LLM によるクイズの自動生成と質問応答への応用

小林俊介 [1]　河原大輔 [1,2]
[1] 早稲田大学理工学術院　[2] 国立情報学研究所 LLM 研究開発センター
{carlike787@toki., dkw@}waseda.jp

概要

生成モデルとして広く使用されるようになった大規模言語モデル (LLM) の使用用途の 1 つとして、学習データ生成が検討、研究されている。本研究では、LLM を用いたクイズデータの自動生成と評価により、学習データの拡張を実施し、質問応答に応用する。生成したクイズについて LLM と人手による評価を比較した結果、LLM は基本的なルールや文法を理解してデータを評価できる一方で、文字数のカウントなど一部に苦手な要素が存在することを確認した。また、生成されたクイズデータを質問応答システムに応用した結果、人手の学習データには及ばないものの、学習効果があることを確認した。

1　はじめに

近年、大規模言語モデル (LLM) は進化するとともに様々な分野・場面で応用されている。その 1 つに LLM を用いた学習データ生成が注目されている。従来の人手によるデータ作成でボトルネックとなっていた、作成コストや専門的人材確保の問題の解決策として提案されている。しかし、これまでの研究では、言語やタスクなどの違いによって、効果の有無が大きく異なっており [1, 2, 3, 4]、その可能性は未知数であるといえる。

本研究では、LLM を用いたデータ生成を、クイズを題材とした質問応答タスクに応用する。この時、プロンプトで質問を段階的に生成するなどの工夫を取り入れる。一方で、生成された質問は、矛盾や不自然な表現を含む場合や、事実と異なる解答を持つ場合があり、そのまま応用するには不適切な場合がある。そのため生成したクイズの評価、フィルタリングを行う。近年、LLM による生成結果の評価能力が向上していることを受け、本研究ではクイズの生成にとどまらず、評価までを一貫して LLM により実施する手法を提案する。さらに、生成されたクイズデータで質問応答システムを学習し、学習の有効性を検証する。

生成されたクイズに対する LLM による評価を人手評価と比較、検証した結果、LLM は人間と同レベルで実施できる評価とそうでない評価を有することを確認した。また、生成されたクイズを質問応答システムに応用した結果、学習データとしての有効性を確認した。本提案手法を用いた学習データの自動生成により、データ量が少ない、もしくは専門的人材が確保できないなどの理由で学習データの確保に困難を伴う分野での応用が期待される。

2　関連研究

英語における質問応答タスクのデータセットには、SQuAD [5]、TriviaQA [6]、Natural Questions [7] などがある。日本語では、JGLUE [8] に含まれている JSQuAD や、運転ドメイン QA データセット [9]、JAQKET [10] などがある。SQuAD と TriviaQA ではデータが 100,000 件前後、Natural Questions では 300,000 件以上用意されている。一方、JSQuAD は 64,000 件弱、運転ドメイン QA データセットは 34,000 件強、JAQKET では 24,000 件前後であり、英語データセットより 1 桁少ないデータ量である。

学習データが十分に得られない場合の対策として、既存のデータを用いて擬似的なデータを生成する場合がある。特に自然言語処理では、擬似データの生成を LLM で 1 から実施する研究が存在する。日本語の例では、藤井ら [1] が、文書分類や感情分析に向けた擬似データ生成のフレームワークを検討しており、タスクによって効果的なものが存在することを報告している。また、Puri ら [2] は、SQuAD 形式の質問応答データセットを GPT-2 によって擬似生成し、質問応答タスクの精度が向上することを確認している。一方で、Li ら [3] は、様々なドメインにおける文書分類で、LLM による擬似データ生成を実施したが、最大で 40%の精度低下が確認されるなど、ドメインによっては擬似データで十分な学習効果が得られないことを報告している。さらに、

