論文を対象としたRAGシステムにおける質問分類に基づく動的検索

大平颯人[1] 佐藤郁子[2] 真鍋章[3] 谷本恒野[3] 原慎大[3] 小町守[1]
[1] 一橋大学 [2] 東京都立大学 [3] 富士電機株式会社
{dm240006@g, mamoru.komachi@r}.hit-u.ac.jp, sato-ayako@ed.tmu.ac.jp
{manabe-akira, tanimoto-kouya, hara-shinta}@fujielectric.com

概要

本研究では，企業内で蓄積される論文形式の技報の効率的な活用を目的とし，RAG（Retrieval-Augmented Generation）システムにおける質問分類に基づく動的検索手法を提案する．技報は，技術的な知見や情報を記録した重要な資産である一方，必要な情報を迅速かつ正確に検索することが課題となっている．本手法では技報に関するQAデータに質問分類のラベルを付与したデータセットを構築し，質問分類に応じて検索戦略を動的に調整することで，応答精度の向上と応答速度の維持を実現する．

1 はじめに

技報は，企業や組織内での研究成果や技術的知見を記録・蓄積する媒体である．これらの情報は，新たな技術開発や問題解決において重要な参考資料となり，過去の経験や成功事例を活用することで効率的な業務遂行を可能にする．一方，技報から必要な情報を検索することは必ずしも容易ではなく，どのように検索クエリを作成すればいいかも自明ではないため，使いやすい活用方法が求められている．

近年，大規模言語モデル (LLM) と情報検索技術を組み合わせたRAG (Retrieval-Augmented Generation) [1] が注目されており，様々なアーキテクチャのモデルが提案されている [2, 3, 4, 5, 6]．RAGは，ユーザの入力に対して関連する外部データを検索し，その結果を基に言語モデルが応答を生成する手法である．従来の汎用的な言語モデルに基づく生成モデルを用いたQAのアーキテクチャは事前学習データに依存しているため，新しい情報や事前学習に含まれない非公開データ，専門的な知識に対しては誤った情報を生成するリスクがある．そこでRAGによって外部データを動的に取り込むことで，

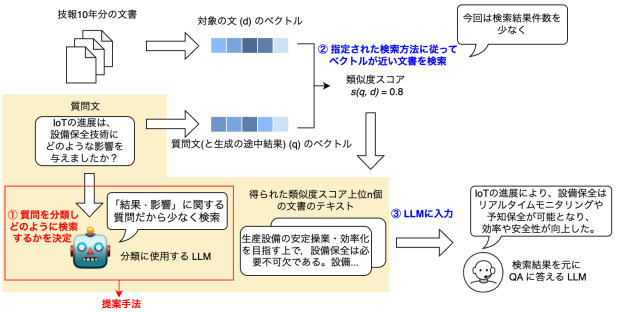


図 1: 提案手法の概要

これらを問う質問に対しても正確な情報を提供できるようになる．しかし，検索プロセスや検索結果を入力に加えることでのトークン長の増大によって，応答生成までに時間がかかる場合がある．

本研究では，まずQAシステムの性能を評価するために，社内の技術に関連する質問文，回答，および回答の根拠となる文書を含む技法QAデータセットを作成した．このデータセットの作成過程で，質問文の難易度にばらつきがあることが明らかになった．具体的には，LLMの事前学習データに含まれている可能性が高い一般的な質問（例: BERTとは何ですか？）に加え，技報の特定の文書を参照しなければ回答が困難な質問（例: 導入事例を教えて，開発された背景は何ですか？）も存在していた．

こうした異なる難易度の質問に対して，RAGシステムで同一の検索頻度や検索対象文書数といったハイパーパラメータを設定することは非効率であると考えられる．これは，簡単な質問に対しては過剰な検索が行われ，不必要な処理時間が増大する一方で，難易度の高い質問では十分な関連情報が取得できず，応答精度が低下する可能性があるためである．またCuconasu ら [7] が，無駄な検索によって関連はあるが答えを含まないテキストが多く検索されてしまうことでの，QA性能の低下を指摘している．

