴田大作, 辻川剛範, 中川敦寛, 小山田昌史, 久保雅洋, 香取幸夫. 大規模言語モデルを用いて診療録から生成した治療経過サマリの評価. 人工知能学会全国大会論文集, 2024.

[15] Rikuto Yamanaka, Tsubasa Saito, Yukoh Wakabayashi, and Norihide Kitaoka. Speech input interface for electronic medical record supporting automatic soap generation using large language models. In **O-COCOSDA 2025: The 28th International Conference of Oriental COCOSDA**, pp. 258–263, 2025.

[16] 佐藤健太. 「型」が身につくカルテの書き方. 医学書院, 2015.

[17] Terry K Koo and Mae Y Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. **Journal of chiropractic medicine**, Vol. 15, No. 2, pp. 155–163, 2016.

A　付録

表 3　生成した経過記録の評価に用いた評価指標（5 段階リッカート尺度）

観点	評価（1 点：最低，5 点：最高）
事実の正確性（会話内容との一致）	1　重大な誤り（事実の捏造・改ざん）が複数ある 2　重大な誤りが 1 件以上ある 3　文脈やタイミング、数値などに一部不正確さがある 4　軽微なずれはあるが、全体として臨床的に許容できる 5　すべての記載が会話と整合している
SOAP 構造の適切性	1　S/O/A/P の区分が崩れており、何がどこに書いてあるか不明瞭 2　いくつかの要素が誤ったセクションに配置されている 3　大まかな区分は合っているが、S と O、A と P の混在が目立つ 4　ほとんどの情報が適切なセクションにあり、時系列もほぼ追える 5　S/O/A/P が明確に区別され、臨床経過と判断の流れが非常に分かりやすい
網羅性（重要情報の抜けのなさ）	1　重大な抜け（リスク悪化、重要な訴えなど）が複数ある 2　重大な抜けがあり、加えて他の重要情報も抜けている 3　重大な抜けが 1 点ある 4　重大な抜けはないが、「あると望ましい」情報がいくつか欠けている 5　重大・潜在的に重要な情報の抜けが特に見当たらない
有用性（他の看護師にとっての実用性）	1　ほとんど役に立たない（実際のケアに使えない） 2　一部は役立つが、重要な点が不足／不要な情報が多く混在 3　内容は概ね関連するが、詳細レベルが適切でない（細かすぎ・粗すぎ） 4　実務上ほぼ問題なく使えるが、改善余地はある 5　引継ぎ・ケア継続に非常に役立つ内容になっている
記述の簡潔性	1　口語体の多用や無駄な接続詞が多く、読むのに時間がかかる 2　同じ意味の内容が複数のセクションで繰り返されている。または敬語表現が多い 3　繰り返し（「～を行った。～も行った。」）が目立つが意味の重複はない 4　概ね簡潔だが、まだ短縮できる表現（不要な修飾語など）が一部残っている 5　専門用語や略語を適切に用いて最小限の文字数で最大限の情報を伝えている