

Question Generation for English Reading Comprehension Exercises using Transformers

Alexander Maas ^{*}, Kazunori D Yamada ^{*}, Toru Nagahama ^{*},
Taku Kawada ^{*}, Tatsuya Horita ^{*}

Abstract

In secondary language education, one tool used by teachers to test students' language ability is reading comprehension. The construction of these problems can take a lot of time as the text needs to contain only the vocabulary and grammar the students know, and the questions also need to test the reasoning skills the teachers want to evaluate. To allow educators to use reading comprehension exercises more frequently, this research aims to alleviate the time constraint of creating these questions by training a controllable transformer-based natural language processing model to create questions of varying types and about a passage of text as specified by the user. After fine-tuning, the questions generated using the new controls either suffered from overfitting or from a lack of diversity between them, however the output of an existing question generation control was altered and became capable of generating questions suitable for use in reading comprehension. To improve the output of the new controls, more data could be used in the training, or an alternative training scheme would need to be utilized.

Keywords: Artificial Intelligence, English Language Education, Natural Language Processing, Reading Comprehension Exercises

1 Introduction

When learning a new language there are four main skills which curriculum are based on: reading, writing, speaking, and listening. Whereas each require their own approach to be taught effectively, there is flexibility available to educators in how they choose to teach these skills to students. One method to test students' understanding of new vocabulary and grammatical concepts is via reading comprehension. Such problems require students to demonstrate not only that they understand the individual vocabulary and grammar used in the problem but also the meaning behind the constructed sentences. Full understanding of the written text further requires students to combine all of these issues and to construct a meaningful whole [1]. It is this constructed meaningful whole that is tested by educators using reading comprehension problems.

As alluded to above, the comprehension of the written word is not attributed to a single skill, such as improving students' vocabularies, and as such there are various reasons why students may struggle to answer such questions. Even if each of the necessary skills for reading comprehension were improved individually, without sufficient practice using them all simultaneously, the students would still struggle to answer reading comprehension questions. Therefore, in

^{*} Tohoku University, Sendai, Japan

addition to improving specific skills, students should continually practice reading comprehension exercises if the teachers wish for the students' reading comprehension ability to improve [2]. However, creating reading comprehension exercises is a very time-consuming process which most educators can only manage when creating examinations due to their already heavy workload. For Japanese teachers, who have many additional duties which often require their attention [3], there is not enough time in the day to assign to this additional exercise creation.

Therefore, to allow for more liberal use of reading comprehension questions, the time constraint on their construction needs to be addressed. With the development of natural language processing (NLP) models which utilise the transformer architecture [4], and their more recent iterations, such as ChatGPT developed by OpenAI [5], it has become possible to automate the construction of these questions.

To allow educators to control the types of questions that will be generated, first the different types of reading comprehension questions need to be identified and the data labelled for use in the controlled generation model. This newly labelled data is then used to fine-tune a controllable, transformer-based generation model capable of outputting questions of the desired category based on the text from a reading comprehension problem. This model could be used with any task where question and answer creation is required, such as for exam creation or allowing students to practice to their heart's content to improve their abilities.

The rest of the article is organised as follows. In Section 2, a brief introduction to Natural Language Processing and ChatGPT will be given along with the impact ChatGPT has had on the field of generative text AI models. Section 3 will introduce the dataset used, explain why it was chosen over other potential datasets and then go into detail explaining how the questions were clustered. Section 4 will give a detailed description of the controllable generation architecture and fine-tuning process as well as the results of this process. Section 5 will discuss the problems encountered with the generation process and possible ways to address the problems found. The paper will conclude with Section 6 and discuss future research opportunities and practical implementation of the developed system.

