

# RAG の応答正確性と連続応答性能の自動評価

岩間 太 竹内 幹雄

日本アイ・ビー・エム株式会社 東京基礎研究所

{gamma, mtake}@jp.ibm.com

## 概要

RAG に基づく質問応答システムによる応答の正確性を自動評価する方法を提案する。特徴は、知識ベース中のコンテキストを、質問、正答例と同時に考慮して、生成回答を LLM で評価する点である。特にコンテキストの選び方に工夫を施す。また、マルチターンの質問応答における連続応答の性能を測るための評価観点を導入し、この観点での評価を自動化する。その後、人手評価との相関、および、連続応答性能の差の検出に基づいた評価実験を実施し、これらの提案手法の有効性、実効性を確認する。

## 1 はじめに

RAG (Retrieval-Augmented Generation, 検索拡張生成) は、事前に準備された独自の知識ベースに対して質問を問い合わせ、関連知識を含む文書等を検索し、その検索結果 (コンテキスト) を質問と同時に LLM に入力して回答を生成する技術である。RAG を用いることで特定領域の専門的で固有の外部知識を利用した自動的な質問応答やチャット機能が実現される。例えば、製品マニュアルやシステムの運用マニュアル、もしくは、組織内の特定業務 (例えば、承認、法務、購買、SDGs など) に関連した文書に基づいた質問応答チャットシステムが、数多く開発され始めてきている。同時に、RAG に基づく質問応答システムの様々な観点からの実践的なテスト/評価技術の必要性も高まっている。RAG の出力には、生成回答だけでなく、文章リストなどの検索結果も含まれる。RAG の評価は、これらの検索結果や検索モジュールの構成を考慮することで、LLM 単体の評価よりも、多様なものとなり得る。

本論文では、第一に、RAG に基づく質問応答システムの応答正確性の評価について論じ、LLM を用いた自動化の方法を提案する。ここで、応答正確性とは、「回答の表現の仕方や粒度の正しさは一意には定まらないが、回答の正確さは (知識ベースの情報

に基づく限り) 明確に判断できる類の質問に対する応答の正確さ」を指す。これは、例えば、DB (データベース) の知識をベースにした際、「SQL において UNION と UNION ALL の違いは?」「デッドロックとは何か?」といった質問を対象にし、「新しい詩とは?」といった質問を扱わないことを意味する。このような意味での応答正確性は、RAG において評価されるべき基本的な能力と考えられる。

応答正確性の自動評価は、例えば、質問ごとに正答例を準備し、生成回答と比較することで行われる。この比較には、ワード単位の recall [1] や BERTScore [2]、または Prometheus [3, 4] 等の評価用の LLM などを用いることができる。また、RAG の自動的な評価フレームワークとして RAGAs [5] が提案されており、生成回答が検索されたコンテキストにどれだけ準拠しているかなど、様々な観点からの評価を LLM を使って自動化している。しかしながら、上述した方法や RAGAs には、コンテキストの情報を考慮して、生成回答が正答例と比較して質問への回答としてどれだけ正確なのかを直接評価する観点は含まれていない。本論文では、この観点を LLM を用いて直接的に評価する方法を提案する。RAG の知識ベースにあるコンテキストは膨大であり、その全てを評価用 LLM のプロンプトに含めることはできない。そこで我々は、特に人手評価との相関を測ることで、どのようなコンテキストを利用した方が良いのかを議論する。

第二に、本論文では、質問応答における連続応答の性能を評価する方法を提案する。近年の RAG システムでは、過去の質問応答の履歴を考慮して検索と回答生成を行うことで、マルチターンの質問応答を可能にしている。我々は、上記の意味での応答正確性において、この連続応答の性能がどのように評価されるべきかを考えたい。マルチターンの対話を評価する方法としては、回答に依存しない一連の質問を標準的な評価用データセットとして準備する方法 [6] や、一連の質問と応答を生成しその自然さを

