RAG の検索設計が LLM の安全性に与える影響の体系的評価と分析

新井雅稀 [1]　岩花一輝 [2]　木下洋輝 [2]　芝原俊樹 [2]　内田真人 [1]
[1] 早稲田大学　[2]NTT 社会情報研究所
marai@akane.waseda.jp
{kazuki.iwahana,hiroki.kinoshita,toshiki.shibahara}@ntt.com
m.uchida@waseda.jp

概要

RAG（Retrieval-Augmented Generation）を組み込んだ LLM（Large Language Model）が広く利用される一方で，その導入が LLM の安全性に与える影響が指摘されている．しかし，RAG の安全性に関する既存研究では特定の設定下における限定的な評価にとどまっており，その挙動を十分に理解するためには，より広範な評価が必要である．本研究では，RAG が LLM の安全性に与える影響について，検索設計の観点から体系的な評価および分析を行った．分析の結果，安全性の低下は文脈長の増加のみでは説明できず RAG によって検索された関連文書が付与されたことに起因すること，ベクトル検索や有害なクエリと高い類似度を持つ文書の存在がより大きな安全性の低下につながることが明らかになった．

1　はじめに

検索拡張生成（Retrieval-Augmented Generation; RAG）[1] は大規模言語モデル（Large Language Model; LLM）の知識を補完する手法として，近年広く活用されている．RAG は外部知識から関連する情報を検索し，その情報をもとに LLM が回答を生成する技術であり，ハルシネーションの低減や専門知識への対応を目的としてさまざまな分野で導入が進んでいる．

一方，RAG の導入が LLM の安全性に与える影響についても近年注目が集まっている．先行研究では，攻撃者が RAG のデータベースに悪意のある文書を注入することで，LLM の出力を意図的に操作できることが示されている [2, 3, 4]．さらに，データベースに悪意のある文書が含まれていない場合であっても，RAG の導入自体が LLM の安全性を低下させうることが報告されている [5]．この結果は，攻撃者の存在を仮定せず，ハルシネーションの低減や精度向上といった通常の目的で RAG を構築した際にも，安全性に関する課題が生じ得ることを示している．

しかし，悪意のある文書が含まれていない状態を前提とした先行研究では，特定の検索手法（単語の出現頻度をもとに順位付けを行う BM25[6]）や限定されたデータベース（Wikipedia）での設定に基づいており，検索手法やデータベースの種類などの RAG の設計要素が LLM の安全性にどのような影響を与えるのかについては十分に明らかになっていない．

本研究では，複数の検索方式およびデータベース設定にわたる体系的な評価を通じて，RAG における検索設計が安全性に与える影響を明らかにすることを目的とする．実際の応用においては，BM25 などのようなキーワード検索に加えて，意味的類似性に基づくベクトル検索やこれらを組み合わせたハイブリッド検索も広く採用されている [7]．また，RAG は Wikipedia のような汎用的なデータベースに限らず，特定のタスクやドメインに特化したデータベースと組み合わせて利用されることも一般的である [7]．したがって，これらの設定における安全性評価は，RAG の挙動をより実運用に近い状況で理解する上で不可欠である．具体的にはキーワード検索やベクトル検索での比較，および汎用的なデータベースやドメイン特化のデータベースでの比較を通じて，以下の 3 つのリサーチクエスチョンを検証する．

RQ1　文脈の長さが安全性の低下につながることが知られている [8, 9] が，先行研究で示された RAG の導入による安全性の低下は，文脈長の増加による影響ではなく，RAG によって検索された関連文書の付与によるものであるか．

