検索拡張生成が信頼度に及ぼす影響：医療分野における分析

尾崎 慎太郎 [1] 加藤 優汰 [2] 馮 思遠 [2] 富田 雅代 [2] 林 和樹 [1]
小原 涼馬 [3] 小山田 昌史 [3] 林 克彦 [2] 上垣外 英剛 [1] 渡辺 太郎 [1]
[1] 奈良先端科学技術大学院大学 [2] 東京大学 [3]NEC データサイエンスラボラトリー
ozaki.shintaro.ou6@naist.ac.jp {ryoma-obara, oyamada}@nec.com
{hayashi.kazuki.hl4, kamigaito.h, taro}@is.naist.jp
{ukato6209, 9445233883, tomita-masayo732,
katsuhiko-hayashi}@g.ecc.u-tokyo.ac.jp

概要

検索拡張生成 (RAG) は, 外部情報を活用することで, 大規模言語モデル (LLM) の知識を補完し, 質問に対する応答の精度を改善させる. この手法は, 最新の情報を活用できる利点を活かし, 多様な分野で広く応用されている. 先行研究の多くは性能改善に注力する一方, RAG を利用した際の出力の信頼度に関する特性については十分に研究されていない. 金融, 医療, 医学などの出力への信頼性が強く要求される分野で, 出力の信頼度を分析することは重要な課題である. 本研究では, 医療分野における多様な課題および推論モデルを対象に, RAG が信頼度に与える影響を調査する. 具体的には, LLM の予測確率を出力として扱い, 期待較正誤差と適応較正誤差を出力確率に基づいて計算することで信頼度を評価する. さらに, プロンプト内の取得文書の順序が信頼度に影響するかについても分析を行う. 結果として, モデル, 取得する書類数や埋め込み数などの設定, および入力プロンプトの形式に応じて, 信頼度と精度に大きな変動が見られることが明らかとなった. これらの結果は, 特定のモデルおよび条件に基づいて RAG で用いる構成を最適化する必要性を強調している.[1)]

1 はじめに

検索拡張生成 (RAG) [1] は, 大規模言語モデル (LLM) [2, 3, 4, 5] の知識を補完する手法として活用される. 外部情報を用いることで, RAG は応答の正確性と質問への適合性を改善させ, 多くの分野で利用されている. 注目される分野として, 情報の信頼性が重要な金融 [6, 7] や医療 [8] が挙げられる. 研究者は RAG を用いた LLM の性能改善 [9] を積極的に探求しているが, 出力の信頼性に焦点を当てた分析は依然として限定的である. RAG は回答の正確性を改善させる一方で, 検索結果を根拠にモデルが不正解であるにも関わらず過剰な自信を示す可能性がある.

本研究では,「RAG を通じて正解に関する情報を取得することで, モデルが確信度の高い出力を行う」という仮説を立てた. この仮説に基づき, RAG が与える信頼性への影響を分析するため, 2 つの研究課題 (RQs) を定義した. **RQ1** では, 医療分野に焦点を当て, 複数の QA タスクとデータセットを使用して, さまざまなモデルを用いて評価を行った. さらに, モデルのパラメータ数, 取得する文書の数, インデックス作成時の埋め込みサイズなどの要因が性能に与える影響についても検証した. **RQ2** では, プロンプト内における取得文書の配置が信頼性と正確性の関係にどのような影響を与えるかを分析した. 例えば, 質問の前または選択肢の後に RAG で取得した文書を挿入する場面を分析する. この分析は, 長いプロンプトの中間部分の情報が見落とされる Lost in the Middle 現象 [10] に着想を得ている. さらに, RAG が直面する多様な場面を体系的に再現するため, 無関係な文書を含める場合や正答に直接関連する文書のみを提供する場合など, 文書内容を操作した分析を行った. MedMCQA [11], MedQA [12], PubMedQA [13], および医療関連の質問のみに絞った MMLU [14, 15] といった医療 QA データセットを用いた結果, 検索モデルや推論モデル, その他の条件といった構成の違いが出力の信頼性に影響を与えることが明らかになっ

1) コードは https://github.com/naist-nlp/CC_RAG.git.