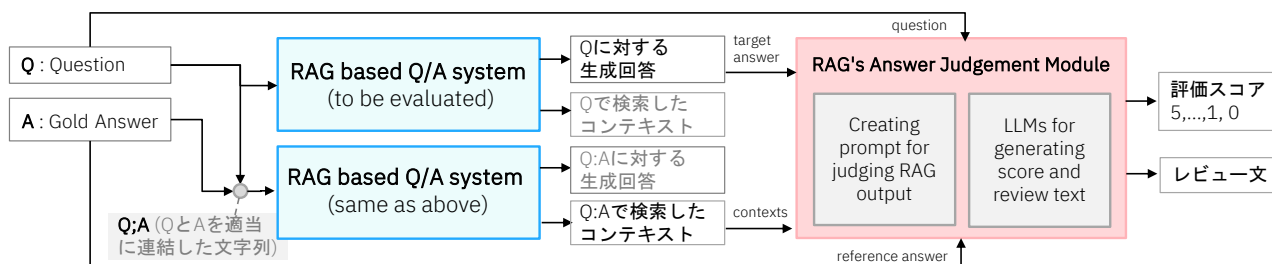


図1 RAGの応答正確性を自動評価するための評価パイプライン

評価する方法 [7] などが提案されているが、これらは我々の目的に十分には適してはいない。RAG は特定領域向けであることが多く、汎用的な評価用データセットでの評価は効果的とは言えず、また、回答の正確性が評価される場面で、対話の自然さを評価することも効果的とは言えないからである。

本論文では、質問と正答例からなる評価データセットを想定し、「評価対象 RAG との連続質問応答で、各質問に対して、その正答例のどの程度の情報を、どの程度の対話回数で、得ることができるか」という観点からの評価を考える。そして、LLM に基づく適切な回答判定、質問生成、回答構成のエージェントを用いて、この観点での評価を自動化する。

## 2 応答正確性の自動評価

評価データとして、質問と正答例からなる組みのセットを想定し、RAG の応答正確性を自動評価する手法を構築する。

基本的なアプローチは、Prometheus [4] と同様に、評価用の LLM のプロンプトに正確さの判定基準を  $k(=6)$  段階で与え、生成回答を  $k$  段階で評価することである。ただし我々は、コンテキスト（検索結果の文章リスト）の情報もプロンプトに含めることとする。結果として、評価用 LLM のプロンプトには、質問、コンテキスト、正答例、生成回答が含まれ、これらの情報に基づき生成回答の正確性が 6 段階: 良い～悪い (5, ..., 1), 無評価 (0) で評価されることとなる。プロンプト概要を付録の図 5 に示す。

ここで、評価時に使用するコンテキストについて考えたい。質問で検索した結果である（回答生成時に使用した）コンテキストを評価時に使用すると、コンテキストが正答の情報を含んでいなくとも、生成回答が、（正しい情報だと見做される）コンテキストに準拠しているという点で、高いスコアが出る可能性が考えられる。そこで我々は、この可能性を軽減するために、質問と正答例を同時に用いて検索し

たコンテキストを評価時に用いることを提案する。

### 2.1 提案手法 I

RAG の応答正確性を自動的に評価する評価パイプラインを図 1 に示す。評価データ（質問、正答例）ごとに、RAG で質問から回答を生成し、次に質問と正答例を用いて RAG でコンテキストを検索し、最後に、このコンテキストを正しい情報と見做し、正答例を最高スコア 5 の参照回答とした際、生成回答が質問の回答としてどの程度正確なのかを 6 段階のスコアで評価する。その上で、例えば、このスコアのデータセットでの（スコア 0 を除外した）平均が、RAG の応答正確性を表す指標になる。

RAGAs [5] とその実装ライブラリ [8] では、RAG の評価指標が、（現状で）8 つ提示されている。例えば、Faithfulness は、生成回答が質問で検索したコンテキストにどれだけ準拠しているかを評価し、Context Recall は、同コンテキストの有効さを、正答例と比較し評価する。しかし生成回答の正確性を直接的に評価する観点は含まれていない。そのためには、上記のような、コンテキストの有効さを前提としている指標と、その有効さを評価している指標を組み合わせる必要が生じる。本提案手法は、その組み合わせの一例と考えることができる。

### 2.2 評価実験と考察 I

提案手法の有効性を確認するために、IBM Z [9] とその関連製品に関する文書を知識ベースとして使用する RAG に基づく Q/A システム (QA) を準備した。これは製品 watsonx Assistant for Z [10] の対話型 Q/A 機能の実験用初期プロトタイプの一つである。

