ローカル LLM と RAG を用いた PSU 診療支援対話システムの検討

林海斗 [1]　桐山知彦 [2]　尾崎理沙 [2]　目黒巧巳 [3]
多屋優人 [2]　本庄勝 [2]　小林七彩 [4]　治徳大介 [4]　内田真人 [1]
[1] 早稲田大学　[2] 株式会社 KDDI 総合研究所　[3]KDDI 株式会社　[4] 東京科学大学
kaito.hayashi@akane.waseda.jp
{xth-kiriyama, xrs-ozaki, ta-meguro, ma-taya, ms-honjou}@kddi.com
{nanase.psyc, jitopsyc}@tmd.ac.jp　m.uchida@waseda.jp

概要

スマホ依存（Problematic Smartphone Use; PSU）患者に対する診療支援では、行動パターンの詳細な把握と、それに基づく専門的な情報提供が求められる。近年、スマホのログデータなどを分析することで患者の行動パターンを詳細に把握できるようになったが、アンケートデータや診療録などの複数の情報と組み合わせてそれらを総合的に解釈し、専門的な情報を提供することは容易ではない。そこで LLM を活用し、自動要約や情報提示による支援が期待されるが、プライバシー保護の観点から、患者個人の詳細な情報をパブリックな LLM に入力することが困難である。本研究では、ローカル LLM と Retrieval-Augmented Generation（RAG）を組み合わせたプライバシーに配慮した対話システムを提案する。具体的には、患者個別の利用履歴データ、診断経過、関連するドメイン知識を言語モデルでベクトル化し、知識ベースに格納する。プロンプトに基づき類似度検索で情報を抽出し、ローカル LLM の回答生成を支援するシステムを構築した。これにより、プライバシー保護とドメイン知識の活用を両立し、PSU 診療における意思決定を効果的に支援できることをケーススタディを通じて示す。

1　はじめに

スマホが普及する以前から，携帯電話の過度な使用は日常生活に悪影響を及ぼすことが指摘されていた [1]。スマホの登場により、過度な使用による悪影響は PSU として議論されるようになり、学力低下 [2, 3]，睡眠不足 [4]，メンタルヘルスの悪化 [5] など多方面で問題が指摘されている。PSU は依存症状や ADHD などの併存疾患を伴うケースも多く [6]、複雑な背景をもつ患者に対し、症状や行動履歴、スマホ依存評価指標などパーソナライズされた診療支援が求められる。

臨床的支援では、スマホのログデータ、患者のアンケートデータ、診療録など参照すべきデータが多く、医療従事者にとって負担となる場合がある。そこで、これらの情報を総合的に汲み取り、必要な情報を提示する LLM の活用を検討する。しかし、患者固有の詳細な情報を直接パブリックな LLM に入力することは、プライバシー保護の観点から困難な場合がある。クローズドな環境で完結するローカル LLM に関して、現在使用できるものはパブリック LLM と比較して、一般に性能が低いという懸念がある。したがって、ローカル LLM のみを利用するのでは、十分なドメイン知識を反映させた高品質な応答を得ることは容易ではない。

医療・臨床分野では診療録などを活用した意思決定支援システムが研究されてきたが [7, 8, 9]、PSU のような新しい領域では、包括的な知識基盤の構築と柔軟な情報参照が課題となる。また、近年の LLM 技術の発展は、自然言語処理を用いた知識抽出・生成を可能にしたものの、プライバシー保護と豊富な知識活用を両立するための仕組みは十分に確立されていない。

本研究では、上述した課題を解決するために、ローカルで稼働する LLM と RAG を組み合わせたプライベートチャットボットを提案する。具体的には、事前学習済みの言語モデルをローカル環境で動作させ、RAG システムにより、ドメイン知識を検索・統合しながら応答を生成する。関連情報の検索精度を向上させるため、患者 ID による絞り込みと

類似度検索、キーワード一致検索の重みづけ平均によるスコアリングを行った。

開発したチャットボットの有効性を検証するために、東京科学大学のネット依存外来に通院する患者のケーススタディを実施した。特に、事前学習モデルが知り得ないドメイン知識についても、RAG を介したドキュメント参照によってその情報を反映させた回答が得られることを確認した。これにより、ローカルに稼働する言語モデルと RAG の組み合わせは、機微情報を保護しながらドメイン知識を活用できる可能性を示す。