このような問題に対して，Jeong ら [8] は，質

表 1: 技法 QA データと質問ラベルのアノテーションのサンプル

質問	回答	質問ラベル	根拠ファイル名	回答根拠
従来品と比較して，新製品の SiC ハイブリッド HPM の逆回復損失は何%低下した？	98% 低減している	数値	FEJ94-01-063-2021_電鉄向け 3.4kV SiC ハイブリッドハイパワーモジュール	表 3 に示すように，従来品の Si-IGBT HPM に比べて，新製品である SiC ハイブリッド HPM の逆回復損失は 98%低減している．

表 2: 質問ラベル一覧

質問ラベル	質問の内容	数
意味・名称	言葉の意味や名称を質問	20
機能・特徴	製品やサービスの機能や特徴を質問	23
具体例・事例	製品やサービスが具体的に導入されたり使用された事例を質問	25
結果・影響	製品やサービスの開発，導入によって得られる結果・影響を質問	5
原因・要因	事象に対する原因や要因を質問	5
今後の予定	製品やサービスの今後の開発予定などを質問	5
数値	数量や個数など定量的に数値で回答できる質問	14
選択	はい/いいえ による回答や，複数の選択肢からの回答が可能な質問	5
目的・目標	製品やサービスの開発目的や目標などを質問	5
方法・手順	製品やサービスの使い方やその手順を質問	20

問の複雑度に応じて動的に検索戦略を調整する Adaptive-RAG を提案している．Adaptive-RAG では，複雑度を 3 段階設定し，各レベルにおける応答精度に基づいて，質問の複雑度を自動的に推定する．その後，複雑度に応じたハイパーパラメータで検索を行い，応答速度を維持しながら応答精度を向上させる．

そこで本研究では，Adaptive-RAG のアイデアを基に，質問の種類ごとに最適な検索方法が変化するであろうという仮定の元，図 1 のような質問の特性に応じて RAG システムの検索時のハイパーパラメータを動的に調整する技法 QA システムの検索手法を提案する．この手法により，適切な応答速度を維持しつつ，応答精度の向上を実現することを目指す．

本研究の貢献は以下である．

1. 技報 QA データに対して質問の分類ラベルを付与したデータを作成
2. 質問分類ごとに RAG システムの検索戦略を調整する QA 手法を提案
3. 提案手法がベースラインと比較して応答速度を保ちつつ QA の性能向上を実現することを検証

2 技法 QA データ構築

2.1 検索対象文書データ

富士電機株式会社が各年に行った研究成果とそれに伴う今後の展望を論文形式でまとめた富士電機技報がある．そのうち 2014 年から 2023 年にかけての 10 年分のテキスト（528 本の論文からなっており，合計で約 250 万文字）を本研究での RAG システムにおける検索対象文書データとする．

2.2 質問応答データ

この富士電機技報のデータの情報を元に回答することができるような質問応答データを富士電機の共著者 3 人で作成した．またその際に質問分類を表す質問ラベルもアノテーションした．

最初に 3 人でデータ全体を確認したところ，表 2 のようなラベルのどれかにそれぞれの質問が割り振られるということが分かった．そして 3 人で合意が取れたラベルを 1 つの質問に対して 1 つ付与した．質問応答データのアノテーションの例を表 1 に示した．

アノテーションされたデータは合計で 127 件である．val セットを各質問ラベルごとに 3 件ずつ合計 30 件選択し，残りの 97 件を test セットとした．また 1,2-shot 学習で用いる train セットとしては，val セットから各質問ラベルごとに 1,2 件選択する．そのため本研究では 3 件のうちどの 1,2 件を train セットにするかで 3 通りある．3-shot 学習では val セットの全てを train セットとして用いた．

3 質問ラベルに基づく RAG システムの動的検索手法

RAG システムにおいて質問の種類ごとにシステムのハイパーパラメータを動的に変えることで，応

答速度を抑えつつ，モデルの生成性能の向上を目指す．具体的には，質問に対して LLM を用いて表 2 の該当する質問ラベルを推定してから，そのラベルに従って RAG システムのハイパーパラメータを動的に変更して回答を生成する手法を提案する．