Db2 に関する 132 個の質問と正答例からなる評価データセット（全て英語）を準備し、QA で回答を生成し、その回答を次の 3 種の方法で評価した: (E1) 専門家による 6 段階評価; (E2) 質問で検索したコンテキストを用いた LLM によ

る評価; (E3) 提案手法での評価 (質問と正答例で検索したコンテキストを用いた LLM による評価)。評価基準は提案手法と全て同じである。その後、人手評価 (E1) と LLM による評価 (E2, E3) との相関を計測した。結果を図 2 に示す。評価用 LLM としては、prometheus-8x7b-v2 [11], mixtral-8x7b-instruct-v01 [12], llama-3-1-70b-instruct [13] を用いた。

平均値(スコア 0 を除外した平均)		専門家評価との相関	
専門家評価(E1):	4.007576	専門家評価(E1):	1.000000
word recall:	0.763517	word recall:	0.494306
rougeL :	0.122036	rougeL :	0.357340
bert score:	0.750147	bert score:	0.082107
prometheus(E2):	4.530303	prometheus(E2):	0.593374
mixtral(E2):	4.348485	mixtral(E2):	0.706555
llama3(E2):	3.742424	llama3(E2):	0.805254
Prometheus(E3):	4.143939	Prometheus(E3):	0.758834
mixtral(E3):	4.128788	mixtral(E3):	0.739532
llama3(E3):	3.734848	llama3(E3):	0.891785

図 2 応答正確性に関する評価実験の結果: E1,E2,E3 の評価スコア (の平均)  $s$  は, 1 (最も悪い)  $\leq s \leq 5$  (最も良い)。

図 2 によると、評価スコアの平均は、3 種のどの LLM を用いたとしても、E2 評価の方が E3 評価よりも高い。一方で、人手評価 (E1) との相関係数は、どの LLM を用いたとしても、E3 評価の方が E2 評価よりも高い。これは、生成回答を、質問と正答例で検索したコンテキストを用いて LLM で評価した方が、E2 評価と比べて、より人間の評価に近くなったことを示している。また、質問で検索した (そして回答生成に使用した) コンテキストを利用すると、E3 評価と比べて、評価スコアが平均的に上振れする傾向にあったことも示している。また、コンテキストを用いた E2, E3 による評価指標は、(コンテキストを利用せず) 生成回答と正答例のある種の類似度を計算する word recall, rouge-long, bert score などの指標よりも、人間の評価との相関が大幅に高くなった。

この実験結果は、E3 による評価、つまり、提案手法の一定の有効性を示していると考えられる。

### 3 連続応答性能の自動評価

次に、RAG の連続応答性能を自動評価する手法を構築する。ここでも評価データセットとしては、質問と正答例からなる組のセットを想定する。

質問応答の正確性において、RAG に基づく Q/A システムの連続応答の性能を評価するために、以下のような観点を考える：

「適当な質問者を仮定したとき、各質問に対して、その質問から始まる Q/A システムとの連続質問応答によって、正答例の情報を (どの程度の情報まで、

どの程度の対話回数で) 得ることができるのか？」

質問応答における Q/A システムとの対話では、質問の意図が正しく解釈されない、回答が十分に詳細でない、または回答が不正確に感じられるといった理由から、質問の表現や内容を修正したり、質問を簡略化して部分的な内容を尋ねる試行が頻繁に見られる。このような段階的な情報取得のプロセスは、質問応答の対話として広く行われている。そのため、上記の観点からの評価は、Q/A システムが持つ、応答の正確性に関する連続応答能力の重要な側面を示していると考えられる。

本論文では、この観点からの評価を、LLM を用いて適切に実現される回答判定、質問生成、回答構成を行うエージェントにより自動化する。

#### 3.1 提案手法 II

RAG の応答正確性に関する連続応答性能を自動的に評価する評価パイプラインを図 3 に示す。全体として、評価データセット中の各ペア (対象質問, 正答例) に対し、以下のタスク (T1..T7) による処理を繰り返し行うことで、各ターンの (暫定) 回答に対する評価スコアを含むリストを返すというものである：