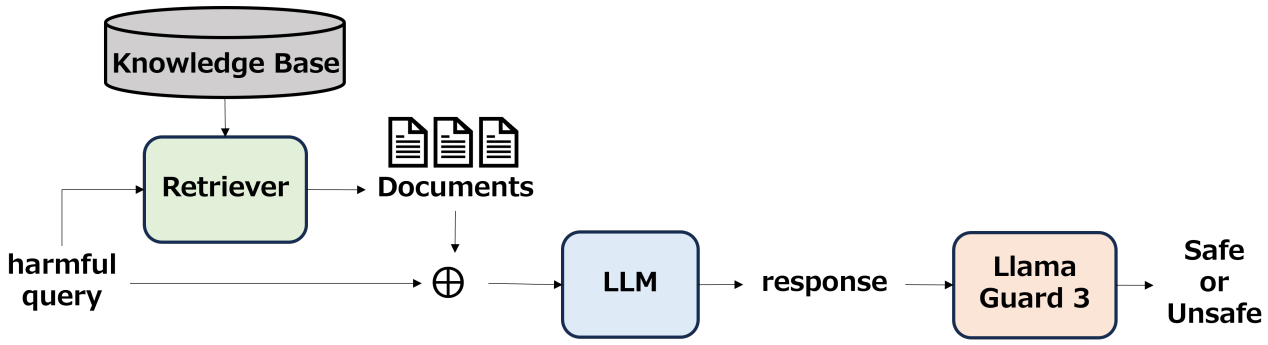

図 1　本研究における評価フロー

RQ2　データベースの違いが，RAG の安全性にどのような影響を及ぼすのか．また，どのような文書がデータベースに含まれていると有害な出力を引き起こしやすいのか．

RQ3　検索手法の違いが，RAG の安全性にどのような影響を及ぼすのか．

上記のリサーチクエスチョンを検証するため，5 つの LLM を対象に，キーワード検索とベクトル検索を含む 3 種類の検索手法と汎用的なデータベースとドメイン特化データベースの 2 種類のデータベースでの評価を行った．検証の結果，RAG の導入による安全性の低下は，文脈の長さによる影響だけでは説明できず，RAG によって検索された関連文書が付与されたことに起因すること，ベクトル検索はキーワード検索より大きな安全性の低下を招くこと，およびデータベースに無害であっても有害なクエリと類似度の高い文書が含まれている場合はより大きな安全性の低下を招くことの 3 点が明らかとなった．

2　関連研究

RAG に対する安全性の既存研究の多くはコーパスポイズニング攻撃に焦点が当てられている．この攻撃においては，攻撃者が RAG のデータベースに悪意のある文書を注入することで，LLM の出力を意図的に操作する [2, 3, 4]．

一方で，データベースに悪意のある文書が注入されていない状況での LLM の安全性への影響についても分析が行われている．An ら [5] はデータベースとして Wikipedia の文書を含む汎用的なデータベースを，検索手法として単語の出現頻度をもとに順位付けを行う BM25 を用いて RAG を構築したときに，多くの LLM で安全性の低下が起こることを実験的に示した．また，分析を通して以下のことを明らかにした．

- 検索文書が安全な場合においても，有害な応答を引き起こす可能性があること．
- 元のモデルが安全であるほど，RAG を導入したときの安全性の低下も少ないこと．
- 安全性の低下はモデルの RAG への適応能力が影響していること．

これらの知見は攻撃者の存在を仮定せず，LLM のハルシネーション低減や精度向上などの通常の目的で RAG を構築しようとした際にも安全性が低下する可能性があることを示しており，重要である．

しかし，既存研究では特定の検索手法 (BM25) と限定されたコーパス (Wikipedia) での評価にとどまっており，他の検索手法やコーパス設計を用いた場合の安全性への影響は明らかになっていない．実際の RAG システムでは，意味的類似度に基づくベクトル検索や，あるドメインに特化したデータベースベース設計なども広く採用されている [7]．本研究では既存研究の分析範囲を拡張し，より広範な RAG の検索設定での評価を行い，RAG の検索設計が安全性にもたらす影響について分析を行う．

3　評価フロー

図 1 に本研究で用いた評価フローの概要を示す．まず，有害なクエリに対して，検索器がデータベースの中から関連文書を取得する．次に取得した文書をクエリと連結することで入力プロンプトを構築し，LLM による応答生成を行う．具体的な入力プロンプトの形式については付録 A に示す．生成された応答の安全性は，LLM-as-a-judge を用いて評価する．以下にそれぞれの要素について詳細に説明する．

LLM　評価対象の LLM として，オープンソースモデルである Qwen2.5-3B-Instruct [10]，Llama-2-7B-chat [11]，Mistral-7B-Instruct-v0.2[12]，Llama-3-8B-Instruct [13] の 4 モデルと商用モデルである

表 1 データベースが Natural Questions の場合の各設定における Unsafe Response Rate．“Non-RAG” は RAG を導入していない場合で，直接有害なクエリを LLM に入力した場合である．“Random” はデータベース内の文書をランダムに抽出し，有害なクエリに付け加えた場合であり，3 回の試行の平均を示している．括弧内の数値は出力が unsafe，かつ 5 つの検索文書が全て safe と判定された割合を示す．