2 Natural Language Processing and ChatGPT

There is a long history behind Natural Language Processing (NLP) techniques, of which the early methods aimed to hard-code language rules to allow machines to translate between languages or generate text, however the statistical models upon which recent methods were built upon started with the Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) architectures developed in the 1980s and 1990s. The field of statistical NLP has continued to advance since then and has recently culminated with the development and release of ChatGPT, based on GPT-3.5, by OpenAI in November 2022 [5] and a premium version based on GPT-4 released in March 2023 [6]. Reference [7] provides a good analysis of both versions of ChatGPT across various tasks and compares the results obtained with the results of the state-of-the-art (SOTA) models for each of those tasks. The authors found that, of the 25 tasks which they tested, ChatGPT underperformed the SOTA models which already had been trained for the specific tasks. Whereas ChatGPT performed well on most of the tasks, its performance was much worse on tasks that were more difficult when compared to the models which had been trained specifically for that task. However, the ability of ChatGPT to generate text, specifically its ability to create reading comprehension exercises for English language learners as this paper is attempting

to do, was not evaluated by the authors. To enable comparison with this paper's model's output, some examples from ChatGPT will be generated and discussed in relation to this paper's model's output in the results section below.

3 Clustering on Latent Traits

In order to successfully fine-tune a generation model capable of creating quality English language reading comprehension questions, training data which is relevant to the subject matter and of a good quality is required. Additionally, in order to control the traits of the generated questions, the dataset must be labelled with those traits to ensure the model can learn the differences between the different types of questions. To this end the dataset and methodology outlined in [8] will be used as a baseline upon which improvements and changes will be implemented.

In the paper [8], the authors used the RACE dataset [9], which contains approximately 28,000 passages of text and 100,000 questions created by English language instructors for examinations in Chinese Junior and Senior High Schools, to train their data labelling AI model. After examining other datasets containing question and answer couplets based on a passage of text [10][11][12], it was decided that the RACE dataset, with its focus on English language reading comprehension exercises authored by actual instructors, was also the most suitable dataset to use for the generation model presented in this paper.

Given this paper's stated purpose of controllable question generation for reading comprehension exercises and the RACE dataset's unlabelled structure, the methodology outlined in [8] to label the dataset will be used. However, several problems were found with the clustering methodology which first needed to be addressed. The first of these was the undue influence the text had on the clustering results; different questions were often clustered together just because they shared the same input passage of text. To rectify this, the model used to find the latent variables was re-trained with the training data changed to have only the question as the primary input with the passage of text and answer included in the decoder's input. This allowed the passage and answer's information to be transmitted backwards via training while keeping the questions information as the primary source of information for constructing the latent variables. The second of the problems involves the inability of a trained t-SNE model to convert new data using the same parameters as was used to cluster the original data [13]. The lack of this ability would make growing the dataset in the future more difficult as the addition of new data would likely change the clusters which would be found due to the changed density and layout of the results. To assign new data to previously found clusters, the t-SNE method was changed to the UMAP method which is capable of converting new data to a previously trained model [14].

The result of these changes on the clustering output can be seen in Figure 1. The figure shows 3 clusters were found by the HDBSCAN algorithm [15] however a quick visual inspection shows some of these are the result of several smaller clusters joined by sufficiently dense groups of questions between them. After sampling several of these smaller groups, the traits of the questions contained within each were identified. The smaller clusters at the bottom and right of the figure both ask the reader to choose the most likely heading for the associated passage of text from the options given. Despite their similar objectives, several differences allow the model to differentiate between them well: the majority of the right-side cluster's questions (~98%) are What or Which questions that end in a "?" whereas the bottom cluster take the form of

cloze-style questions (sentence completion). Analysis of the left and right sides of the top cluster also shows a tendency toward separation based on physical traits of the questions. The right side of the cluster is composed predominately of WH-style questions ($\sim 65\%$) compared to $\sim 5\%$ in the left side of the cluster. Additionally, cloze questions account for $\sim 96\%$ of the left side of the cluster while only accounting for $\sim 3.5\%$ of the right side. Whereas the physical traits of the questions appear to be responsible for the separation into distinct clusters, how the questions are organised within each cluster and why the questions connecting the 2 sides of the upper cluster were placed in between the 2 sides remains unknown and requires further investigation.

Whereas the clusters found from the latent variable analysis appear to favour rudimentary and simple separation, these can still be used to label the questions for use in the controlled generation training. To label the questions, 3 clusters will be used to broadly represent different types of questions which educators could ask their students. The right and bottom clusters are combined to represent questions which require students to understand the whole passage to choose an appropriate title for the text, the left side of the top cluster will be used to represent cloze-style questions and the right side for general purpose questions.