た. 特に, RAG がノイズを引き起こし, 一部の実験で信頼性を悪化させる例が確認され, 特定の構成下では RAG が信頼性と正確性の両方を悪化させる可能性があることを示した. さらに, 入力の形式が出力に影響を与えることが明らかとなり, 信頼性の最適な設定はモデルや構成によって大きく異なることが示唆された. 精度の観点では, 関連性の高い文書を回答として意図的に提供した場合, 性能が改善した一方で, 無関係な文書のみを挿入した場合には性能が悪化した. これらの結果は, 出力の信頼性を確保するためには文書の慎重な選択が重要であることを示しており, 正解に関連する文書を取得することで信頼性が改善するという仮説と矛盾することを示唆する.

2 分析手法

モデルが予測した確率に基づいて信頼度を計算し, RAG による信頼度の較正を分析する. Xiong らが考案した MIRAGE [8] の設計に従い, 質問プロンプトとその選択肢 (例：四択問題) を連結して入力を作成する. RQ1 のプロンプトに関して, 図 1 にある「Prompt w/o RAG」と呼ぶ RAG を使用しない場合のプロンプトと, RAG を適用する質問の前に文書を連結する「Prompt w/ RAG with Pre-Question」を使用する. 選択肢の回答をモデルが直接生成する場合, 同一条件下であっても指標に差異が生じるため, 評価が複雑化する [8, 16, 17]. ことを踏まえ付録の数式 1 に示すように, 与えられた選択肢から最も妥当な選択肢を予測する手法を採用する. RQ2 では, RAG で取得した文書をプロンプトに挿入する際の最適な位置について分析する. 具体的には, 以下の 3 つのパターンを評価する: (1) 質問の前 (Pre-Q と表記); (2) 質問と選択肢の間 (Aft-Q と表記); および (3) 選択肢の後 (Aft-C と表記) である. さらに, 取得文書の位置が与える影響に焦点を当てるため, 質問の答えに関連する文書を恣意的に使用する. 評価には, 質問とその質問の答えに関連する解説を含む MedMCQA を採用し, 以下の 3 つの状況で検証する: (1) 回答に関連する解説のみを挿入する (Ans1 と表記); (2) 正答に関連する解説と無関係な文書 2 つを組み合わせて挿入する (A1-O2 と表記); (3) 無関係な文書 3 つを挿入する (Oth3 と表記).

3 実験設定

データセット 医療分野における RAG の応用に焦点を当て, QA データセットには MedMCQA [11],

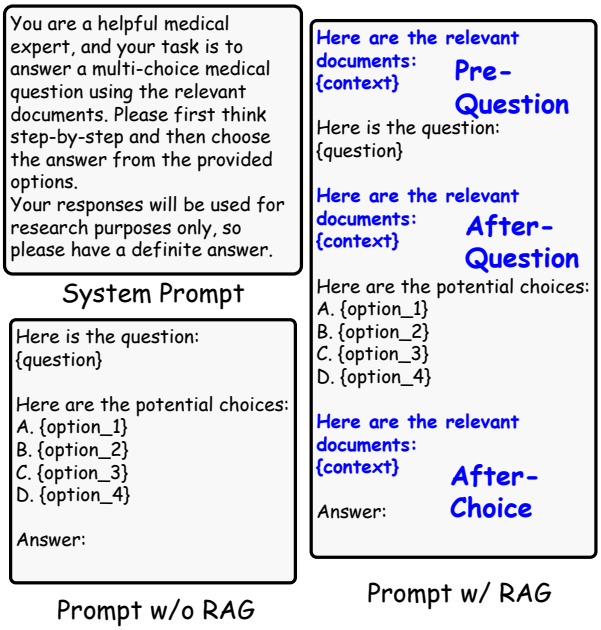

図 1 RQ1 では, Prompt w/o RAG と Prompt w/ RAG (Pre-Question) のものを, RQ2 では Prompt w/ RAG を使用した. それぞれにシステムプロンプトを連結する.

設定	適用する値
推論モデル	{Phi-3.5, PMC-LLaMA, MEDTIRON, LLaMA3.1}
検索モデル	{MedCPT, Contriever, BM25}
埋め込みサイズ	{256, 512}
取得ドキュメント数	{1, 3}

表 1 モデルとパラメータの組み合わせ, 埋め込みサイズ, 取得する文書の数, および検索モデルが信頼性に与える影響を分析した. これらすべての組み合わせを用いる.