(T1) 評価対象の RAG に質問をし回答を生成し、質問と生成回答を Q/A 履歴に保存する; (T2) Q/A 履歴から対象質問に対する (暫定的な) 回答を構成する; (T3) その構成回答を評価して評価スコアとレビュー文を生成する。評価スコアが 5 であれば正答が得られたので終了; (T4) レビュー文に基づいて、完全な回答を得るためのさらなる質問が必要であるかどうかを判断する; (T5) 質問が必要であれば次の質問を生成し、最大会話ターン数に達していなければ (T1) へ; (T6) 質問が必要なければ、現在の構成回答を正答例の粒度に書き直して最終回答にする; (T7) 最終回答を評価して評価スコアを出し終了する。

上記タスク (T3, T7) は回答評価エージェントが、(T2, T6) は回答構成エージェントが、(T4, T5) は質問生成エージェントが、それぞれ担当する処理である。参考として各エージェントが基づく LLM が使用するプロンプトの概要を付録の図 5, 6, 7 に示す。回答評価エージェントのプロンプトは 2 節の応答正確性を評価する LLM のプロンプトと同じであり、そこで使用されるコンテキストも元の対象質問と正答例から検索して得られるコンテキストである。

結果、最大ターン数  $n$  に対し、質問と正答例から評価スコアリスト  $\sigma = [s_0, \dots, s_k]$  ( $0 \leq k < n$ ) が得ら



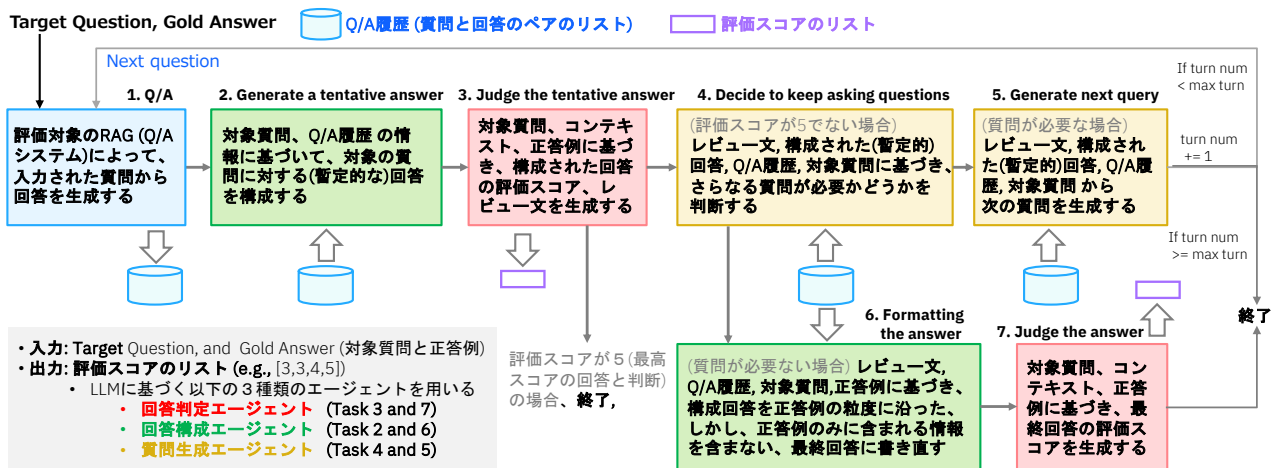


図3 RAGの応答正確性に関する連続応答性能を自動評価するための評価パイプライン

れる。 $\sigma$  は以下の3種の指標スコアで評価される。

$$\text{wscore}(\sigma) = \frac{ns_0 + \dots + (n-k)s_k + \sum_{i=1}^{n-k-1} (n-k-i)s_k}{n + \dots + 1}$$

$$\text{lscore}(\sigma) = \text{len}(\sigma) \quad \text{mscore}(\sigma) = \max(\sigma)$$

ここで、 $\text{len}(\sigma)$  はリスト  $\sigma$  の長さ、 $\max(\sigma)$  は  $\sigma$  が含む値の最大値を求める関数である。