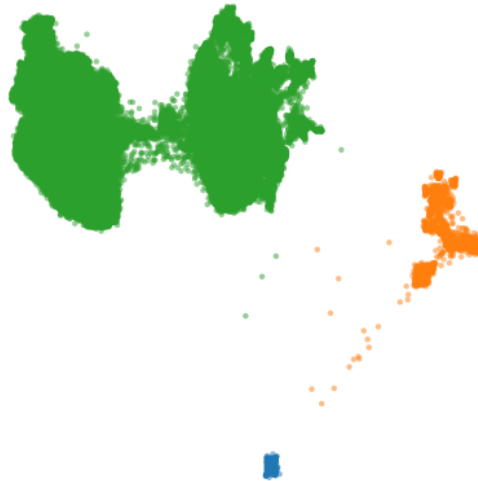


Figure 1: Latent variables of the RACE dataset questions after PCA and UMAP dimension reduction and HDBSCAN clustering showing 3 primary clusters. The x- and y-axis have no interpretable meaning due to the non-linear nature of the UMAP algorithm. The visualisation is just to identify similarity between the data points.

4 Controlled Question Generation

4.1 Background of Existing Text Generation Architectures

This section will begin with a discussion of a review [16] which covered advances in automatic question generation techniques for educational purposes between 2015 and 2019. During this period, machine learning and AI approaches were almost non-existent. Additionally, there is little existing literature on generating questions of controlled difficulty or possessing desired traits with the existing papers relying on expert input of rules for what constitutes difficulty thus limiting generalizability of the approaches used.

Since the review was conducted, AI and more specifically large language models have been getting increased attention from academics and the public as demonstrated by the unveiling of ChatGPT [5]. Below, several papers will be introduced that have attempted to use AI methods for automatic question generation. These papers are not focused on generation for educational purposes; however, they still provide a useful starting point for AI focused methodologies.

The first of these papers which will be discussed [17] developed a model that can generate both questions and answers from a multi-paragraph document using the Bidirectional Encoder Representations from Transformers (BERT) architecture. Whereas the model can process multi-paragraph content, the questions and answers are drawn from a single sentence. In addition, some of the questions and answers either make no sense or are very basic. Another consideration is the lack of control the user has over the type of questions generated. From an education standpoint it would not be unusual to see a question like "Where could this passage be found?" being asked about the text. The model can generate a variety of questions regarding the text but does not possess the ability to allow the user to specify the types of questions they want to be generated. Given how their focus is not on language education this detailed control of the generation process is not unexpected.

The next paper [18] utilises a Graph Convolutional Network (GCN) with a Bidirectional Long Short-Term Memory (Bi-LSTM) component. Unlike the model introduced in the previous paragraph, this approach is better able to account for information in different sentences/paragraphs when generating questions. However, this model requires as input an answer in addition to the text passage in order to generate the questions. This requirement limits the usefulness of the model to educators as if they already had an answer regarding the passage, they would already have an idea of what type of question it came from. This negates the time saved by an automatic question generation model which is a primary benefit to educators. If the model could be altered to also generate answers from multi-paragraph text the usefulness of this approach would be greatly increased.

The final paper to be introduced [19], have trained a large language model, called CTRL, which is capable of controlled text generation. During training, they do this by conditioning the language model on a control code so that the model learns the distribution $P(x|c)$ where c is the control code and x is a prompt consisting of a sequence of symbols (tokens) created by encoding the text using Byte Pair Encoding (BPE). By training their model on a large corpus of data, 140GB of text data taken from a variety of different sources, it is capable of generating unique text which follows the style associated with that control code. In the examples shown in the paper, the authors demonstrate how the same input prompt can produce very different output based on the control code chosen for the generation. However, whereas the authors have implemented a control code to generate questions and answers, the results are closer to trivia questions than those one would find in a reading comprehension context. See Table 1 for an example using text taken from a Japanese Junior High School's 3rd grade English textbook [20]. As can be seen from Table 1, the current CTRL model interprets the prompt as a question to be answered and, despite the length of the prompt, will generate questions following the theme of the prompt but not the actual content. For example, the word "yurt" and the country Guatemala are not mentioned in the prompt, however because of the data the model was trained on, the model considers these topics semantically similar and therefore candidates for generation. This behaviour limits the model's usefulness as an educational tool since educators would only want to ask questions to students that they could reasonably infer the answer of from the given text.

