

# 教育現場における質問の性質分析と大規模言語モデルを活用した質問回答システムの検討

原田憲旺<sup>1</sup> Edison Marrese-Taylor<sup>1,2</sup>

岩澤有祐<sup>1</sup> 松尾豊<sup>1</sup>

<sup>1</sup> 東京大学, <sup>2</sup> 産業技術総合研究所

{keno.harada, emarrese, iwasawa, matsuo}@weblab.t.u-tokyo.ac.jp

## 概要

大規模言語モデルの発展により、作業の効率化・自動化や創造性の拡張が期待される。教育分野においても応用例の一つとして学習者の質問に答え学びを促進させる QA システムの開発が期待されるが、現実の利用シーンに即したベンチマークが整備されていないため、QA システムが対処すべき質問の性質が不明なため、開発で注力すべき機能の考慮が行えない。本論文では、教育現場での QA システム開発の一助となることを目指し、実際に行われた講義中において生徒から提出された質問を分析した結果を公開する。また大規模言語モデルを活用して得られた回答文を分析することで今後の課題を明らかにする。

## 1 はじめに

大量のテキストデータで学習が行われ、タスクを表す指示文への応答を柔軟に行うように訓練された大規模言語モデルは、さまざまな自然言語処理タスクにおいて高い性能を発揮している [1, 2]。大規模言語モデルを活用することによって人間が日々行うテキストによる知的作業を効率化・自動化し、また創造性を拡張することが期待される。そのため、さまざまな業界において業務効率化によるコスト低減や、新たな価値創出の機会が探られている。教育業界においても、大規模言語モデルを活用することによって学びを行う生徒の理解度を促進するために、生徒の質問に答える質問回答システムの開発が期待される [3]。

質問と回答、場合によっては回答根拠となる文書群のペアを集めた質問回答データセットは数多く整備されているものの、教育現場での応用を見据えた質問を集めたデータセットは少なく [4]、さらに日

本語でのデータセットは少ない。本論文では、教育現場での QA システム開発の一助となることを目指し、実際に行われた講義中において生徒から提出された質問を分析した結果を公開する。既存の質問回答データセットの質問とは異なり、教育現場で得られる質問には、質問文単体では回答のためのコンテキストが足りないものが多く、質問回答の際には受講生の理解度を推し量った上での回答、理解の手助けが必要であるという点を明らかにする。また大規模言語モデルを活用した質問回答システムの検討を行い、今後の課題について述べる。

## 2 関連研究

### 2.1 Long-form Question Answering データセットと評価指標

質問回答データセットには、回答の性質によって分類を行うと単語・文章の抜き出しによって回答を行う Extractive QA、選択肢によって回答を行う Multi-choice QA、Yes/No の分類によって回答を行う Categorical QA、自由記述によって回答を行う Freeform QA に分類される [4]。Freeform QA の中でも、質問に対して、回答根拠となるような文章を参照し、複数文からなる回答をする必要のある QA は Long-form QA と呼ばれる。教育現場で生徒から得られる質問に対する回答では、生徒が十分に理解を行っていないトピックの詳細な説明が必要だと考えられるため Long-form QA に位置付けられる。Long-form QA にはオンライン掲示板において、前提知識を持たない人へ何かしらのトピックについて説明を行うやり取りから作成された ELI5 [5] や、検索エンジンにユーザーから実際に入力されたクエリを集めて作成された Natural Questions [6] などのデータセットがあるものの、実際の教育現場において見られる質問を集めたデータセットは少ない。

**質問:** 大規模言語モデルにおけるバイアスの問題はなぜ起こるのですか？カラクリをあまり理解できませんでした。

**回答:** 大規模言語モデルにおけるバイアスの問題は、主に以下の理由で起こります。1. データ選択バイアス：トレーニングコーパスを構成するテキストの選択によって生じるバイアスです。モデルが学習するデータが偏っていると、その偏りがモデルに反映されます。2. 社会的バイアス：トレーニングデータに含まれる社会的偏見がモデルに継承されること、性別、年齢、性的指向、民族、宗教、文化などに関するバイアスが生成されたテキストに現れます。...(生成文続)

Queried results

Biases in Large Language Models: Origins, Inventory, and ... (URL: <https://dl.acm.org/doi/10.1145/3597307>)