Shumailov ら [4] は、擬似データによるモデルの学習が、最終的にはモデルの推論能力を失わせることを示し、言語モデルのファインチューニングにおいてこれが表出することを確認している。

3 クイズ生成の手法

本研究ではクイズ生成を、質問と解答を生成するステップと、生成された質問・解答のペアを評価するステップに分けて行う。

3.1 LLM によるクイズ生成

まず、LLM を用いて質問と解答のペアを生成する。クイズ生成においては、質問と解答の対応、質問そのものに事実と異なる内容が含まれていないことが重要である。そのため、質問・解答の根拠とする文書として、Wikipedia 記事を LLM の入力 (プロンプト) に含める。その際、一部の記事は非常に長くなることを踏まえ、第一段落以外で分量の多いセクションを除去する前処理を行う。プロンプトでは、「質問は 40 文字以上とすること」などの基本的なルールを指定するとともに、Chain-of-Thought [11] や few-shot learning [12] などを利用し、質問生成において有益な情報を含めるなどの工夫を施す。以下にその一部を示す。

- 質問の生成を前半・後半で分割し、独立に行う
- 「〜ですが、」のようなクイズに特有の文体についての説明を導入する
- クイズ作家が作成したクイズの例をプロンプト内で示す

さらに、クイズ生成段階で同時に、質問の難易度を 1 から 5 までの 5 段階で自己評価する。この評価は後述の 3.2 節で使用する。以上の手順で、各記事から質問と解答のペアを 3 つ生成する。その際、記事中のどの部分を質問、解答の根拠にしたかを同時に出力させる。これを「根拠情報」と呼ぶ。

3.2 LLM によるクイズ評価

次に、生成された質問と解答のペアを LLM で評価する。評価基準は、質問が解答との対応や文法面などで正確であるか、根拠情報が意外性を含むものであるか、の 2 点を設定する。これに加え、より多くの人にクイズへの解答意志を生むには、極端に簡単、難解でないクイズであることが必要と考え、自己評価した難易度が 3 に近いかどうかについても評価する。5 点を基準として、これらの評価基準に沿った評価を実施し、スコアの加算・減算を行う。想定される最高点は 12 点とするが、LLM がこれを超える点数をつける場合もある。ここで算出されたスコアは、質問応答システムの学習時 (5 節) のフィルタリングに用いる。

4 クイズ生成の実験

4.1 実験設定

2020 年から、質問応答タスクのコンペティションとして「AI 王」が開催されており、データセットとして JAQKET が用いられている。本研究では、同コンペティションの第 3 回大会[1)]で使用されたデータセット[2)]を用いる。このデータセットに含まれる質問と解答のペアの中から、クイズ生成の例示に用いるペアを十数点抽出する。根拠情報とする Wikipedia の記事は、mediawiki[3)] ライブラリを用いて取得する。クイズの生成と評価では、OpenAI が提供している GPT-4 Turbo[4)] を LLM として使用する。

クイズの生成では、20,000 件の Wikipedia 記事から生成し、合計で 60,000 問を得る設定とする。

生成したクイズの評価では、まず生成されたペアを、3 点ごとのスコア帯別に分類する。その後、それぞれからサンプリングを行い、100 件のクイズを抽出する。これらに対し、人手による定量的および定性的評価を行う。定量的評価は専門家による評価と、クラウドソーシング[5)]による評価の 2 つを実施する。前者では、LLM にプロンプトとして与えた評価基準とサンプリングしたクイズ、根拠情報を用いて、2 名の専門家が LLM と同様の評価を実施する。この専門家評価と LLM による評価がどれだけ一致するかを調査する。後者のクラウドソーシングによる評価は、比較的簡単に評価ができ、クイズとして成立するか否かを決定づける要素を対象とし、文法と情報の引用の正確さの 2 点で実施する。異なる 5 名のワーカーが、2 つの観点について「はい」「わからない」「いいえ」の 3 択で評価する。これらをそれぞれ 1, 0, -1 の数値に変換したうえで平均を