$\text{wscore}(\sigma)$  は、重み付き評価スコアで、どれだけ早いターンで高い評価スコアの回答を出せるかを表す。例えば  $n=3$  のとき、 $\sigma' = [1, 5]$  に対して  $\text{wscore}(\sigma') = \frac{3 \cdot 1 + 2 \cdot 5 + 1 \cdot 5}{3 + 2 + 1} = 3$  である。 $\text{lscore}(\sigma)$  は、対話が終了するまでのターン数であり低いほどよい。 $\text{mscore}(\sigma)$  は、連続対話で得られた最良の回答の評価スコアを表す。全てのスコアの値は0以上  $n$  以下。尚、 $\text{lscore}(\sigma') = 2$ 、 $\text{mscore}(\sigma') = 5$  である。

最終的に、各々の指標スコアの評価データセットでの平均値等が連続応答の性能を表す指標となる。

### 3.2 評価実験と考察 II

提案手法の実効性を評価するために QA を標準的な形でマルチターンに対応させた MQA を準備した。MQA は、質問と Q/A 履歴から検索クエリを生成し、回答生成時に Q/A 履歴を利用する。MQA の連続応答性能は QA より高いと想定できる。

IBM Z 関連の10個の製品に関する432個の質問/正答例(全て英語)を準備し、QA, MQA の連続応答性能を提案手法(最大ターン数5)を用いて計測した。結果を図4に示す。各エージェントの出力を見ると、回答判定によるレビュー文に基づいて、質問生成エージェントが、新たな細かい質問/間違い部分を確認する質問を生成し、Q/A システムから追加

QA: 各スコアの平均	MQA: 各スコアの平均
wscore (↑): 3.535207	wscore (↑): 3.606557
lscore (↓): 3.147541	lscore (↓): 3.049180
mscore (↑): 4.185012	mscore (↑): 4.266979

図4 連続応答性能に関する評価実験の結果(↑/↓は、値が高い/低い方が連続応答性能が高いことを示す。)

情報を得て、獲回答構成エージェントが最終的に正答例に相当する回答を構成するといった会話が頻繁に見受けられた。会話の実例を付録の図8に示す。

図4によると、wscore と mscore の平均値は MQA の方が QA より高く、lscore の平均値は MQA の方が低い。また、この傾向は、製品ごとの10個のデータセット各々での計測においても、一貫して成立していた。付録の図9にこれらの計測値を示す。

この実験結果は、MQA が QA よりも、平均的に、より早くより正確な回答を生成し、最終的により短い対話数でより正確な回答を生成したことを示している。このことは、提案手法が、応答の正確性に関する連続応答性能の差を検出可能であることを示しており、本手法が一定程度、実効的であることを示唆していると考えられる。

## 4 まとめ

RAG による応答の正確性を「質問と正答例に基づいて検索されたコンテキスト」を用いて自動評価する方法、および、連続応答の性能を「連続的な質問応答で正答例の情報をどの程度、どれだけ早く引き出せるか」という観点から自動評価する方法を構築した。また、予備的な実験により、その有効性と実効性を一定程度確認した。他の評価指標(RAGAs など)との比較や、連続応答における評価観点の更なる考察、評価手法の洗練化が今後の課題である。

## 参考文献

- [1] Eric J. Breck, John D. Burger, Lisa Ferro, Lynette Hirschman, David House, Marc Light, and Inderjeet Mani. How to evaluate your question answering system every day ... and still get real work done. In M. Gavriliadou, G. Carayannis, S. Markantonatou, S. Piperidis, and G. Stainhauer, editors, **Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)**, Athens, Greece, May 2000. European Language Resources Association (ELRA).
- [2] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In **8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020**. OpenReview.net, 2020.
- [3] Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing fine-grained evaluation capability in language models. In **12th International Conference on Learning Representations, ICLR 2024, 2024**.
- [4] Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 4334–4353, Miami, Florida, USA, November 2024. Association for Computational Linguistics.
- [5] 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.
- [6] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In **Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23**, Red Hook, NY, USA, 2024. Curran Associates Inc.
- [7] Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. MT-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 7421–7454, Bangkok, Thailand, August 2024. Association for Computational Linguistics.
- [8] Ragas : List of available metrics, Accessed on 01/04, 2025. [https://docs.ragas.io/en/stable/concepts/metrics/available\\_metrics/](https://docs.ragas.io/en/stable/concepts/metrics/available_metrics/).
- [9] IBM. IBM Z, Accessed on 01/04, 2025. <https://www.ibm.com/jp-ja/z>.
- [10] IBM. watsonx Assistant for Z, Accessed on 01/04, 2025. <https://www.ibm.com/jp-ja/products/watsonx-assistant-for-z>.
- [11] Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. prometheus-eval/prometheus-8x7b-v2.0, Accessed on 12/30, 2024. <https://huggingface.co/prometheus-eval/prometheus-8x7b-v2.0>.
- [12] Mistral AI. mistralai/mixtral-8x7b-instruct-v0.1, Accessed on 12/30, 2024. <https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1>.
- [13] Meta. meta-llama/llama-3.1-70b-instruct, Accessed on 12/30, 2024. <https://huggingface.co/meta-llama/LLama-3.1-70B-Instruct>.