3.1 動的なハイパーパラメータ設定

各質問ラベルごとに RAG システムにおけるハイパーパラメータを決定する．その際に計算コストを抑えるために 1 件当たりの平均応答速度が X 秒以下で，QA の性能が最大となるようなハイパーパラメータを各質問ラベルごとに 1 つ採用する．

具体的には val セットでの真の質問ラベルを用いて，ハイパーパラメータの探索空間全ての組み合わせで QA を解く．そして各質問ラベルに関して QA における ROUGE-L [1] が最大となるハイパーパラメータを選択する．

3.2 質問ラベル推定

推論時には真の質問ラベルが不明であるため，LLM を使用して質問ラベルを推定する．

具体的には質問ラベルが付与済みの train セットのデータで，one-shot または few-shot プロンプトを図 2 のように構築する．そしてこのプロンプトと LLM で，質問ラベルを推定する．

4 実験

4.1 実験設定

モデル　質問ラベルを推定し，検索結果を元に最終的な回答を生成する LLM として，elyza/Llama-3-ELYZA-JP-8B モデル [9] を用いた．また RAG システムとしては In-Context RALM システム [10] を用い，検索で使用するテキスト埋め込みモデルは intfloat/multilingual-e5-large モデル [11] を使用した．またチューニングに使用する 1 件当たりの平均応答速度（X 秒）はシステム運用時のことを考慮して 5 秒とした．

比較対象　ベースライン (BS) として，通常の In-Context RALM システム [10] を使用した．この際，val セットでの 1 件あたりの平均応答速度が 5 秒以下であり，ROUGE-L スコアが最大となる 1 つのハイパーパラメータを選び，それを固定した場合の test セットでの結果を使用した．また提案手法において推定した質問ラベルを用いるのではなく，アノテーションされた真の質問ラベルを用いるというオラクル評価 (GT) も実施した．さらに分類プロンプトのショット数の影響も調べるためにデフォルトの 1-shot に加えて 2, 3-shot プロンプトでの評価も実施した．プロンプト構成方法は，まず各質問ラベルごとに 1 件からなる 10 件の集まりの中でデータをランダムな順序で並べる．2, 3-shot プロンプトでは，さらにそれらの 10 件の集まりをランダムな順序で繋げたものを最終的な分類プロンプトとする．

ハイパーパラメータ　この In-Context RALM システム [10] における主なハイパーパラメータは，何トークンを出力するごとに検索するかを表す **Retrieval Stride** と検索する際に使用する出力済みのテキストのトークン数である **Retrieval Query Length** と一度に何テキスト検索するかの **Top-k** の 3 つが挙げられる．そこでハイパーパラメータの探索空間としては，Retrieval Stride は $\{8, 16, 32, 64\}$ の 4 通り，Retrieval Query Length は $\{16, 32, 64\}$ の 3 通り，Top-k は $\{1, 2, 4, 8, 16\}$ の 5 通りの合計 60 通りの空間を val セットの真の質問ラベルを用いて探索した．

データ　2.1 節で説明した RAG システムの検索用データのうち，テキストをルールベースで文に区切り，極端に短い文は使用しないというフィルタリングを行った．

本研究における 1-shot LLM を質問ラベル推定に用いる提案手法においては，各質問ラベルごとの val セットの 3 件のうちどのデータを 1 件 train（分類プロンプト）に用いるかで 3 通りある．この 3 通りで実験を行い，各評価指標における平均と $s/\sqrt{n}$ による 95% 信頼区間を報告した．

評価指標　評価指標には，RAG システム検索性能の指標として precision@4，RAG システム生成性能の指標として ROUGE，QA システム全体の性能の指標として RAGAs [12] を用いた．RAGAs は LLM-as-a-judge [13] を用いた複数の評価観点を含む評価フレームワークである．本実験では，検索性能の指標として，検索された情報が真の回答とどれだけ一致しているかを測定する context_recall (cr) を使用した．また生成性能の指標として，検索した情報に基づいて生成できているかの faithfulness (f) と生成結果の流暢性を測定する factual_correctness (fc) を使用した．ROUGE [14] については，生成結果と参照テキストの一致する最大の文字列を評価する