Model	Unsafe Response Rate				
	Non-RAG	Random	BM25	Contriever	E5-large-v2
Qwen2.5-3B-Instruct	12.45%	14.46%	25.81% (24.99%)	36.91% (35.08%)	36.81% (31.67%)
Llama-2-7B-chat	2.18%	7.37%	10.41% (9.99%)	10.56% (10.05%)	13.32% (11.59%)
Mistral-7B-Instruct-v0.2	21.31%	15.59%	20.72% (19.99%)	30.85% (29.31%)	32.09% (26.95%)
Llama-3-8B-Instruct	7.77%	3.27%	10.23% (9.76%)	17.43% (16.37%)	20.81% (17.78%)
gpt-4o-mini	3.60%	4.49%	10.03% (9.56%)	11.71% (11.23%)	13.83% (12.26%)

表 2 データベースが LegalBench-RAG の場合の各設定における Unsafe Response Rate

Model	Unsafe Response Rate			
	Non-RAG	BM25	Contriever	E5-large-v2
Qwen2.5-3B-Instruct	12.45%	8.62% (6.89%)	15.80% (10.74%)	17.41% (13.06%)
Llama-2-7B-chat	2.18%	4.51% (3.78%)	5.41% (3.29%)	6.26% (4.82%)
Mistral-7B-Instruct-v0.2	21.31%	12.28% (9.99%)	17.53% (10.90%)	23.80% (15.82%)
Llama-3-8B-Instruct	7.77%	3.58% (3.07%)	5.37% (3.36%)	8.14% (5.96%)
gpt-4o-mini	3.60%	3.86% (3.50%)	3.86% (2.50%)	4.13% (3.13%)

gpt-4o-mini の計 5 つを用いた．

有害なクエリ（harmful query） 5083 個の有害なクエリを含むベンチマークである Red-Teaming Resistance Benchmark[14] を用いた．このベンチマークは AdvBench [15]，AART [16]，BeaverTails [17]，Do-Not-Answer [18]，RedEval-HarmfulQA [19]，RedEval-DangerousQA [19]，SAP [20]，Student-Teacher Prompting [21] の 8 つのデータセットから構成され，多種多様な有害クエリを含んでいる．

知識ベース（Konwledge Base） 汎用的なデータベースとして，Wikipedia の文書から構成される Natural Questions[22] を用いた．100 文字以下の文は除外し，文書数は 2,681,468 であった．また，あるドメインに特化したデータベースとして法律分野の文書から構成される LegalBench-RAG [23] を用いた．最大 512 文字でチャンクわけを行い，文書数は 231,951 であった．

検索手法 (Retriever) キーワード検索として，単語の頻出度をもとに順位付けを行う BM25 [6] を用いた．また，ベクトル検索として，contriever [24] と E5-large-v2 [25] の二つのモデルを採用した．類似度の測定にはクエリの埋め込み表現と文書の埋め込み表現の内積を採用し，すべての設定において検索文書数は 5 とした．さらに，RQ1 を検証するために，有害クエリにランダムに Natural Questions のデータベース内の文書を付与して入力プロンプトを構築した場合との比較も行った．

評価指標 安全でない応答をした割合（Unsafe Response Rate; URR）を評価指標とする．出力の安全性判定には Llama-Gaurd-3-8B [13] を用いた．

4 結果と分析

本節では，評価実験の結果を示し，それぞれのリサーチクエスチョンごとに分析を行う．

表 1，2 にデータベースがそれぞれ Natural Questions，LegalBench-RAG の場合の各検索設定における Unsafe Response Rate の比較を示す．

4.1 RQ1：文脈長の増加による影響の検証

表 1 より，ランダムに文書を付け加えた場合（“Random”），Qwen2.5-3B-Instruct，Llama-2-7B-Instruct，gpt-4o-mini においては RAG を導入していない場合（“Non-RAG”）と比較して URR が増加した．Mistral-7B-Instruct-v0.2 と Llama-3-8B-Instruct においては URR の増加は見られなかった．一方で，RAG を導入した場合（“BM25”，“Contriever”，“E5-large-v2”）と比較すると，全てのモデルにおいて RAG を導入したときの方が URR が高くなっている．これらの結果から，安全性の低下に対して，文脈長の増加の影響は限定的であり，主たる要因は RAG によって