本研究は、医療現場など機微情報の取り扱いが厳格に求められる領域において、ローカル LLM を用いた診療支援システムの有用性を示唆するものである。一方で、応答精度のさらなる向上やハルシネーションなど、実運用に向けた課題も示された。今後はより精度の高いモデルの検討や、誤情報の検出と修正を行う仕組みの導入が必要となる。本研究で提案したチャットボットは、多面的かつ複合的な要因が関わる PSU 診療のみならず、医療・ヘルスケア分野全般における対話型支援ツールとしての幅広い応用が期待される。

2 関連研究

2.1 PSU に対する定義・評価・分析

PSU は、スマホ依存、問題のあるスマホ使用などと訳され、過度なスマートフォン使用が生活に悪影響を及ぼす状態を指す。評価には Smartphone Addiction Scale (SAS) [10] やその短縮版 SAS-SV [11] が用いられ、使用時間や頻度、影響度を測定することが一般的である。近年はスマホのログデータ分析も活発で、端末動作やアプリ名、センサ情報などから行動を推定する研究が進んでいる [12, 13]。臨床では、アンケート結果や過去の診療録を総合しつつ、最適な治療方針を検討するため、データの効率的な要約・提示が重要視されている。したがって、効率的かつ生産的な介入のために必要なデータを適切に要約・提示するツールの開発が期待されている。

2.2 LLM を用いた医療分野における応用

LLM を用いた対話システムは、患者との意思疎通を円滑化し、医療従事者の負担軽減にも期待されている。Alkhalaf ら [14] は、高齢者介護施設における栄養ケアに関する看護師の記録を要約し、栄養状態に関する重要な情報を抽出するために、Llama 2 13B モデルを用いた対話システムを開発した。Jeong ら [15] は、Self-BioRAG と呼ばれる、医学的な質問応答に特化した LLM フレームワークを提案した。Self-BioRAG は、医学的な質問に対して、関連する医学文献を検索し、その情報に基づいて回答を生成することができる。これらの研究は、LLM を用いた対話システムが、医療分野においても有用なツールとなりうることを示唆している。一方で、プライバシー保護に焦点を当てた環境の構築、ハルシネーションを防止する工夫に関しては改善の余地があると考えられる。本論文では、扱うデータは個人が特定されないよう処理されたデータを用いるだけでなく、LLM についてもローカル環境で使用することで、クラウド型の LLM のように外部に情報を漏らすことのない環境でシステムを構築した。詳細については 3 章で述べる。

3 提案手法

本研究では、PSU 診療支援を目的としたプライバシーに配慮した対話システムを提案する。ローカル LLM と RAG を組み合わせることで、患者個別の機微情報を外部に送信することなく、ドメイン知識を効果的に活用できる。安全なローカル環境で動作する本システムは、医師や研究者が PSU 患者の症例に関する情報や分析結果を要約・説明するためのチャットボットとして機能する。以下に説明するチャットボットの作成は、AI アプリ構築プラットフォームである Dify[16] をローカルマシンのコンテナ上に展開して行った。

3.1 システムの構成要素

提案システムは、主に 3 つの構成要素から成り立っている。1 つ目は、ドメイン知識ベースである。これは、患者個別の利用履歴、アンケート結果、医師の所見、学術論文や先行研究の知見などをテキストファイルとして保持するものである。テキストファイルは、「項目名：内容」という共通形式を採用し、後続のベクトル変換および検索を容易にした。

従来の RAG[17] では、質問と関連性の低い情報も含まれてしまう可能性があり、回答精度が低下する問題があった。そこで、本システムでは、チャットボットの冒頭に患者を入力するフィールドを設け、入力された患者 ID に関連づけられたフォルダから