A Trip Towards Fairness: Bias and De-Biasing in Large Language ... (URL: <https://arxiv.org/abs/2305.13862>)

There's More to AI Bias Than Biased Data, NIST Report Highlights ... (URL: <https://www.nist.gov/news-events/news/2022/03/theres-more-ai-bias-biased-data-nist-report-highlights>) OpinionGPT demonstrates the impact of training data on AI bias (URL: <https://the-decoder.com/opiniongpt-demonstrates-the-impact-of-training-data-on-ai-bias/>) ...(検索結果続)

**図1** 質問文中にて言及されるトピックについて検索エンジンを用いて記述を抽出し、記述を踏まえた上で回答を行う質問回答システムによって得られた出力の例。

質問回答において、ユーザーの入力する質問自体が完全でなく質問意図がわからないため回答に困難が生じる場合や、質問が完全であったとしても回答の候補が複数あるものなど曖昧性が生じる。質問・回答における曖昧性については Ambig QA[7] のようなデータセットで考慮がなされているものの、教育現場において見られる質問・回答の曖昧性については整理されていない。

Long-form QA は回答が自由記述のため、回答文の評価が難しいことが課題である。前もって準備した正解文と生成文を照らし合わせて評価を行う ROUGE[8] や MAUVE[9] といった指標や、人手評価によって回答の好ましさを段階評価やランクづけによって行う方法がある。評価の際には文章の自然さや、回答の正確性・信頼性など多角的な軸で評価すべきであることが指摘され [10]、回答文で触

れられている項目が正しいか信頼性の評価を行う FActScore[11] や、判断根拠として列挙した文章が質問に関連する文章であるかを評価する指標 [12] などが提案されている。

### 3 データセットの作成

データセットの作成手順について説明する。2023年9月に東京大学において開催された大規模言語モデルに関する講座の期間内で、実際に受講生から提出された質問をもとに作成された。講座の受講生は学生（社会人学生も可。ただし大学生と社会人学生の場合、学士以上の学位取得ができる学生が対象）・東京大学の教職員・国家、地方公務員・研究者・スタートアップ企業の社員の合計約2000人である。受講生の予備知識や事前準備の目安として、「Zoom, Google サービス (google drive, google form, google colab 等), Web ブラウザ, Slack を利用できる PC が用意できること, 毎回の講義を確実に視聴できること, 最低でも毎回3時間程度の自習時間が確保できること, 大学の理系学部レベルの線形代数, 微分積分, 確率論・統計学に関する知識を有すること, 機械学習の基礎知識を備え, 基礎的な実装が自身でできること (Pytorch, TensorFlow, JAX などのフレームワークを使ってモデルの構築・訓練・評価を行った経験があること)」を案内しているが、事前知識を問うクイズなどによって受講可否は判断されず、講義受講希望者は所属によって受講可否が伝えられた。

受講生はリアルタイムで開催されている講義中、あるいは後日公開されるアーカイブ動画を視聴中に質問を受け付ける専用のフォームから質問を提出した。1回2時間の講義が計7回開催され、受講生から提出された質問の合計は884件である。質問に対して回答は講義の講師から口頭によって、あるいはTA・講師によってテキストで行われた。884件中、講義の事務連絡やシステム障害などに関連する問い合わせを除いた、講義の内容に関する質問は715件であった。講義の内容に関する質問715件のうち質問・回答ともに「スライドを参照」「こちらを参照」などの文言が含まれておらずコンテキストが十分に含まれている238件を大規模言語モデルを活用した質問回答システムによる回答文の分析のために抽出した。

表1 ランダムに50問サンプルした質問のタイプ分け。一つの質問が複数のタイプに分類される場合もある。質問文の例は、著者が実際に得られた質問文の構成を参考に作成した文である。

タイプ	例
講義で扱った内容のさらなる深掘り (60%)	Q: 大規模言語モデルにおけるバイアスの問題はなぜ起こるのですか？カラクリをあまり理解できませんでした。 / Q: In-Context Learning を行う能力はどのように学習されるのでしょうか？
解釈・理解の確認 (38%)	Q: Prompt Tuning の話を聞いて、文脈内学習と似たようなことをやっているように思えたのですが、モデルの重みを更新する以外の違いは何かありますか？ / Q: Transformer における QKV のうちの Query と呼ばれる部分は、質問文の Query とは性質が違うのでしょうか。
インタビュー (8%)	Q: 大規模な言語モデルの訓練と運用は膨大な計算資源を必要とします。これに伴う環境への影響やエネルギー消費についての懸念はありますか？持続可能な大規模モデルの開発に向けた現在の取り組みはどのようなものですか？ / Q: 大規模言語モデルの技術は今後どのように進化すると予想されますか？将来的に期待される主なイノベーションやブレイクスルーは何ですか？