Table 1: Output of the CTRL model [19] using the “Questions” control code with a longer prompt to determine effectiveness at generating reading comprehension questions

Prompt: “This is a book that shows houses in Asia. This is a ger, a traditional house in Mongolia. Some people in Mongolia move their gers with the seasons. This way of life is interesting. So I’d like to visit Mongolia and stay in one.”

| | Question | Answer |
|---|---|------------------|
| 1 | - | yurts |
| 2 | The first president of this country was named after this man who founded it in 1911 | Guatemala |
| 3 | In his autobiography, he wrote about how he had been an alcoholic for years before becoming sober & writing “The Autobiography of Malcom X” | Martin Luther... |

4.2 Methodology

The model chosen for the controlled question generation task is the CTRL model [19] which can not only generate output based on a specified input but also control the form of the output via the inclusion of control codes. The authors have made the pretrained model publicly available on their GitHub page (<https://github.com/salesforce/ctrl>) in addition to instructions and code describing how to prepare user data and use it to fine-tune the model to add additional control codes. The control codes "fill", "normal" and "title" were selected to represent the 3 clusters of question types as found in Section 3: with "fill" being used for the cluster composed predominately composed of cloze-style or sentence completion questions, "normal" for the cluster composed predominately of WH-type questions and "title" for the remaining questions which ask for the best title or headline of the input text. To prevent overfitting on small datasets, the authors recommend keeping the number of iterations low during fine-tuning. To this end, 2 different models were trained: 1 with the iterations set to 250 as directed by the authors and another with iterations set to 25,000.

4.3 Results

Sample output from both models based on the same prompt as used in Figure 1 is shown in Table 2. Generated questions from the models trained for 250 and 25,000 iterations are displayed at the top and bottom respectively on the pre-existing "Questions" control code and the newly created "fill", "normal" and "title" control codes.

Table 2: Generated Questions

Prompt: "This is a book that shows houses in Asia. This is a ger, a traditional house in Mongolia. Some people in Mongolia move their gers with the seasons. This way of life is interesting. So I'd like to visit Mongolia and stay in one."

| Code | Question | Answer |
|--------------------------|--|---|
| Questions 250 iter | The writer wrote this passage mainly for _ . According to the author, we should not judge others by our own standards because _ . | students they may be different from us |
| fill 250 iter | Which of the following may be true about the author? What can we know from this passage about the author's attitude towards life? | He was born on April Fool's Day. Optimistic |
| normal 250 iter | Which of the following may be true about the author's life? What can we know from this passage? | She was born on April Fool's Day. The author had been living in Beijing for some time. |
| title 250 iter | Which of the following may be true about the author's life? What can we know from this passage? | She was born on April Fool's Day. The author has been living abroad for some time. |
| Questions 25,000 iter | According to this passage, "I" am not an American but _ . It is possible for us to learn English from each other because we _ . | I love America very much have something similar in culture |
| fill 25,000 iter | According to this passage, if you want to call your friend for help but he or she doesn't respond, you'd better _ . The author mentioned "the ostrich approach" (Para. 5) because it describes ostriches as _ . | hang up the phone stupid |
| normal 25,000 iter | According to Jolle Jolles, how did she get her first house? What do you know about Mr. Smith? | She was working at an American hospital. He can't remember his wife's name. |
| title 25,000 iter | According to this passage, which one do you think is right? In Leo's opinion, what should he do? | Ted wants his dad to buy him some video games. Talk to his parents about this. |

Several revealing conclusions can be drawn from these results. Firstly, analysing the generated output for the new control codes, the 250 iterations model shows that the questions are not only similar between the different control codes but also create questions and answers whose information is not present in the provided prompt. The inability of the model to capture the traits of the data in each control code category could be attributed to a lack of training iterations pre-

venting the model from distinguishing between them. However, the same control codes, when fine-tuned for 25,000 iterations only fix some of these problems. Each code is now generating questions closer to the type contained in their respective data, but the content of the questions and answers are even further removed from the provided prompt often mentioning information which is clearly very different from the prompt, for example the question "The author mentioned "the ostrich approach" (Para. 5) because it describes ostriches as _." is clearly taken from one of the training data points thus indicating overfitting is occurring.