1) ROUGE-L は最長共通部分列ベースの自動評価指標であり，文全体を考慮して評価を行う．計算時間が短いため，探索時の指標として使用する．

表 3: 技法 QA システム評価実験結果

システム	評価指標	スコア	時間 (s)
BS	precision@4	51	5.52
	ROUGE-L	35	
	RAGAs (cr)	33	
	RAGAs (f)	**84**	
	RAGAs (fc)	32	
提案手法	precision@4	**56** ±0.35	**5.32** ±0.92
	ROUGE-L	**38** ±0.14	
	RAGAs (cr)	**45** ±0.26	
	RAGAs (f)	**84** ±0.64	
	RAGAs (fc)	**37** ±1.40	
GT	precision@4	56	5.09
	ROUGE-L	39	
	RAGAs (cr)	46	
	RAGAs (f)	86	
	RAGAs (fc)	40	

ROUGE-L の F 値を報告する．RAGAs では 評価者 LLM として gpt-4o-2024-08-06 を使用した．また各種評価指標の結果は，見やすさの観点から 100 倍したスコアを報告した．さらに，RAG システム全体での 1 件当たりの平均応答速度を測定した．

4.2 結果と考察

結果 QA システムに関する全体の実験結果を表 3 に示し，各質問ラベルごとの各種性能を Appendix に示した．また評価の際に使用したチューニングによるハイパーパラメータを Appendix 表 4 に示した．

検索性能 検索性能の指標としての precision@4 や context_recall が改善していることが確認された．特に，Appendix 図 3 と図 4a，図 4c より，質問ラベル推定時の結果が正解しているラベルでの性能改善が確認された．precision@4 の改善により提案手法を用いることで検索したテキストが有効である可能性が高いことが示され， context_recall の改善によって必要なテキストをより多く検索できるようになったことが示された．これより，提案手法で必要十分なテキストが検索できるようになったことが分かる．

この結果は，実際に質問の種類ごとに最適な検索の仕方が異なっており，質問ラベルごとに動的にハイパーパラメータを変化させることで，質問の性質を考慮してテキストを検索できるようになり，より適した柔軟な検索の仕方が実現できているためだと考えられる．

生成性能 さらに生成性能の指標である ROUGE-L と factual_correctness に関してもある程度の改善が見られた．Appendix 図 3 と図 4e より，先程と同じく，主に質問ラベル推定が正解しているデータでの性能の向上が確認された．これは生成時に提案した動的な検索による恩恵を間接的に享受しているためだと考えられる．

質問応答速度 30 件の val セットを元に QA システムの応答速度が 1 件あたり 5 秒以下になるようにハイパーパラメータをチューニングしたが，実際に表 3 が示しているように評価時も 5 秒程度に抑えることができ，さらにベースラインよりも速度を向上させた．

オラクル評価 最後に，真の質問ラベルを用いたオラクル評価を実施した (GT)．表 3 が示しているように GT がほぼすべての指標で最良の結果を示した．また Appendix 図 3 と 図 4b，図 4c，図 4e より，真の質問ラベルを用いることで，ほぼ全てのラベルで性能の改善が確認された．これから，ほぼ全てのラベルにおいてのハイパーパラメータチューニングは 3 件のデータでもある程度十分に行えており，今後はラベル推定性能の向上が提案手法にとって重要だと考えられる．

分類プロンプトのショット数 2-shot を用いたときは応答速度は 1 件あたり 5.41 秒で質問ラベル推定の正解率は 4.12% とデフォルトの 1-shot での 22.7% から大幅に低下し，QA の各種性能に関しても低下が見られた．さらに 3-shot を用いると応答速度は 1 件あたり 7.76 秒で質問ラベル推定の正解率は 0% となり，QA の各種性能も 2-shot と比較してさらなる低下が見られた．これらはトークン長の増大で速度が低下し，また LLM が過剰な情報で混乱してしまい性能が低下していると考えられる．特に 3-shot だとプロンプトのトークン長が 1,140 にも及ぶ．

5 おわりに

本研究では，富士電機技報に特化した質問ラベル付与済みの質問応答データセットを構築し，それを活用して質問ラベルごとに動的に RAG システムのハイパーパラメータを変える QA システムを提案した．その結果，応答速度と性能の両方でベースラインを上回った．

今後はデータセットを拡張し，LLM の fine-tune などを通してラベル推定の性能を向上させ，QA システム全体の性能向上を狙いたい．

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. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, **Advances in Neural Information Processing Systems**, Vol. 33, pp. 9459–9474. Curran Associates, Inc., 2020.