表2 質問文単体では回答のためのコンテキストが足りない質問の例。質問文の例は、著者が実際に得られた質問文の構成を参考に作成した文である。

タイプ	例
スライドを基にした質問	Q: スライドの30枚目でモデルのパラメータ数とメモリについての説明頂きましたが、そこに記載されているモデルの占めるメモリ量は fp32/bf16 のどちらでしょうか？また、両者では計算速度はどう違うのでしょうか？ / Q: スライド12ページの図の横軸は何を指していますか？
口頭での説明の流れを基にした質問	Q: 動画のように？3次元畳み込みのようなことはできないのでしょうか？ / Q: 先ほどの話に関連して、文を短く要約する処理を組み込むことで長期の関係性を捉えられないのでしょうか？

## 4 データセットの分析

講義の内容に関する715件のうちランダムに50件をサンプルし、質問のタイプ分けを行った。質問には大きく分けて、「講義で扱った内容のさらなる深掘り」、学んだ内容に対しての「解釈・理解の確認」、何かしらのトピックに対して講師の考えを尋ねる「インタビュー」が存在した。それぞれのタイプの例は表1に挙げた。「講義で扱った内容のさらなる深掘り」タイプの質問に関しては、講義で紹介したトピックの元論文の詳細な説明部分やサーベイ論文を紹介することで受講生への回答が十分だと考えられるが、「解釈・理解の確認」に関しては、受講生がなぜそのような解釈を行うのか、挙げている複数の概念が同じ・あるいは違うと受講生が考えている理由は何か、それぞれのトピックを理解しているのかといった深掘りをする必要がある。

また質問文は、受講生が投影している資料や口頭での説明・前回までの講義での説明を受けて作成したものであるため、質問文単体だけでは回答のためのコンテキストが足りない場合も見受けられた。質問の例は表2に挙げた。

現状整備したデータセットは質問と回答のペア

のみであるが、より現実に即したデータセットの構築のために、投影しているスライド情報、講義で扱っている論文群の整備が必要であることが考えられる。

## 5 大規模言語モデルを活用した質問回答システムの検討

### 5.1 質問回答システム

大規模言語モデルを活用することで学習者の質問に答え学びを促進させることが期待できるかの検証のため、講師・TAの回答と大規模言語モデルを活用して得られた回答を比較し、質問に対して正しく答えられたのかを評価する。大規模言語モデルを活用した質問回答システムは2つ用意した。1つ目は”あなたは機械学習・自然言語処理・深層学習の専門家です。質問に対してわかりやすく回答してください。ただし与えられた質問が十分な情報を含んでおらず正確な回答ができない場合は、「質問に対して回答できません。」と回答してください。”という指示文と共に質問文を与え、大規模言語モデルに回答を出力させる方法である。

2つ目は本論文で分析を行った「講義で扱った内容のさらなる深掘り」、学んだ内容に対しての「解

積・理解の確認」への質問へ対処できるような工夫を施したシステムである。「講義で扱った内容のさらなる深掘り」に対して、質問文中にて言及されるトピックについての詳細な情報を提供することが必要であり、「解釈・理解の確認」への質問に対しては質問文中にて言及されるトピックについての詳細な情報を踏まえた上での回答を行うことで受講生の理解が正しいかどうか回答する必要があると考えた。どちらの質問に対しても質問文中にて言及されるトピックについての詳細な記述を検索することが重要であるため Web 検索を取り入れた。作成したシステムは大規模言語モデルを活用し、質問文が与えられたら、その質問文へ答えるために調査するトピックを検索できるような検索クエリを出力し、Google 社の Custom Search API を使用し検索エンジンの検索結果から Top10 個の Web 記事の記述を踏まえた上で回答を生成させる Retrieval Augmented Generation(RAG)[13] のシステムである。プロンプトの例は付録 A に記載している。質問回答の 1 例を図 1 に示す。大規模言語モデルは両方のシステムとも OpenAI 社の GPT-4 モデル (gpt-4-1106-preview) を使用し seed 値を 1208, temperature 値を 0 に設定して他はデフォルトの設定で使用し実験は 2024 年 1 月 11 日に行った。