1) https://sites.google.com/view/project-aio/competition3
2) https://github.com/cl-tohoku/AIO3_BPR_baseline/blob/main/download_data.sh に記載のスクリプトでダウンロードできる。
3) https://github.com/barrust/mediawiki
4) 2024 年 1 月 25 日発表のもの
5) プラットフォームに Yahoo!クラウドソーシングを利用。

取り、以下のように分類する。

1. 否定的: 平均値が-0.2 未満
2. 中立的: 平均値が-0.2 以上 0.2 以下
3. 肯定的: 平均値が 0.2 より大きい

また、定性的評価として、サンプリングされたクイズの中からいくつかを抽出し、専門家の評価と LLM の評価の差を分析する。

4.2 生成結果の分析

生成の結果、一部の入力で LLM が結果を出力しない場合があり、56,469 件のクイズを得た。

4.2.1 LLM と専門家による評価の比較

まず、専門家の評価と LLM による評価のスコアを比較した結果、両者の評価の相関係数は 0.20 と、わずかに正の相関がみられた。以下は両者の評価に差があるクイズと、根拠情報の例[6]である。

質問: 「スリラー」ミュージック・ビデオでジャクソンの恋人役を演じた女優は誰でしょう?
解答: オーラ・レイ
情報: 1. オーラ・レイがジャクソンの恋人役として共演した
2. ジャクソンがビデオで初めて女性と共演した作品

この例では、LLM が 13.5 点と評価したのに対し、専門家は平均で 1 点を付けた。LLM が採点したスコアが想定した最高点を上回っているが、その原因の 1 つとして、情報の意外性を評価する段階において、範囲を超えた加点を行ったことが考えられる。また、生成時に「質問は 40 文字以上とすること」と定めたため、39 文字しかないこの例は減点が必要である。しかし LLM の評価スコアからはこの減点がなかったことが推測され、LLM は文字数を捉えるのが苦手であることも推察される。

一方で、文法の誤りは人間・LLM がともになしと判定するなど、LLM が人間と同水準の評価を行っていること、また質問自体はクイズとして持つべき要素を有しており、クイズの生成自体は実行できていることも観察された。次の例は、根拠情報とした記事が 2024 年 6 月に初めて掲載されたもの[7]である。

質問: 『ふつうの軽音部』で、ちひろが初めて結成したバンドの名前は何でしょう?
解答: ラチッタデッラ
情報: 1. ラチッタデッラメンバーはちひろ、厘、カッキー、ヨンス。ちひろが初めて結成したバンドだったが色々あって早々と解散。
2. 鳩野ちひろ（はとの ちひろ）主人公。1 年。

評価では LLM が 11 点、専門家の評価の平均は 7.5 点と多少の差があるものの、双方の評価が基準点である 5 点を上回っている。また、本実験で使用した LLM は 2023 年 12 月までのデータで学習されており、学習範囲にないデータからクイズを生成できている例となっている。

以上の分析から、LLM によるクイズの生成はできている一方で、生成時に指定したルールに従っているかの評価は人間と異なり、苦戦する要素があることがうかがえる。

表 1 クラウドソーシングによる評価の結果

		情報		
		否定的	中立的	肯定的
文法	否定的	0	4	1
	中立的	0	3	5
	肯定的	7	24	56

4.2.2 クラウドソーシングによる評価

次に、クラウドソーシングによる各項目の評価結果をクロス集計したものを表 1 に示す。双方の評価が肯定的であったクイズが 56 件、否定的な評価を 1 つも得なかったクイズは 89 件であった。多くの生成されたクイズがルールに従って生成されていることが、クラウドソーシングによる評価でも確認された。評価項目別にみると、文法の評価はほとんどが肯定的な評価になっているのに対し、情報の評価では中立的、または否定的な評価を受けたクイズが多かった。このことから、LLM は複数の情報を組み合わせて出力を生成すること、または自身が生成した出力の根拠を引用することが得意ではないことが推察される。