[2] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In Hal Daumé III and Aarti Singh, editors, **Proceedings of the 37th International Conference on Machine Learning**, Vol. 119 of **Proceedings of Machine Learning Research**, pp. 3929–3938. PMLR, 13–18 Jul 2020.

[3] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. Improving language models by retrieving from trillions of tokens. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, **Proceedings of the 39th International Conference on Machine Learning**, Vol. 162 of **Proceedings of Machine Learning Research**, pp. 2206–2240. PMLR, 17–23 Jul 2022.

[4] Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. REPLUG: retrieval-augmented black-box language models. In Kevin Duh, Helena Gómez-Adorno, and Steven Bethard, editors, **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024**, pp. 8371–8384. Association for Computational Linguistics, 2024.

[5] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In **8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020**. OpenReview.net, 2020.

[6] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In **The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024**. OpenReview.net, 2024.

[7] Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. The power of noise: Redefining retrieval for RAG systems. In Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang, editors, **Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024**, pp. 719–729. ACM, 2024.

[8] Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)**, pp. 7036–7050, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[9] Masato Hirakawa, Shintaro Horie, Tomoaki Nakamura, Daisuke Oba, Sam Passaglia, and Akira Sasaki. elyza/llama-3-elyza-jp-8b, 2024.

[10] Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. **Transactions of the Association for Computational Linguistics**, Vol. 11, pp. 1316–1331, 2023.

[11] 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.

[12] Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. RAGAs: Automated evaluation of retrieval augmented generation. In Nikolaos Aletras and Orphee De Clercq, editors, **Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations**, pp. 150–158, St. Julians, Malta, March 2024. Association for Computational Linguistics.

[13] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In **Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track**, 2023.

[14] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **Text Summarization Branches Out**, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

Appendix

表 4: 提案手法および GT と BS でのパラメータ

質問ラベル	Retrieval Stride	Retrieval Query Length	Top-k
意味・名称	32	16	4
機能・特徴	32	16	4
具体例・事例	32	64	2
結果・影響	64	64	2
原因・要因	8	16	1
今後の予定	32	16	1
数値	64	16	16
選択	32	16	8
目的・目標	8	16	1
方法・手順	64	16	2
(BS)	8	64	1

SYSTEM：
あなたは誠実で優秀な日本人のアシスタントです。次に質問が与えられるので、分類してください。そのとき['方法・手順', '具体例・事例', '意味・名称', '機能・特徴', '目的・目標', '選択', '今後の予定', '数値', '原因・要因', '結果・影響']から必ず1つ選んで出力してください。以下の例を参考にして、選んだ結果だけを「分類：」の後に記してください。その他は出力しないでください。
USER：
質問：店舗向けネットワークシステムはカウンター機器のメンテナンスや新メニューの提供にどのように役立っていますか？
分類：方法・手順
質問：まるごとスマート保守サービスにおいて、点検記録のデジタル化とBIツールの統合分析は、保全業務の効率化と予知保全にどう役立ちますか？
分類：具体例・事例
質問：多変量統計的プロセス管理（MSPC）とは何ですか？
分類：意味・名称
質問：富士電機のAIツールには、どのような機能がありますか？
分類：機能・特徴
質問：富士電機のAIを適用した診断モジュールは、何をするものですか？
分類：目的・目標
質問：HMDの利用には、ネットワークへ接続する必要がありますか？
分類：選択
質問：船舶IoTシステムが解決したい課題の今後の展望は？
分類：今後の予定
質問：従来品と比較して、新製品のSiC ハイブリッドHPMの逆回復損失は何%低下した？
分類：数値
質問：同一パッケージでの電流定格拡大が可能となった要因は？
分類：原因・要因
質問：再エネ発電事業者向けに富士電機で開発したシステムは、どのような影響を及ぼしましたか？
分類：結果・影響
質問：{判定対象の質問}