## A 付録

```

###Task Description:
An instruction, a question, context documents, a response to evaluate, a reference answer that gets a score of 5, and a score rubric representing an evaluation criteria are given.
1. Write a detailed feedback that assesses the quality of the response strictly based on the given score rubric, not evaluating in general.
2. After writing a feedback, write a score that is an integer between 0 and 5. You should refer to the score rubric.
3. The output format should look as follows: "Feedback: (write a feedback for criteria) [RESULT] (an integer number between 0 and 5)"
4. Please do not generate any other opening, closing, and explanations.
###The instruction to evaluate:
Evaluate the below given response based on the contexts and the reference answer. The evaluation should be based strongly on the fact that the reference answer is a correct answer to the question. So make sure if the response implies the reference answer, the response is also correct.
###Contexts: ¥n {CONTEXTS}
###Question: ¥n {QUESTION}
###Response to evaluate: ¥n {TARGET_ANSWER}
###Reference Answer (Score 5): ¥n {REFERENCE_ANSWER}
###Score Rubrics:
[If the context contains the correct information and the reference answer is completely correct, what is score of the response based on the following criteria?]
Score 0: The response states it is not sure.
Score 1: The response is completely wrong, and fatal.
Score 2: The response is mostly wrong, but not very fatal.
Score 3: The response is partially correct and partially wrong.
Score 4: The answer is largely correct, but incomplete.
Score 5: The response is correct and complete.
###Feedback:

```

図 5 生成回答の評価 LLM のプロンプトの概要

```

### System prompt
You are trying to ask a series of questions to the Question and Answer System (QA), which uses the entire session history for generating answers, and get an answer that the Evaluation System (EVAL) would judge the answer as a complete correct answer (of score 5) for the following target question: " {SORG_QUESTION} ".

### User prompt (for task 3)
The below is a history of questions and answers between you and QA:¥n {QA_HISTORY}

Generate a correct and complete answer to the target question by selecting and summarizing information in the history of Q/A session. Note that the responses from the QA are not necessarily correct, but only information contained in, or that can be inferred from, the answers from QA must be used.

Please be sure to use the following format in your response:
Answer: <answer>, Explanation: <explanation>

### User prompt (for task 7)
The following is a question and answer history between you and QA in its current state. (Note that the responses from the QA are not necessarily correct.): {QA_HISTORY}
Based on this Q/A history, you have the tentative answer " {TGT_ANSWER} " to the target question.
A reference answer (of score 5) for the target question is that " {GOLD_ANSWER} ".

Revised the tentative answer to a complete and correct answer. The revised response may delete or rephrase information in the tentative response or add information inferred from previous responses from the QA, according to reference answer, but may not contain information described solely in the reference answer.

Please be sure to use the following format in your response:
Answer: <revised-answer>, Explanation: <explanation>