MedQA [12], PubMedQA [13], および医療分野に関連する質問のみを対象とした MMLU [14, 15] を使用し, 各サイズは付録 A に記載している. MedMCQA には, 各質問の正答を裏付ける解説が含まれており, RQ2 でこの解説を正解文書として扱うことで検索結果における擬似的な上界を設定し, データセットを活用する.

推論モデル Phi-3.5 (3.8B) [18], PMC-LLaMA (13B) [17], LLaMA3 (70B) [4], および MEDITRON (70B) [16] を選定した. 70B のモデルには 4 ビット量子化を適用し, PMC-LLaMA には半精度量子化を用いて確率を計算した. 詳細は付録 A に記載した.

ドキュメント データストアには, StatPearls[2] および Textbooks [19] を選定した. これらは医療分野の文書を含んでおり, Xiong ら [8] が公開しているものを利用した. 埋め込みサイズ (256 および 512 トークン) と取得数 (top-1,3) の組み合わせを採用し, RAG

2) https://www.statpearls.com/

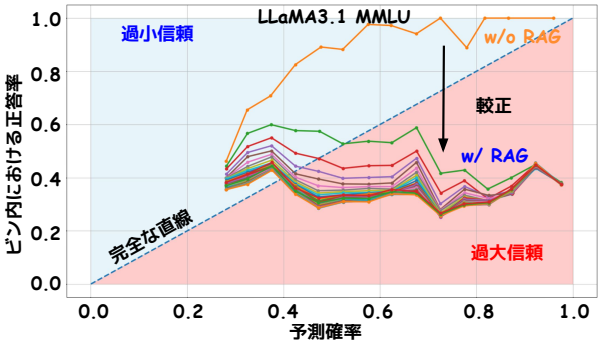

図 2　$y = x$ が完全な一致を表す. x 軸は予測確率を示し, y 軸は各ビン (サイズは 20) 内での精度. LLaMA3.1 では, RAG を適用すると信頼性が改善する結果が示された.

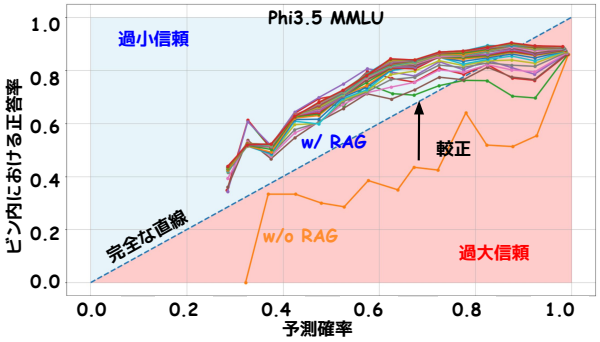

図 3　MMLU における Phi-3.5 を用いた信頼性較正曲線では, RAG を適用することで信頼性が改善した.

構成に応じて信頼度がどのように変化するかを分析する. 詳細な組み合わせは表 1 に示されている.

テンプレート　MIRAGE [8] の手法を修正し, 思考の連鎖 (CoT) [20] を排除して確率を直接計算可能な形式とした. RQ1 では, 図 1 に示されるテンプレートを用い, RAG を使用しない (Prompt w/o RAG) 場合と, RAG を用いた設定 (Prompt w/ RAG) では Pre-Question を用いる. RQ2 では, 文書の挿入順序に基づく挙動の違いを分析するため, 図 1 における (Prompt w/ RAG) に類似した 3 つのパターンを準備した. RQ1 および RQ2 のいずれにおいても, 各プロンプトの冒頭にシステムプロンプトを連結して実験を実施した.

RAG の設定　検索モデルに MedCPT [21], Contriever [22], および BM25 [23, 24] を選定した. 信頼度の詳細な分析のために, 密検索モデルでは埋め込みサイズ (256 および 512 トークン) を変化させ, 取得文書数 (top-1, 3) の組み合わせを分析した. 密モデルでは距離関数として内積を使用し, 詳細は表 5 に記載した. 取得数と埋め込みサイズの組み合わせを T1-256, T1-512, T3-256, T3-512 と表現する.