The pre-existing control code "Questions", however, produces some interesting results. Whereas the original output of the CTRL model appeared to follow the logic of trivia questions the fine-tuned model is causing the code to create questions closer to those found in reading comprehension problems despite the "Questions" code not receiving any direct fine-tuning itself. This contrasts with the other pre-existing control codes, such as "Wikipedia," which continue to produce output in the same manner as before the fine-tuning occurred. For the "Questions" control code, whereas the output of the 25,000-iteration model also appears to be suffering from overfitting, the 250-iteration model is producing good reading comprehension questions which could reasonably be expected to accompany the associated prompt. To confirm this is not just a coincidence, additional questions were generated using the 250-iteration model with the "Questions" control code a prompt taken from a prior 3rd grade Eiken examination [21], with minor modifications to ensure UTF-8 compatibility. The additional results can be seen in Table 3.

Table 3: Additional Generated Questions

Prompt: "In 1909, Lilian's uncle sent her postcards. One of them had a picture of Louis Bleriot on it. Bleriot was a pilot, and he built his own plane. He was the first person to fly a plane across the English Channel. His plane had an accident when it landed, but Bleriot wasn't hurt. His story soon became famous."

| | Question | Answer |
|---|---|--|
| 1 | The writer is probably referring to _. | the history of air travel |
| 2 | According to the author, we can infer that _. | people are more likely to be killed in accidents if they have been drinking too much alcohol |
| 3 | We know from the passage that _. | there were no fatalities caused by the crash because all passengers survived except one who died later |

These examples provide some additional information about the generation capabilities of the model. Firstly, question 1-1 is the only question to have a strong connection with the content of the prompt, correctly determining that the prompt is about the history of air travel. This likely arose from the year mentioned at the start of the prompt, 1909, and the various words associated with aviation, plane, pilot, landed and Louis Bleriot. However, the remaining questions in example 1 have created their own content while generating the questions. No mention of alcohol is in the prompt yet question 1-2 drew upon the presence of the word "accident" in the prompt to generate the question. Similarly, question 1-3 likely drew from the phrase "wasn't hurt" and the aviation theme to generate the question mentioning the lack of casualties from the flight. However, it fails to correctly link this with the prompt as Bleriot was the only person onboard the plane, no other passengers were mentioned in the prompt, and no mention of anyone dying was mentioned yet was included in the generated response.

These results show that relevant reading comprehension questions can be generated however the output's connection with the content of the prompt is often lacking or the prompt's meaning is misinterpreted by the model resulting in output that resembles reading comprehension problems but with inaccuracies. This may be due to the training data used by the CTRL authors [19]. The multiple sources, including 45 subreddit forums, possess very different themes which may be inadvertently appearing in the output. Finally, given the iterative generation process, it is likely that questions generated later in the sequence are influenced by the earlier questions thereby causing any errors or misinterpretations to compound and the theme set by those earlier questions to be continued, as evidenced by questions 1-2/1-3's focus on death.

Table 4: ChatGPT generated questions and answers

(1) *For the following prompt, please create some reading comprehension questions for students. "In 1909, Lilian's uncle sent her postcards. One of them had a picture of Louis Bleriot on it. Bleriot was a pilot, and he built his own plane. He was the first person to fly a plane across the English Channel. His plane had an accident when it landed, but Bleriot wasn't hurt. His story soon became famous."*

(2) *Can you also provide the answers for those questions?*

| | Question | Answer |
|---|---|---|
| 1 | Who sent postcards to Lilian in 1909? | Lilian's uncle sent her postcards in 1909. |
| 2 | What was depicted on one of the postcards Lilian received? | One of the postcards had a picture of Louis Bleriot on it. |
| 3 | Who was Louis Bleriot, and what was his notable achievement? | Louis Bleriot was a pilot who built his own plane. |
| 4 | What did Louis Bleriot do that made him famous? | Louis Bleriot was the first person to fly a plane across the English Channel. |
| 5 | Can you recall the outcome of the accident that happened to Bleriot's plane when it landed? | The passage does not provide details about the outcome of the accident when Bleriot's plane landed. |
| 6 | How did Bleriot escape from the accident with his plane unharmed? | The passage mentions that Bleriot wasn't hurt in the accident that happened when his plane landed. |
| 7 | What do you think might have been the significance of Bleriot's story becoming famous? | The passage does not explicitly state the significance of Bleriot's story becoming famous, but one can infer that it was likely due to his pioneering achievement in flying across the English Channel. |