```

図 6 回答構成 LLM のプロンプトの概要

```

### System prompt
You are trying to ask a question to the Question and Answer System (QA), which uses the entire session history for generating answers, and get an answer that the Evaluation System (EVAL) would judge the answer as a complete correct answer (of score 5) for the following target question: " {TGT_QUESTION} "

### User prompt
The following is a question and answer history between you and QA in its current state. (Note that the responses from the QA are not necessarily correct.): {QA_HISTORY}

Based on this Q/A history, you have the tentative answer "{TGT_ANSWER}" to the target question.
EVAL's evaluation of this tentative answer is that " {Judgement_RESULT} ".
A reference answer (of score 5) for the target question is that " {GOLD_ANSWER} ".

Complete the below tasks:
1. Decide if additional queries are needed to the QA to obtain or confirm any unknown information for revising the tentative answer to ones that EVAL would judge to be the complete correct answer to the target question. Here, you must not use information contained only in the reference answer or evaluation from EVAL without first validating such information with QA.
2. If additional queries are needed, generate a query that should be made to QA to directly obtain or directly confirm information that is still missing or unknown in order for EVAL to judge to be the complete correct answer to the target question. The query must be different from previous questions and can include information contained in the EVAL evaluation results. (If no additional questions are needed, leave blank).

Please be sure to use the following format in your response:
Query: <query-statement>, Explanation: <explanation>.

```

図 7 質問生成用 LLM のプロンプトの概要

会話例:実線部分は間違い、破線部で間違いを指摘、点線部で間違いを確認する質問を生成

**質問:** What do SQL statements UNION and UNION ALL do and what are the difference between them ?

**正答例:** They are both used to combine the results of SELECT statements. UNION eliminates duplicates, UNION all does not.

**[Task1] MQAの生成回答:** SQL statements UNION and UNION ALL are used to combine the results of two or more SELECT statements into a single result table. The difference between them is that UNION will combine the results without eliminating duplicates, while UNION ALL will combine the results and eliminate duplicates.

**[Task3] 回答判定エージェント:** The response is partially correct, but incomplete. It correctly states that UNION and UNION ALL are used to combine the results of SELECT statements, and that UNION ALL eliminates duplicates. However, it incorrectly states that UNION does not eliminate duplicates, when in fact, UNION eliminates duplicates. **Score: 3**

**[Task4, 5] 質問生成エージェント:** Query: Does UNION eliminate duplicates or not?  
Explanation: The tentative answer is incomplete and incorrect according to EVAL's evaluation. It states that UNION does not eliminate duplicates, but EVAL says it does. To confirm this, a direct question is needed to clarify the behavior of UNION regarding duplicates.

**[Task1] MQAの生成回答:** Yes, the SQL keyword UNION does eliminate duplicates. By itself, UNION returns all of the values from the result table of each SELECT statement, but if you specify UNION ALL, it will return all duplicate rows in the result table.

**[Task2] 回答構成エージェント:** Answer: SQL statements UNION and UNION ALL are used to combine the results of two or more SELECT statements into a single result table. The difference between them is that UNION eliminates duplicates, while UNION ALL does not eliminate duplicates and returns all rows, including duplicates.

**[Task3] 回答判定エージェント:** The response is correct and complete. It accurately explains the difference between UNION and UNION ALL, stating that UNION eliminates duplicates while UNION ALL does not. The response is faithful to the reference answer and the context information. **Score: 5**

図 8 エージェントの出力例 (会話例)

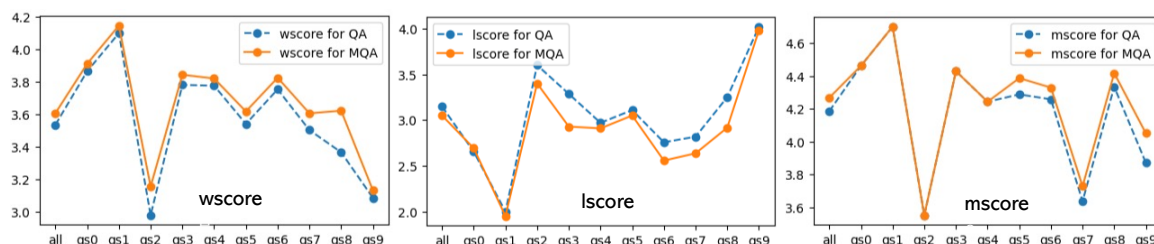


図 9 各指標スコア wscore, lscore, mscore の平均を製品ごとのデータセット  $qa_i (i = 0 \dots 9)$  に対して求めてプロットしたグラフ。青点 (点線) が QA、黄点 (実線) が MQA を評価対象とした際の指標スコアの平均を示す。