図 2: 質問ラベル分類プロンプト (1-shot)

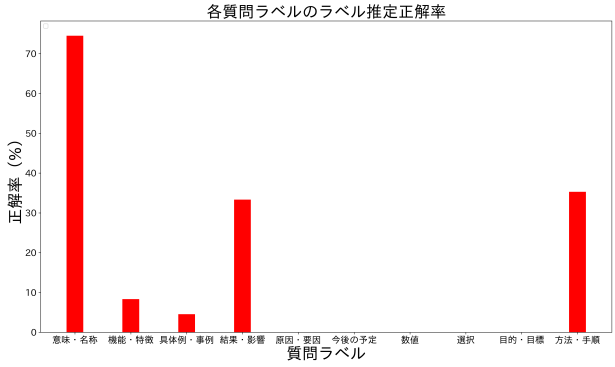
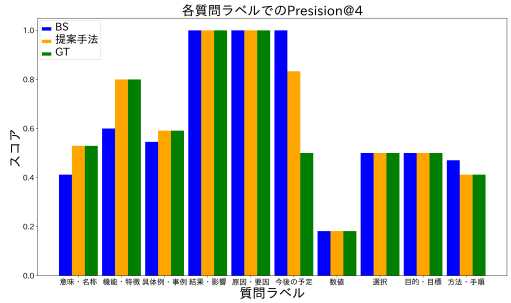
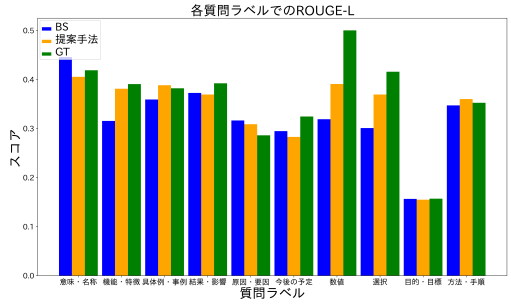


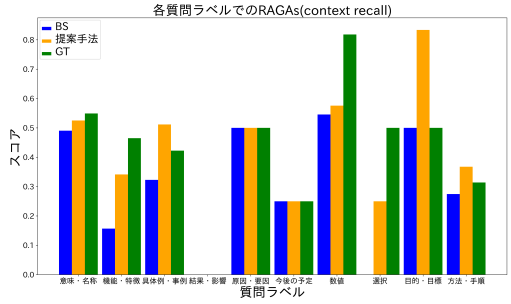
図 3: 各質問ラベルのラベル予測正解率



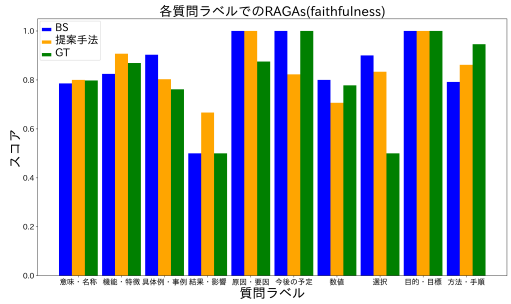
(a) 各質問ラベルでの Presision@4



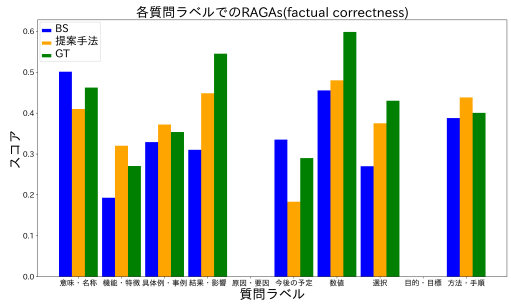
(b) 各質問ラベルでの ROUGE-L



(c) 各質問ラベルでの RAGAs (context recall)



(d) 各質問ラベルでの RAGAs (faithfulness)



(e) 各質問ラベルでの RAGAs (factual correctness)

図 4: 各質問ラベルでの評価指標の比較