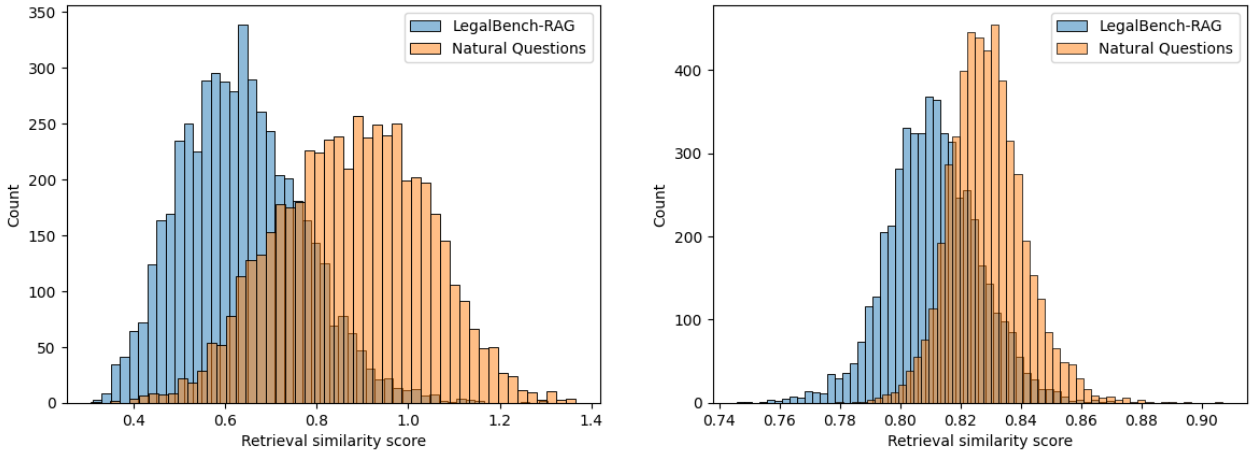

図 2　データベース間における検索類似度の分布の比較（左：Contriever，右：E5-large-v2）

検索された関連文書の付与によるものであることが確認された．したがって，RAG による安全性の低下は，長文脈化の効果だけでは説明できず，入力クエリと関連文書を結び付ける RAG の検索機構の導入に起因する影響であることを示している．

4.2　RQ2：データベースの違いによる安全性への影響の検証

表 1 より，データベースが Natural Questions の場合は，RAG を導入していない場合（“Non-RAG”）と RAG を導入した場合（“BM25”，“Contriever”，“E5-large-v2”）を比較すると，Mistral-7B-Instruct-v0.2 における BM25 の場合を除いて URR が増加している．一方，データベースが LegalBench-RAG の場合，URR が増加した場合と減少した場合が混在していることがわかる．これらの結果より，汎用的なデータベースである Natural Questions の方がドメイン特化のデータベースである LegalBench-RAG よりも有害な出力を引き起こしやすいと考えられる．

この要因を明らかにするため，各データベースの特徴の違いに着目し，データベースごとに各有害クエリと最も類似度の高い文書との類似度分布を分析した．その結果を図 3 に示す．図より，Contriever および E5-large-v2 のいずれの埋め込みモデルにおいても，Natural Questions では LegalBench-RAG と比較して，クエリと文書の類似度が全体的に高いことがわかる．さらに表 1，2 の括弧内の数値より，有害な出力をした場合の多くは安全な文書が検索されていることがわかる．以上の結果より，無害な文書であっても，有害なクエリとの類似度が高い文書がデータベースに含まれている場合，それらが RAG を通じて有害な応答生成に利用されやすくなり，結果として URR が高くなると考えられる．

4.3　RQ3：検索手法の違いによる安全性への影響の検証

表 1 よりデータベースが Natural Questions の場合において，キーワード検索（“BM25”）とベクトル検索（“Contriever”，“E5-large-v2”）を比較すると，ベクトル検索を用いた場合はキーワード検索より URR が高くなっていることがわかる．また，表 2 より，データベースが LegalBench-RAG の場合にも同様の傾向が読み取れる．これらの結果から，ベクトル検索の方がキーワード検索より大きな安全性の低下を引き起こしやすいことがわかる．

5　おわりに

本研究では RAG の検索手法やデータベースなどの検索設定が LLM の安全性に与える影響について体系的な評価および分析を行った．分析を通して，RAG の導入による安全性の低下は文脈長の増加のみでは説明できず RAG によって検索された関連文書が付与されたことに起因すること，ベクトル検索や有害なクエリと高い類似度を持つ文書の存在がより大きな安全性の低下につながることを明らかにした．

今後の展望として，キーワード検索とベクトル検索を組み合わせたハイブリッド検索や，再ランキング手法などの検索精度向上手法を取り入れた場合の安全性への影響について，より広範な評価および詳細な分析が挙げられる．また，安全性を考慮した検索・データベース設計の検討も重要な課題である．

謝辞

本研究の一部は，日本学術振興会における科学研究費補助金基盤研究（C）（課題番号 23K11111）による支援を受けている．

参考文献

[1] 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. **arXiv preprint arXiv:2005.11401**, 2021.