5 質問応答への応用実験

5.1 実験設定

生成したクイズデータを用いて、質問応答システムの学習を行う。システムには、事前学習された日

6） Wikipedia 記事「スリラー (ミュージック・ビデオ)」から生成、CC BY-SA 4.0 国際 パブリック・ライセンス
7） Wikipedia 記事「ふつうの軽音部」から生成、CC BY-SA 4.0 国際 パブリック・ライセンス

表 2　質問応答システムの学習で用いるデータセット

データセット名	データ数
AI 王学習データ	15,556
全生成データ	30,642
上位 50%データ	21,105
上位 30%データ	12,678

本語 T5 [13][8] をベースとした、質問応答モデルの一つである Fusion-in-Decoder [14][9] (FiD) を用いる。FiD では、質問と、これに関連した 100 件の文書を入力し、解答を出力する。

前処理として、関連文書の収集と、生成したクイズデータのフィルタリングを行う。関連文書の収集については、AI 王で提供されている BPR [15][10] と日本語 Wikipedia 記事の文書集合を用いて、各質問に関連する文書を 100 件ずつ抽出する。この中に解答が存在しないクイズは除外する。

生成されたクイズには品質が低いものが含まれうることを踏まえ、3.2 節で算出したスコアにより、クイズデータをフィルタリングし、上位スコアのデータを抽出する。このスコアの上位割合を変化させ、生成データによるデータセットを 3 種類構築した。

比較として、専門家が作成した AI 王の学習データを用いる。これについても関連文書の収集を行う。実験で用いるデータセットの詳細を表 2 に示す。

得られたデータセットを用いて、FiD の学習を実施する。評価においては、AI 王が提供する評価用データセットを利用する。まず評価用データセットの質問を BPR に入力し、関連する日本語 Wikipedia の文書 100 件を抽出する。次に、学習データの選定と同様に、この 100 件中に解答が書かれた文書が存在する 1,791 問のクイズを抽出する。最後に、抽出された質問を文書とともに FiD に入力し、解答を得る。そのうえで、正解と、生成された解答の完全一致の割合である Exact Match (EM) により評価する。この実験を各データセットで、異なるシード値を用いて 5 回実施し、評価値の平均を算出する。

5.2　学習したモデルの評価

図 1 に、それぞれのデータセットを用いて学習した場合の、EM 正答率の変化を示す。生成されたデータで学習した結果、いずれのデータも最高で 70%前後の正答率に達し、学習が進んでいること

8）https://huggingface.co/retrieva-jp/t5-base-long
9）https://github.com/facebookresearch/FiD
10）https://github.com/cl-tohoku/AIO3_BPR_baseline

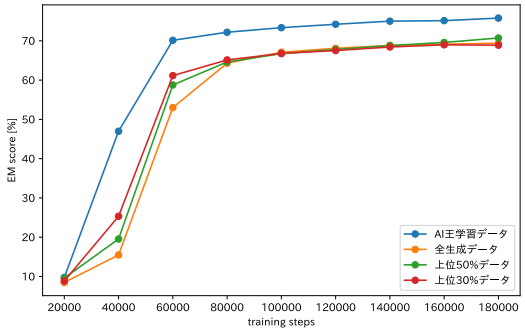

図 1　学習中のモデルの評価データによる精度の変化

が確認できる。AI 王学習データの精度には及ばなかったが、生成されたデータは一定の品質を有していることが推察される。

生成データで学習した結果のみで比較すると、最終的な正答率は上位 50%データで一番よく、全生成データ、上位 30%データと続いた。フィルタリングの程度と一貫性が取れない結果になっているが、LLM 及び専門家による評価の相関係数が低かったことを踏まえれば、この差は評価の品質に起因するものと考えられる。一方で、8 万 step までの正答率に注目すると、概ね厳しいフィルタリングを施したデータセットで早い段階から向上している。高品質なデータに絞って学習することで、学習速度が向上するという点では、フィルタリングを行うことに一定の効果があったといえる。