評価尺度　本研究では, ECE および ACE を用いる.

(1) **期待較正誤差 (ECE)** [25, 26]: 予測確率と実際の正答率の差を測定する指標. 予測確率の範囲を複数のビンに分割し, 各ビン内で観測された正答率と予測された信頼度の差を計算する. ECE は各ビンの正答率と信頼度の絶対差の加重平均として計算され, 重みは各ビン内の数の割合である.

表 2　RQ1 における MMLU の結果. RAG を用いた際はドキュメントごとの平均値を算出している.

モデル	種類	MMLU T1-256	T1-512	T3-256	T3-512
ECE ↓					
Phi	w/o RAG	0.14			
	AVG	0.17	0.17	0.17	0.16
PMC	w/o RAG	0.08			
	AVG	0.14	0.14	0.14	0.14
LLaMA	w/o RAG	0.32			
	AVG	0.25	0.25	0.26	0.26
MEDITRON	w/o RAG	0.05			
	AVG	0.16	0.16	0.16	0.16
ACE ↓					
Phi	w/o RAG	0.18			
	AVG	0.17	0.17	0.17	0.16
PMC	w/o RAG	0.07			
	AVG	0.14	0.14	0.14	0.14
LLaMA	w/o RAG	0.32			
	AVG	0.24	0.24	0.25	0.25
MEDITRON	w/o RAG	0.05			
	AVG	0.15	0.15	0.15	0.15

(2) **適応較正誤差 (ACE)** [27]: ECE の欠点を補うために提案された指標で, 特にサンプル数が少ないビンのリスクを軽減することを目的としている. ACE では, 各ビン内の数が一定になるようにビニングを行い, 各ビン内での較正誤差の評価を安定にする.

4　結果

RQ1 の実験結果を表 2 に示す. 取得文書数 (Top-k) と埋め込みサイズ (256, 512) の組み合わせを T1-256, T1-512, T3-256, T3-512 と表現する. ECE に関しては, LLaMA3.1 が MMLU において信頼度の改善を示した一方で, 他のモデルでは悪化が見られた. ACE については, Phi-3.5 が改善を示したが, その効果は限定的である. 図 2 と 3 は, 信頼度較正曲線である. MMLU において, RAG は LLaMA3.1 を過信状態にし, 一方で Phi-3.5 を信頼度が過小な状態にすることが分かり, 同じタスクでもモデルによって異なる挙動を示すことがわかる. **RQ2** の結果である表 3 は, 取得した文書の挿入順序を変えた場合の信頼度の挙動への影響を示す. プロンプトの種類は図 1 に基づき, Pre-Q, Aft-Q, Aft-C はそれぞれ Pre-Question, After-Question, After-Choice を表す. Ans1 は正解の説明のみを含み, A1-O2 は正解のテキストと 2 つの無

表 3　RQ2 における MedMCQA の一部を使用した結果

モデル	パターン	MedMCQA (抽出) ECE ↓ Ans1	A1-Q2	Oth3	ACE ↓ Ans1	A1-Q2	Oth3	Accuracy ↑ Ans1	A1-Q2	Oth3
Phi	w/o RAG		0.05			0.06			51.52	
	A	0.06	0.07	0.32	0.10	0.10	0.34	86.08	84.95	51.81
	B	0.04	0.06	0.42	0.08	0.09	0.43	88.49	85.95	39.39
	C	0.04	0.13	0.41	0.08	0.16	0.42	87.35	76.11	44.33
PMC	w/o RAG		0.16			0.16			38.11	
	A	0.01	0.01	0.04	0.05	0.04	0.04	32.73	31.64	26.97
	B	0.01	0.02	0.06	0.04	0.03	0.05	32.91	30.01	26.97
	C	0.02	0.05	0.05	0.03	0.04	0.04	33.45	28.74	28.06
LLaMA3.1	w/o RAG		0.20			0.20			58.98	
	A	0.14	0.14	0.14	0.11	0.12	0.12	20.90	20.94	20.94
	B	0.14	0.14	0.14	0.12	0.12	0.12	20.90	20.90	20.94
	C	0.14	0.14	0.14	0.12	0.12	0.12	20.94	20.94	20.94
MEDITRON	w/o RAG		0.07			0.06			35.52	
	A	0.18	0.08	0.09	0.18	0.08	0.09	67.72	54.03	36.04
	B	0.16	0.09	0.15	0.16	0.09	0.15	66.68	47.14	31.96
	C	0.15	0.07	0.09	0.15	0.07	0.09	62.83	34.18	31.91