## 5.2 質問回答システムによって得られた回答文の分析

質問回答システムによる回答文の分析のために抽出した、質問文と講師・TA による回答のペアが揃っている 238 件のうちシステムによって回答が得られた 159 件に対して ROUGE-2 F1 値を算出した結果、大規模言語モデルに質問文を与えて出力させたシステムは 38.9, RAG を利用したシステムは 24.8 であった。質問文・回答文・大規模言語モデルの回答文・RAG を利用したシステムの回答文の文字数の長さの平均はそれぞれ 84, 90, 757, 1636 であった。159 件のうちランダムに抽出した 10 件に対して回答の正確性という軸で人手による評価を行った。システムによる回答文と講師・TA が作成した回答を照らし合わせ、講師・TA が作成した回答と矛盾していたり、事実と異なる記述が含まれる場合は講師・TA が作成した回答が好ましいとし、講師・TA が作成した回答で触れられていないトピックに関しても正しい記述が含まれている場合はシステムの回答が好ましいとする評価を行った。10 件中講師・TA が作成した回答より大規模言語モデルの回答文が好ま

しいと評価されたのは 9 件であり、講師・TA が作成した回答より RAG を利用したシステムの回答文が好ましいと評価されたのは 8 件であった。今回システムに使用したモデルは 2023 年 4 月までのデータを使用したと報告されており、そのため大規模言語モデルに関連する質問に対しても正確な回答が行われたと考えられる。システムより講師・TA が作成した回答の方が好ましいという評価がなされた回答については講師・TA が回答で触れていた側面が明確に触れられておらず、RAG システムでは記事から回答根拠が見つからず「回答は見つかりませんでした。」と回答されたものであった。

システムによる回答文を観察した結果、質問文中に十分なコンテキストが含まれていられないような質問に対して、正確な回答はやはり難しいが、質問文中に正式名称ではなく略語が書かれていたとしても、機械学習分野の特定の単語だと仮定し回答を行なっている例も見受けられた。回答文がもっともらしいことを記述しており機械学習研究者である評価者にとっても正誤の判断が難しい点も観測された。シンプルなプロンプティングによるシステムでは存在しない単語についてももっともらしく回答をしているが、検索結果を組み合わせた回答システムではそのような間違いをしない例も見受けられた。講師や TA の回答に比べ大規模言語を活用した回答システムの回答は詳細な記述がなされており、受講生にとっては丁寧な説明になっているように見受けられるが実際に受講生が回答を見てどのように評価するかは調査すべき点である。

## 6 おわりに

本論文では、教育現場での QA システム開発の一助となることを目指し、実際に行われた講義中において生徒から提出された質問を分析した結果を公開した。既存の質問回答データセットの質問とは異なり、教育現場で得られる質問には、質問文単体では回答のためのコンテキストが足りないものが多く、質問回答の際には受講生の理解度を推し量った上での回答、理解の手助けが必要であるという点を明らかにした。また大規模言語モデルを活用した質問回答システムの検討を行い、今後の開発において回答の正確性を担保するような枠組みや提示された情報から学習者がさらに理解を深めていけるような仕組みとその評価を行う必要性があることを指摘した。