Some results of ChatGPT's [5] generation are displayed in Table 4. To best approximate a reading comprehension question and answer structure, a specific and detailed prompt was used as shown in the table. Since the end users of this paper's generation model are intended to be English language teachers, knowledge of how to best phrase prompts to obtain the most desired outputs would be a skill that most educators would be lacking without further training. For this reason, the results a teacher could obtain may be a lower quality or of less relevance than those presented here. When comparing the generated questions and answers with those of this paper's model presented earlier, several patterns reveal themselves. The first is the vocabulary used in the questions, with the ChatGPT questions being of a higher general level (for example, de-

pected) when compared to this paper's model (for example, referring). This could pose a problem to students who are still unfamiliar with the words being used but could be fixed by teachers manually if necessary. As long as the generated questions and answers are simple, ChatGPT does well but struggles when composing questions on more difficult concepts. This is in-line with the findings of previous findings that ChatGPT struggles with more complex tasks [7]. Additionally, ChatGPT's question and answer 5 are not answerable with only the given prompt. This is similar to this paper's model which can create questions and answers which are not relevant to the prompt. The answers provided by ChatGPT are also longer on average and constantly refer back to the passage to justify itself, a trait not seen in normal reading comprehension questions. However, despite all these problems, ChatGPT does, on average, produce questions and answers with a more direct link to the prompt than this paper's model can consistently output. This indicates that more fine-tuning or alternate approaches are required for this paper's model to consistently reference the provided prompt in its generated questions and answers.

5 Discussion

The fine-tuned CTRL model is simpler for non-experts to use than other alternatives due to only needing the passage of text, on which it will generate questions, and the selection of the type of question desired. Given the ability of the model to read a passage of text in approximately 1 second and to create a question-and-answer pair in a fraction of a second, the model offers the potential for much faster creation of reading comprehension learning materials. To improve the model, the results described in the previous section provide a guide as to where future efforts would be most beneficially focused. The fact that the model struggles with associating the prompt with the generated questions has several possible solutions. Since the original CTRL model [19] was focused on very short prompts with control codes that conveyed the style of the generated text and not the specific content, the inability of fine-tuning to overcome this limitation implies that architectural changes or the fine-tuning approach may need to be reconsidered to obtain better results. One way to accomplish this could be by conducting fine-tuning not only on the control code and question/answer text but also by incorporating the prompt into training efforts. Whereas this may also suffer from overfitting, it could alleviate the disconnect between the prompts and generated texts observed in the outputs. Another possible approach to address this problem is to include the theory behind the controlled generation approach shown used in the CTRL model [19] with the work of another author whose model could associate their generation with the input prompt [17][18] or an even newer approach utilising the generational capabilities of ChatGPT [5]. This would allow the full prompt to be utilised by the generation model while also allowing for some level of controllability.

Finally, the dataset used for training could be increased by incorporating other question answering datasets into the training data. Whereas the RACE dataset's focus on actual reading comprehension problems is highly relevant to the task, its size is still too small for fine-tuning large language models, as evidenced by the overfitting results shown in Table 2. Whereas other datasets may not have this focus on reading comprehension, it may still be possible to incorporate them into the fine-tuning of the model to allow for more iterations to be run before overfitting occurs. Datasets like NewsQA [11], despite the problems discussed in Section 3, are potential candidates to increase the volume of training data.

6 Conclusion

The automatic generation of questions and answers with desired traits is a research topic that has received little focus partly due to the onerous expert knowledge requirements. However, the ability of machine learning methods to produce state-of-the-art results in natural language processing tasks has made this a possibility. The results of the clustering on latent traits lends support to the idea that a machine learning method can differentiate between different traits of questions allowing for datasets to be labelled for use in controlled generation models, as was done for the RACE dataset. However, the controlled generation task using the CTRL model, while promising, encountered several problems which need further research to overcome. These being the limited size of datasets for reading comprehension problems causing overfitting at medium to high training iterations and the lack of connection between the provided prompt and the generated output. Future research will continue to search for solutions to these problems in order to create a tool of use to English teachers allowing them to focus their efforts more on addressing problems encountered by students instead of on exam creation.