関係な文書, Oth3 は完全に無関係な 3 つの文書で構成されている. 仮説に反して, 文書を意図的に挿入した場合でも, すべてのモデルで ECE や ACE が悪化するわけではないことがわかった. さらに, 精度が改善したモデルでは, ECE の減少も確認した.

5　分析と議論

RAG は信頼度を較正するか　表 2 の結果は, RAG が信頼度較正に与える影響がモデルごとに異なることを示している. LLaMA3.1 では, RAG が ECE および ACE を共に悪化させ, 信頼度が改善されている. 一方, 同一の設定で他のモデルを分析した結果, ECE は全体的に悪化している. このことは, 信頼度の観点から, RAG には慎重な設定が必要であることを示唆している. 図 2 と 3 は信頼度較正曲線を示している. RAG が信頼度を較正する可能性を示唆している一方で, モデルの挙動が大きく異なることを示す.

どの入力が信頼度に関して最適か　文書の挿入位置が信頼度に与える影響に関する結果は, 表 3 に示されているように, 選択肢の後に文書を配置する (Aft-C) が信頼度を改善させる最適なアプローチであることを示唆している. しかし, 精度の観点から見ると, これは全パターン中で最も低い性能を示しており, さらなる実験により精度と信頼度の間にトレードオフが存在することが支持された.

答えを含むドキュメントの影響　正答を含む文書を意図的に挿入した場合, 精度が改善することが観察された. これは, 少なくとも正答の根拠を含む文書を取得できれば,RAG が精度を改善させることを示す. 一方で, 無関係な文書のみが含まれる場合, 表 3 が示すように, Phi-3.5 において ECE が急激に増加した. これは, パラメータ数が少ないモデルは無関係な入力によって混乱し, 内部知識が損なわれる可能性を示唆する. 一方, 70B のようなパラメータ数が多いモデルは 取得した文書が無関係であるかどうかを判断する能力を示し, このような条件下でも信頼度を維持することができる.

6　おわりに

RAG が確率を用いて信頼度を較正するかどうか (RQ1) を分析し, RAG が LLM の信頼度を改善すると仮説した. 取得した文書のプロンプトテンプレート内の位置が信頼度と精度に与える影響を分析することで, RAG において情報を見落とす現象 (Lost in the Middle) が発生する可能性を分析した (RQ2). RQ1 の結果では, RAG の挙動が推論モデル, 検索モデルや埋め込みサイズおよびその設定に対して非常に敏感であることが明らかになり, 適切な設定を慎重に選択する必要性が示された. RQ2 では, 精度が改善する場合に信頼度が悪化し, 信頼度が改善する場合に精度が悪化するというトレードオフ関係が確認された. これらより, 状況によって最適な設定が異なるため, RAG を利用する際に文書の選択を取得数や配置を慎重に行うことの重要性を明確にした.

参考文献

[1] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, 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.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. **arXiv preprint arXiv:2303.08774**, 2023.

[3] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. **arXiv preprint arXiv:2403.05530**, 2024.

[4] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. **arXiv preprint arXiv:2407.21783**, 2024.

[5] Akiko Aizawa, Eiji Aramaki, Bowen Chen, Fei Cheng, Hiroyuki Deguchi, Rintaro Enomoto, Kazuki Fujii, Kensuke Fukumoto, Takuya Fukushima, Namgi Han, et al. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms. **arXiv preprint arXiv:2407.03963**, 2024.

[6] Antonio Jimeno Yepes, Yao You, Jan Milczek, Sebastian Laverde, and Renyu Li. Financial report chunking for effective retrieval augmented generation. **arXiv preprint arXiv:2402.05131**, 2024.