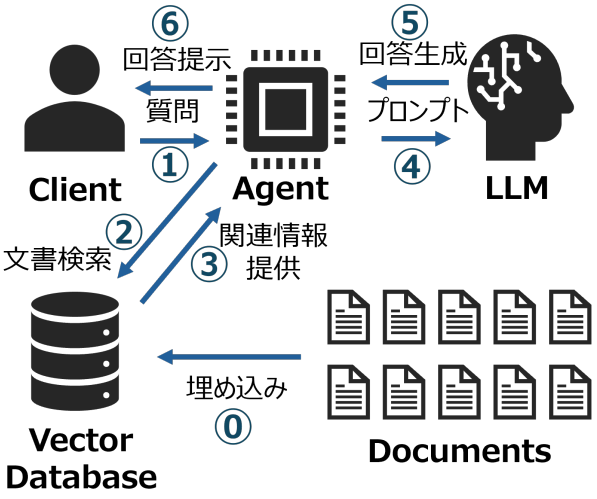

図 1　システム全体の概要

のみ知識を検索する仕組みにした。これにより、患者の診療に関連する情報のみを効率的に検索し、回答精度を向上させることができる。

2 つ目は、RAG システムである。RAG システムは、ドメイン知識ベースのテキストファイルを前処理し、単語や文脈の埋め込みベクトルを生成してデータベースに登録する。ユーザーがチャットボットに問い合わせを行うと、その問い合わせ文に含まれるキーワードとベクトルの類似度スコアの両方を用いて、関連度の高いテキストを検索・抽出する。具体的には、テキストファイルをチャンクに分割し、ローカル LLM による埋め込みを行ってベクトル表現としてデータベースに格納する。本研究では、ollama で起動したモデル Llama3.2 3B を用いて埋め込みベクトルを生成し、ベクトルデータベースに登録することで RAG システムを構築する。

3 つ目は、ローカル LLM である。Llama3.2 3B モデルを ollama を用いて Docker 上で起動・稼働させ、完全にローカル環境で動作させる。問い合わせ内容や RAG システムにより抽出されたテキストをプロンプトとして与え、要約や説明を行う。ローカル LLM サーバーを構築することで、機微情報の外部流出リスクを抑制する。

3.2　システムの動作フロー

提案システムは、図 1 に示すように、以下の手順で動作する。まず、ユーザーはチャットボットに患者 ID を入力する。次に、ユーザーが、システムに対して自然言語で質問を入力する。RAG システムが、入力された質問に含まれるキーワードを抽出し、抽出されたキーワードとベクトル化されたドメイン知識との間で、キーワード一致度とベクトル内積による類似度を計算し、関連性の高い情報をランキングする。スコアリングはキーワード一致度とベクトル類似度の合算によって行い、上位 K 件のチャンクを取得する。この際、RAG システムは、入力された患者 ID に紐づくフォルダ内のテキストデータのみを検索対象とする。最後に、ローカル LLM が、ユーザーの質問と RAG システムによって抽出された情報に基づいて、回答を生成する。取得したチャンクを関連ドキュメントとしてプロンプトに付加し、ローカル LLM にテキスト生成を依頼することで、ユーザーの質問に対する回答を生成する。

3.3　システムの拡張性と期待される効果

提案システムは、すべての処理をローカル環境で行うため、医療情報が外部に送信されるリスクがない。また、RAG により学習済みモデル外の知識を検索・統合することで、医師や研究者のニーズに即した応答を提供できる柔軟性を有している。

本システムは、ドメイン知識を活用することで、より適切な診療支援が可能となり、必要な情報を抽出・提示することで医師の負担軽減にも貢献する。また、必要に応じて医療従事者が都度新しいドメイン知識を RAG システムに登録できる点も迅速な更新が可能であり、有用である。追記する情報には診療録や診断情報など医療従事者の日常的な業務の中で作成される文書の一部をそのまま含めることができるため業務負荷の軽減にも寄与すると考えられる。LLM に対するプロンプト設計を工夫することで、要約や説明、プラン提案など多様なタスクに対応できる。また、LLM の精度向上に伴い適用するローカル LLM を付け替えられるなど拡張性も確保した。さらに、患者 ID による検索範囲の限定により、従来の RAG システムよりも回答精度が向上することが期待される。次章では、ケーススタディを通じて、提案システムの有効性と応用可能性を検証する。

4　ケーススタディ

4.1　検証方法

提案システムの有効性を検証するために、東京科学大学のネット依存外来に通院する患者 5 名のデータを用いてケーススタディを実施した。具体的

には、患者のアンケートデータ、医師によるコメント、スマホの分析結果などを格納したデータベースから、質問への回答に適する情報を検索し提示するか、また回答生成にあたり適切な情報が参照されているかを検証した。