References

- [1] J. Oakhill, K. Cain, and C. Elbro, *Understanding and teaching reading comprehension: A handbook*, Routledge, 2014, pp. 15.
- [2] K. Aizawa, “Relationship between vocabulary coverage and comprehension in reading comprehension [読解における語彙カバー率と理解度の関係],” *Educational materials research* 22, 2011, pp. 22.
- [3] T. Kanbayashi, “The influence of peripheral duties on the sense of busyness and burden of public elementary and junior high school teachers: Focusing on the workload per unit of time [周辺の職務が公立小・中学校教諭の多忙感・負担感に与える影響: 単位時間あたりの労働負荷に着目して],” *Bulletin of the Japanese Society of Educational Management* 57, 2015, pp.79.
- [4] A. Vaswani et al., “Attention is all you need,” *Advances in neural information processing systems* 30 (NIPS 2017), 2017, pp. 5998.
- [5] OpenAI, “Chatgpt: Optimizing language models for dialogue.” OpenAI.
<https://openai.com/blog/chatgpt/> (Accessed Feb. 7, 2023).
- [6] OpenAI, “GPT-4.” OpenAI. <https://openai.com/research/gpt-4> (Accessed Oct. 24, 2023).
- [7] J. Kocoń et al. "ChatGPT: Jack of all trades, master of none." *Information Fusion* (2023): 101861.
- [8] A. Maas, T. Kawada, K. Yamada, T. Nagahama, and T. Horita, “Identifying Latent Traits of Questions for Controllable Machine Generation,” *EdMedia+ Innovate Learning (EDIL 22)*, 2022, pp. 42.
- [9] G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy, “Race: Large-scale reading comprehension

- dataset from examinations,” Proc. of the 2017 Conf. on Empirical Methods in Natural Language Processing, 2017, pp. 785.
- [10] M. Richardson, C.J. Burges, and E. Renshaw, “Mctest: A challenge dataset for the open-domain machine comprehension of text,” Proc. of the 2013 Conf. on Empirical Methods in Natural Language Processing, 2013, pp. 193.
- [11] A. Trischler et al., “Newsqa: A machine comprehension dataset,” Proc. of the 2nd Workshop on Representation Learning for NLP, 2017, pp.191.
- [12] K.M. Hermann, et al., “Teaching machines to read and comprehend,” Advances in neural information processing systems 28 (NIPS 2015), 2015, pp. 1693.
- [13] L. Van Der Maaten, “t-SNE – FAQ.” Laurens van der Maaten Github.io.
<https://lvdmaaten.github.io/tsne/> (Accessed Oct. 18, 2022).
- [14] L. McInnes, “Transforming New Data with UMAP.” UMAP.
<https://umap-learn.readthedocs.io/en/latest/transform.html> (Accessed Sep. 18, 2022)
- [15] L. McInnes, J. Healy, and S. Astels, “hdbscan: Hierarchical density based clustering,” Journal of Open Source Software, vol. 2, no. 11, 2017, pp. 205.
- [16] G. Kurdi, J. Leo, B. Parsia, U. Sattler, and S. Al-Emari, “A systematic review of automatic question generation for educational purposes,” International Journal of Artificial Intelligence in Education, vol. 30, 2020, pp. 121.
- [17] M. Roemmele, D. Sidhpura, S. DeNeefe, and L. Tsou, “AnswerQuest: A system for generating question-answer items from multi-paragraph documents,” Proc. of the 16th Conf. of the European Chapter of the Association for Computational Linguistics: System Demonstrations, 2021, pp. 40.
- [18] D. Su, et al., “Multi-hop question generation with graph convolutional network,” Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 4636.
- [19] N.S. Keskar, B. McCann, L.R. Varshney, C. Xiong, and R. Socher, “Ctrl: A conditional transformer language model for controllable generation,” arXiv preprint:1909.05858, 2019.
- [20] M. Negishi et al. “New Crown English Series New Edition 3,” Sanseido, 2015, pp. 54.
- [21] Eiken Foundation of Japan, “Eiken test in practical english prociency, grade 3.” EIKEN.
https://www.eiken.or.jp/eiken/exam/grade_3/pdf/202203/2022-3-1ji-3kyu.pdf (Accessed Mar. 14, 2023).