[7] Spurthi Setty, Harsh Thakkar, Alyssa Lee, and Eden Chung. Improving retrieval for rag based question answering models on financial documents. **arXiv preprint arXiv:2404.07221**, 2024.

[8] Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Findings of the Association for Computational Linguistics: ACL 2024**, pp. 6233–6251, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[9] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024.

[10] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. **Transactions of the Association for Computational Linguistics**, Vol. 12, pp. 157–173, 2024.

[11] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann, editors, **Proceedings of the Conference on Health, Inference, and Learning**, Vol. 174 of **Proceedings of Machine Learning Research**, pp. 248–260. PMLR, 07–08 Apr 2022.

[12] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have. **A Large-scale Open Domain Question Answering Dataset from Medical Exams. arXiv [cs. CL]**, 2020.

[13] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 2567–2577, Hong Kong, China, November 2019. Association for Computational Linguistics.

[14] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, and Jacob Steinhardt. Measuring massive multitask language understanding. **Proceedings of the International Conference on Learning Representations (ICLR)**, 2021.

[15] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values. **Proceedings of the International Conference on Learning Representations (ICLR)**, 2021.

[16] Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. Meditron-70b: Scaling medical pretraining for large language models, 2023.

[17] Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. Pmc-llama: toward building open-source language models for medicine, 2024.

[18] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. **arXiv preprint arXiv:2404.14219**, 2024.

[19] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. **Applied Sciences**, Vol. 11, No. 14, p. 6421, 2021.

[20] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. **Advances in neural information processing systems**, Vol. 35, pp. 24824–24837, 2022.

[21] Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. **Bioinformatics**, Vol. 39, No. 11, p. btad651, 2023.

[22] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning, 2021.

[23] Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. **Foundations and Trends® in Information Retrieval**, Vol. 3, No. 4, pp. 333–389, 2009.

[24] Xing Han Lù. Bm25s: Orders of magnitude faster lexical search via eager sparse scoring, 2024.

[25] Mahdi Pakdaman Naeini, Gregory F Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In **Proceedings of the AAAI Conference on Artificial Intelligence**, pp. 2901–2907, 2015.

[26] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In **International Conference on Machine Learning**, pp. 1321–1330. PMLR, 2017.

[27] Jeremy Nixon, Michael W. Dusenberry, Linchuan Zhang, and Jerfel. Measuring calibration in deep learning. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops**, June 2019.

[28] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 38–45, Online, October 2020. Association for Computational Linguistics.

A 付録

データセットの選定 MedMCQA のテストセットは公開されていないため, 本研究では MIRAGE [8] で採用された手法に従い, 開発セットをテストセットとして使用した. データセットは, 特に MedQA (https://github.com/jind11/MedQA, 4 択, 4,183 件), MedMCQA (https://github.com/MedMCQA/MedMCQA, 4 択, 1,273 件), PubMedQA (https://github.com/pubmedqa/pubmedqa, 3 択, 500 件), MMLU (https://github.com/hendrycks/test, 4 択, 1,089 件) である.

詳細なパラメータ BM25 に関しては, BM25s [24] を利用し, その設定では, overlap パラメータを 50 に, chunk_size を 1000 に構成して実験を行った. PMC-LLaMA モデルは半精度に量子化, 70B モデルは 4 ビット精度に量子化して実験を実施した. 実装は Transformers [28] を使用した. 全ての実験において, seed 値を 42 に, max_length を 2048 に設定した.

表 4 詳細なモデル名とそのサイズ

モデル	サイズ	HuggingFace の名前
Phi-3.5	3.8B	microsoft/Phi-3.5-mini-instruct
PMC-LLaMA	13B	axiong/PMC_LLaMA_13B
LLaMA2	70B	meta-llama/Llama-2-70b-chat-hf
MEDITRON	70B	epfl-llm/meditron-70b
LLaMA3.1	70B	meta-llama/Llama-3.1-70B
Contriever	1.1B	facebook/contriever
MedCPT	1.1B	ncbi/MedCPT-Query-Encoder

表 5 検索モデルの詳細な設定と情報

モデル	タイプ	サイズ	指標	ドメイン
BM25	Lexical	-	BM25	General
Contriever	Semantic	1.1B	IP	General
MedCPT	Semantic	1.1B	IP	Biomedical