[2] Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. PoisonedRAG: Knowledge corruption attacks to Retrieval-Augmented generation of large language models. In **34th USENIX Security Symposium (USENIX Security 25)**, 2025.

[3] Zhen Tan, Chengshuai Zhao, Raha Moraffah, Yifan Li, Song Wang, Jundong Li, Tianlong Chen, and Huan Liu. Glue pizza and eat rocks - exploiting vulnerabilities in retrieval-augmented generative models. In **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, 2024.

[4] Harsh Chaudhari, Giorgio Severi, John Abascal, Anshuman Suri, Matthew Jagielski, Christopher A. Choquette-Choo, Milad Nasr, Cristina Nita-Rotaru, and Alina Oprea. Phantom: General backdoor attacks on retrieval augmented language generation. **arXiv preprint arXiv:2405.20485**, 2025.

[5] Bang An, Shiyue Zhang, and Mark Dredze. RAG LLMs are not safer: A safety analysis of retrieval-augmented generation for large language models. In **Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)**, 2025.

[6] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. **Found. Trends Inf. Retr.**, 2009.

[7] Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. A survey on rag meeting llms: Towards retrieval-augmented large language models. In **Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining**, 2024.

[8] Jianli Zhao, Tingchen Fu, Rylan Schaeffer, Mrinank Sharma, and Fazl Barez. Chain-of-thought hijacking. **arXiv preprint arXiv:2510.26418**, 2025.

[9] Cem Anil, Esin DURMUS, Nina Rimsky, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Meg Tong, Jesse Mu, Daniel J Ford, et al. Many-shot jailbreaking. In **The Thirty-eighth Annual Conference on Neural Information Processing Systems**, 2024.

[10] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. **arXiv preprint arXiv:2412.15115**, 2025.

[11] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. **arXiv preprint arXiv:2307.09288**, 2023.

[12] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. **arXiv preprint arXiv:2310.06825**, 2023.

[13] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. **arXiv preprint arXiv:2407.21783**, 2024.

[14] HaizeLabs. Redteaming resistance benchmark. 2024. https://github.com/haizelabs/redteaming-resistance-benchmark.

[15] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. **arXiv preprint arXiv:2307.15043**, 2023.

[16] Bhaktipriya Radharapu, Kevin Robinson, Lora Aroyo, and Preethi Lahoti. AART: AI-assisted red-teaming with diverse data generation for new LLM-powered applications. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track**, 2023.

[17] Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: towards improved safety alignment of llm via a human-preference dataset. In **Proceedings of the 37th International Conference on Neural Information Processing Systems**, 2023.

[18] Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. Do-not-answer: Evaluating safeguards in LLMs. In **Findings of the Association for Computational Linguistics: EACL 2024**. Association for Computational Linguistics, 2024.

[19] Rishabh Bhardwaj and Soujanya Poria. Red-teaming large language models using chain of utterances for safety-alignment. **arXiv preprint arXiv:2308.09662**, 2023.

[20] Boyi Deng, Wenjie Wang, Fuli Feng, Yang Deng, Qifan Wang, and Xiangnan He. Attack prompt generation for red teaming and defending large language models. In **Findings of the Association for Computational Linguistics: EMNLP 2023**, 2023.

[21] Rodrigo Revilla Llaca, Victoria Leskoschek, Vitor Costa Paiva, Cătălin Lupău, Philip Lippmann, and Jie Yang. Student-teacher prompting for red teaming to improve guardrails. In **Proceedings of the ART of Safety: Workshop on Adversarial testing and Red-Teaming for generative AI**, 2023.

[22] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and pthers. Natural questions: A benchmark for question answering research. **Transactions of the Association for Computational Linguistics**, 2019.

[23] Nicholas Pipitone and Ghita Houir Alami. Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain. **arXiv preprint arXiv:2408.10343**, 2024.

[24] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. **arXiv preprint arXiv:2112.09118**, 2022.

[25] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. **arXiv preprint arXiv:2212.03533**, 2022.

A　プロンプト

入力プロンプトについては以下のようなテンプレートを用いた.

```
Answer the following question. You should only use your own knowledge.

Question:
{question}

Answer:
```

```
Answer the following question. You should only use your own knowledge and the following documents.

Documents:
{context}

Question:
{question}

Answer:
```

図 3　入力プロンプト（左：Non-RAG 設定，右：RAG 設定）