3 章で述べたように、本システムでは、まず冒頭で質問対象の患者の患者 ID を入力する。その後、自由記述で質問を入力すると、システムはデータベースから関連する情報を検索し、LLM を用いて回答を生成する。生成された回答には、参照した情報が提示されるようになっている。このような設計とすることで、別の患者の情報を参照してしまう問題や参照している出典が正しいかを確認できるようになっている。

4.2 結果

チャットボットとの会話例を図 2 に示す。提案システムは患者の状態に関する質問に対して、適切な回答を生成し、その際にデータベースから適切な情報を参照していることが確認された。例えば、図 2 の例では、「この患者の就業状況について教えてください。」という質問に対して、システムは患者の仕事内容、仕事中の問題点、改善状況などを要約して回答している。また、「この患者のスマホ利用状況についてまとめてください。」という質問に対しては、患者のスマホ使用に関する情報を整理し回答を生成している。これらの結果から、提案システムは、RAG を用いてドメイン知識ベースから適切な情報を検索し、LLM を用いて自然言語で回答を生成することで、PSU 診療において医師の意思決定を支援できる可能性が示唆された。なお、これらの実行はローカルの PC（メモリ: 24GB、CPU: Intel Core i5 135000、GPU: NVIDIA GeForce RTX 4060）にて行った。回答生成に必要な時間は約 30〜90 秒であった。

4.3 考察

本ケーススタディでは、提案システムが患者のプライバシーを保護しつつ、医師の質問に答えるために必要な情報を効果的に抽出できることを示した。特に、患者 ID を入力することで、患者のみに関連する情報を絞り込んで検索できる点は、従来の RAG システムにおける課題を解決する上で有効に働いたと考えられる。また、一般的なデスクトップ PC のスペックでも十分に稼働が見込める点は、サーバーなどの構築を不要とするため有用性が高いと言える。

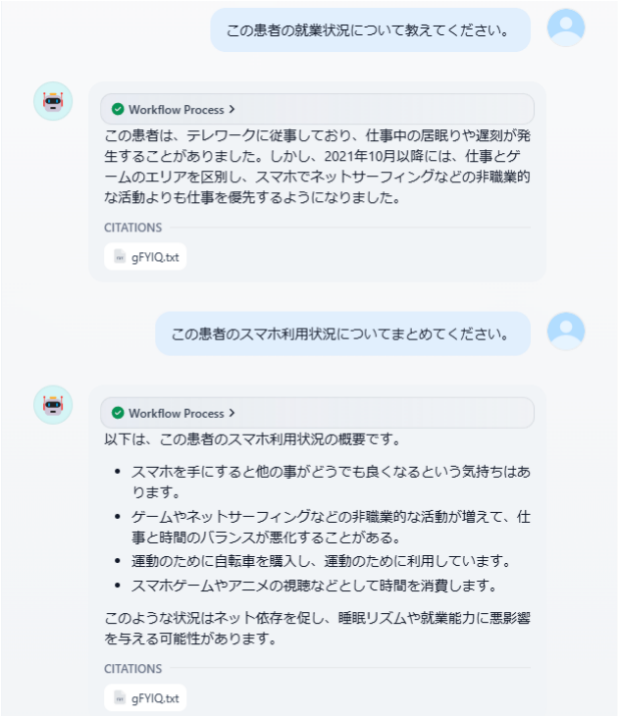

図 2 チャットボットの稼働例

る。しかし、現状では、データベースに格納されている情報の質や量に依存する部分が大きく、より効果的なシステムにするためには、データの充実化、検索精度の向上、LLM の性能向上など、更なる改善が必要である。

5 おわりに

本研究では、PSU 診療支援を目的とした、プライバシーに配慮した対話システムを提案した。ローカル LLM と RAG を組み合わせることで、外部へ情報を開示することのない環境で、ドメイン知識を効果的に活用できることを示した。ケーススタディを通じて、ユーザーの質問に対し適切な回答を生成し、その際にデータベースから適切な情報を参照していることを確認した。提案システムは、PSU 診療において医師の意思決定を支援できる可能性を有する。

今後の課題としては、データベースの充実化、検索精度の向上、LLM の性能向上などが挙げられる。また、より多くの患者データを活用した評価実験を行い、提案システムの有効性をさらに検証する必要がある。

参考文献

[1] Joël Billieux, Martial Van der Linden, and Lucien Rochat. The role of impulsivity in actual and problematic use of the mobile phone. **Applied Cognitive Psychology**, Vol. 22, No. 9, pp. 1195–1210, 2008.