6　おわりに

本研究では、質問応答タスクの 1 つであるクイズを題材に、LLM を用いたデータ生成、ならびに質問応答システムの学習への応用を試みた。LLM で生成したデータの評価では、専門家と評価が乖離する要素もあったが、基本的な部分では人間と同等のレベルで評価できていることが確認できた。生成されたデータの応用では、専門家が制作したデータには及ばなかったが、学習データとしての利用可能性が示された。また、評価スコアによるフィルタリングにより、学習の前半段階での精度向上が早まることを確認した。

今後の課題として、LLM がより評価しやすい項目による評価の実施や、生成段階でより学習効果が期待できる質問を生成できるような工夫、そして Wikipedia 記事に限らない、多様な入力からの質問生成が挙げられる。

謝辞

本研究は文部科学省補助事業「生成 AI モデルの透明性・信頼性の確保に向けた研究開発拠点形成」の支援を受けた。本研究の生成クイズデータの評価についてアドバイスいただいた、早稲田大学の佐々木斗海氏に心から感謝申し上げる。

参考文献

[1] 藤井 巧朗, 勝又 智. 日本語タスクにおける llm を用いた疑似学習データ生成の検討. 言語処理学会第 30 回年次大会 発表論文集, pp. 2284–2289. 言語処理学会, March 2024.

[2] Raul Puri, Ryan Spring, Mohammad Shoeybi, Mostofa Patwary, and Bryan Catanzaro. Training question answering models from synthetic data. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 5811–5826, Online, November 2020. Association for Computational Linguistics.

[3] Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. Synthetic data generation with large language models for text classification: Potential and limitations. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 10443–10461, Singapore, December 2023. Association for Computational Linguistics.

[4] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. The curse of recursion: Training on generated data makes models forget. arXiv, 2024. abs/2305.17493.

[5] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics.

[6] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[7] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. **Transactions of the Association for Computational Linguistics**, Vol. 7, pp. 452–466, March 2019.

[8] Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. JGLUE: Japanese general language understanding evaluation. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 2957–2966, Marseille, France, June 2022. European Language Resources Association.

[9] Norio Takahashi, Tomohide Shibata, Daisuke Kawahara, and Sadao Kurohashi. Machine comprehension improves domain-specific Japanese predicate-argument structure analysis. In **Proceedings of the 2nd Workshop on Machine Reading for Question Answering**, pp. 98–104, Hong Kong, China, November 2019. Association for Computational Linguistics.

[10] 鈴木正敏, 鈴木潤, 松田耕史, 西田京介, 井之上直也. JAQKET: クイズを題材にした日本語 QA データセットの構築. 言語処理学会第 26 回年次大会 (NLP2020) 発表論文集, pp. 237–240, Online, March 2020. 言語処理学会.

[11] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. arXiv, 2023. abs/2201.11903.

[12] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. **Advances in neural information processing systems**, Vol. 33, pp. 1877–1901, 2020.

[13] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. **Journal of Machine Learning Research**, Vol. 21, No. 140, pp. 1–67, 2020.

[14] Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume**, pp. 874–880, Online, April 2021. Association for Computational Linguistics.

[15] Ikuya Yamada, Akari Asai, and Hannaneh Hajishirzi. Efficient passage retrieval with hashing for open-domain question answering. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)**, pp. 979–986, Online, August 2021. Association for Computational Linguistics.

A　学習設定

5 節で行った実験時に設定したハイパーパラメータを表 3 に示す。

表 3　学習時に設定したハイパーパラメータ

合計ステップ数	180,000
合計バッチサイズ	100
学習率	2.0×10^{-5}
トークン数上限 (reader)	200