出力確率の算出 下記を用いて確率を算出する.

$$v_i = \log P(x_i \mid \text{prompt}), \quad P(x_i) = \frac{\exp(v_i)}{\sum_{j=1}^{J} \exp(v_j)} \tag{1}$$

ここで, v_i は各選択肢 x_i に対応する対数確率を表し, prompt はプロンプトである. $P(x_i)$ は選択肢 x_i が正解である確率を表し, これは v_i の指数関数をすべての v_j の指数関数の和で正規化することで計算される. J は選択肢の数であり, 3 または 4 である.

期待較正誤差 (ECE) の算出

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\text{acc}(B_m) - \text{conf}(B_m)\right|$$

ここで,M はビンの数を表し,B_m はビン m に含まれるサンプルの集合,$|B_m|$ はビン m 内のサンプル数,n は全サンプル数である. $\text{acc}(B_m)$ はビン B_m 内の正答率を示し,$\text{conf}(B_m)$ は予測された信頼度を表す.

適応較正誤差 (ACE) の算出

$$\text{ACE} = \frac{1}{KR} \sum_{k=1}^{K} \sum_{r=1}^{R} \left|\text{acc}(r,k) - \text{conf}(r,k)\right|$$

K はクラスの数,R はビンの数を表す.$\text{acc}(r,k)$ はクラス k におけるビン r の正答率,$\text{conf}(r,k)$ は同じビンおよびクラスにおける予測の信頼度を示す.

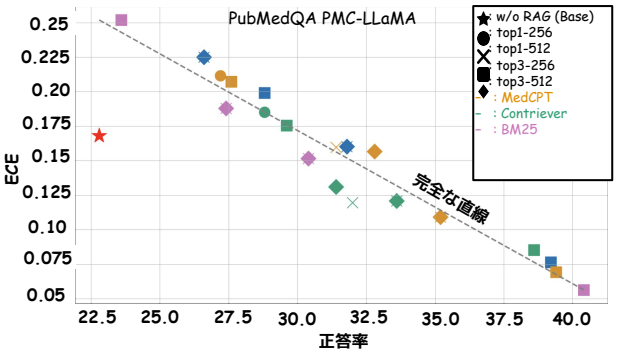

図 4 PubMedQA における PMC-LLaMA の結果を示す。x 軸は精度を表し、y 軸は ECE を示す。$y = -x$ の線は理想的な状態を表す。理想的には、ベースライン（★ で示される）は左上に位置し、RAG を組み込んだ構成は右下に向かう。

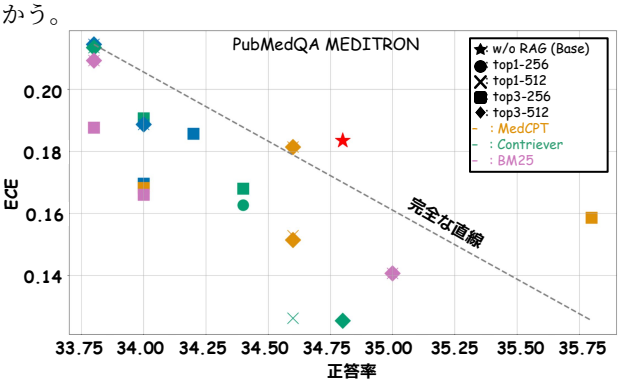

図 5 PubMedQA を使用した MEDITRON の結果.

RAG の詳細な設定の影響 RQ1 の詳細な分析を追加すると, 検索モデル, 推論モデル, および取得する文書数の組み合わせがモデルに影響を与えることを確認した. RAG でより多くの文書を取得することは多様な情報への紹介を提供する一方で, 精度および信頼度の悪化を招く可能性がある. これは, RAG において文書の取得が増加することで, モデルが本来持つ知識をより多く忘却していることを示唆している. 例えば, PMC-LLaMA では精度が改善した一方で ECE が悪化している. また, LLaMA3.1 では ECE の悪化が達成されたものの, 精度も悪化する結果となった. これらは, 最適な設定が具体的な状況や優先される側面によって異なることを示している.