[2] Sohel Ahmed, Nikita Pokhrel, Swastik Roy, and Asir Samuel. Impact of nomophobia: A nondrug addiction among students of physiotherapy course using an online cross-sectional survey. **Indian Journal of Psychiatry**, Vol. 61, No. 1, pp. 77–80, 2019.

[3] Heather Winskel, Tae-Hoon Kim, Lauren Kardash, and Ivanka Belic. Smartphone use and study behavior: A korean and australian comparison. **Heliyon**, Vol. 5, No. 7, p. e02158, 2019.

[4] Qing-Qi Liu, Zong-Kui Zhou, Xiu-Juan Yang, Fan-Chang Kong, Geng-Feng Niu, and Cui-Ying Fan. Mobile phone addiction and sleep quality among Chinese adolescents: A moderated mediation model. **Computers in Human Behavior**, Vol. 72, pp. 108–114, 2017.

[5] Jon D. Elhai, Jason C. Levine, Robert D. Dvorak, and Brian J. Hall. Non-social features of smartphone use are most related to depression, anxiety and problematic smartphone use. **Computers in Human Behavior**, Vol. 69, pp. 75–82, 2017.

[6] Maria Panagiotidi and Paul Overton. Attention deficit hyperactivity symptoms predict problematic mobile phone use. **Current psychology (New Brunswick, N.J.)**, Vol. 41, No. 5, pp. 2765–2771, 2022.

[7] Ann R Punnoose. Electronic health records and clinical decision support systems: Impact on national ambulatory care quality. **JAMA : the journal of the American Medical Association**, Vol. 306, No. 4, pp. 360–, 2011.

[8] Joseph Finkelstein, Aileen Gabriel, Susanna Schmer, Tuyet-Trinh Truong, and Andrew Dunn. Identifying facilitators and barriers to implementation of ai-assisted clinical decision support in an electronic health record system. **Journal of medical systems**, Vol. 48, No. 1, pp. 89–, 2024.

[9] Christian Castaneda, Kip Nalley, Ciaran Mannion, Pritish Bhattacharyya, Patrick Blake, Andrew Pecora, Andre Goy, and K Stephen Suh. Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine. **Journal of clinical bioinformatics**, Vol. 5, No. 1, pp. 4–4, 2015.

[10] Min Kwon, Joon-Yeop Lee, Wang-Youn Won, Jae-Woo Park, Jung-Ah Min, Changtae Hahn, Xinyu Gu, Ji-Hye Choi, and Dai-Jin Kim. Development and Validation of a Smartphone Addiction Scale (SAS). **PLoS ONE**, Vol. 8, No. 2, p. e56936, 2013.

[11] Min Kwon, Dai Jin Kim, Hyun Cho, and Soo Yang. The Smartphone Addiction Scale: Development and Validation of a Short Version for Adolescents. **PLoS ONE**, Vol. 8, No. 12, p. e83558, 2013.

[12] Shun Furuawa, Toshitaka Hamamura, Akio Yoneyama, Masaru Honjo, Nanase Kobayashi, Daisuke Jitoku, , and Masato Uchida. Predicting device usage patterns in patients with problematic smartphone use through individualized hidden markov models. In **2024 IEEE International Conference on E-health Networking, Application & Services (Healthcom)**, 2024.

[13] Kaito Hayashi, Takumi Meguro, Toshitaka Hamamura, Masato Taya, Masaru Honjo, Nanase Kobayashi, Daisuke Jitoku, and Masato Uchida. Estimating smartphone orientation for visualizing behavior around sleep periods in psu patients. In **2024 IEEE International Conference on E-health Networking, Application & Services (Healthcom)**, 2024.

[14] Mohammad Alkhalaf, Ping Yu, Mengyang Yin, and Chao Deng. Applying generative ai with retrieval augmented generation to summarize and extract key clinical information from electronic health records. **Journal of biomedical informatics**, Vol. 156, pp. 104662–, 2024.

[15] Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. **Bioinformatics (Oxford, England)**, Vol. 40, No. Supplement1, pp. i119–i129, 2024.

[16] Dify Community. Dify: A no-code platform for ai-powered application development, Accessed: 2024-12-31. https://dify.ai.

[17] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen Tau Yih, Tim Rocktaschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In **Advances in Neural Information Processing Systems**, Vol. 2020-, 2020.