## 参考文献

- [1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In **Advances in Neural Information Processing Systems**, Vol. 33, pp. 1877–1901. Curran Associates, Inc.
- [2] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In **Advances in Neural Information Processing Systems**, Vol. 35, pp. 27730–27744. Curran Associates, Inc.
- [3] Wensheng Gan, Zhenlian Qi, Jiayang Wu, and Jerry Chun-Wei Lin. Large language models in education: Vision and opportunities, 2023.
- [4] Anna Rogers, Matt Gardner, and Isabelle Augenstein. QA Dataset Explosion: A Taxonomy of NLP Resources for Question Answering and Reading Comprehension. **ACM Comput. Surv.**, Vol. 55, No. 10, 2023.
- [5] Fan, Jernite, Angela and, Perez, Yacine and, Grangier, Ethan and, Weston, David and, Auli, Jason and, and Michael. ELI5: Long Form Question Answering. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 3558–3567, Florence, Italy. Association for Computational Linguistics.
- [6] Kwiatkowski, Palomaki, Tom and, Redfield, Jennimaria and, Collins, Olivia and, Parikh, Michael and, Alberti, Ankur and, Epstein, Chris and, Polosukhin, Danielle and, Devlin, Illia and, Lee, Jacob and, Toutanova, Kenton and, Jones, Kristina and, Kelcey, Llion and, Chang, Matthew and, Dai, Ming-Wei and, Uszkoreit, Andrew M. and, Le, Jakob and, Petrov, Quoc and, and Slav. Natural Questions: A Benchmark for Question Answering Research. **Transactions of the Association for Computational Linguistics**, Vol. 7, pp. 452–466, 2019.
- [7] Min, Michael, Sewon and, Hajishirzi, Julian and, Zettlemoyer, Hannaneh and, and Luke. AmbigQA: Answering Ambiguous Open-domain Questions. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 5783–5797, Online. Association for Computational Linguistics.
- [8] Lin and Chin-Yew. ROUGE: A Package for Automatic Evaluation of Summaries. In **Text Summarization Branches Out**, pp. 74–81, Barcelona, Spain. Association for Computational Linguistics.
- [9] Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. MAUVE: Measuring the Gap Between Neural Text and Human Text using Divergence Frontiers. In **Advances in Neural Information Processing Systems**, Vol. 34, pp. 4816–4828. Curran Associates, Inc.
- [10] Krishna, Roy, Kalpesh and, Iyyer, Aurko and, and Mohit. Hurdles to Progress in Long-form Question Answering. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 4940–4957, Online. Association for Computational Linguistics.
- [11] Min, Krishna, Sewon and, Lyu, Kalpesh and, Lewis, Xinxin and, Yih, Mike and, Koh, Wen-tau and, Iyyer, Pang and, Zettlemoyer, Mohit and, Hajishirzi, Luke and, and Hannaneh. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 12076–12100, Singapore. Association for Computational Linguistics.
- [12] Gao, Yen, Tianyu and, Yu, Howard and, Chen, Jiatong and, and Danqi. Enabling Large Language Models to Generate Text with Citations. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 6465–6488, Singapore. Association for Computational Linguistics.
- [13] 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 **Advances in Neural Information Processing Systems**, Vol. 33, pp. 9459–9474. Curran Associates, Inc.

## A RAG で使用したプロンプト

質問文が与えられてから回答文を出力するまでの流れは以下の通りである。

1. 質問文から検索エンジン API へ入力するための検索クエリを生成する (大規模言語モデルの使用, 1 つデモンストレーション例を与えた one-shot 生成で行う。)

2. 検索エンジン API から検索結果を取得し, 各 Web 記事の記述を取得する

3. 各 Web 記事の記述とともに質問文を追加し, 記述を踏まえた文を生成する (大規模言語モデルの使用)

4. 生成文に各 Web 記事のタイトルと URL を追記し回答文として出力

質問文から検索エンジン API へ入力するための検索クエリを生成する際に大規模言語モデルへ与えたプロンプト, 各 Web 記事の記述とともに質問文を追加し, 記述を踏まえた文を生成する際に大規模言語モデルへ与えたプロンプトを以下で示す。

### A.1 検索クエリを出力するためのプロンプト

Given a question, please provide search query to find the answer from Google search results. In order to find the relevant information, please break down the question and provide the search query for each part. Return lists of search queries in following format. Search Query: [bulleted list of search queries]

図 2 質問文から検索エンジン API へ入力するための検索クエリを生成する際に大規模言語モデルへ与えたプロンプト。

### A.2 検索結果を踏まえた上で回答を出力するためのプロンプト

Use the provided articles delimited by triple quotes to answer questions. If the answer cannot be found in the articles, write "回答は見つかりませんでした。" Please answer in Japanese.  
-Excerpted from [article title] (URL: [URL]) """"[article paragraph]"""" ...(Web 検索結果のセクション続)  
Question: [質問文]

図 3 各 Web 記事の記述とともに質問文を追加し, 記述を踏まえた文を生成する際に大規模言語モデルへ与えたプロンプト。 [article title] には検索結果の Web 記事のタイトル, [URL] には記事の URL, [article paragraph] には記事の記述, [質問文] には質問文がそれぞれ入る。