LLM exam generation
Project Proposal for NLP Course, Winter 2024

Nikita Kozlov
nikita.kozlov.stud@pw.edu.pl

Zofia Łągiewka
zofia.lagiewka.stud@pw.edu.pl

Jakub Świstak
jakub.swistak.stud@pw.edu.pl

Jacek Zalewski
jacek.zalewski.stud@pw.edu.pl

Supervisor: Anna Wróblewska
Warsaw University of Technology
anna.wroblewska1@pw.edu.pl

Abstract

This report focuses on the preliminary review of existing solutions and datasets relevant to automated question-answer pair generation, a critical area in educational technology for exam creation and personalized learning. We examine current methodologies, including traditional rule-based systems and state-of-the-art approaches utilizing Large Language Models (LLMs) such as GPT, T5, and BERT. Particular attention is given to the types of questions these models generate, such as multiple-choice and fact-based queries, and the limitations they encounter in addressing more complex question types. Additionally, we explore commonly used datasets like SQuAD and NewsQA, analyzing their strengths and constraints for this domain. By consolidating insights from existing solutions and datasets, this report establishes a foundation for developing an advanced framework in subsequent phases of the project, aiming to generate diverse and contextually accurate question-answer pairs.

Report Outline

This report is organized as follows:

- **Section 1**: Introduction
- **Section 2**: Literature review of state-of-the-art methods and relevant datasets.
- **Section 3**: Review of relevant datasets.
- **Section 4**: Solution concept and proposal.
- **Section 5**: Evaluation of different available LLMs

By exploring these objectives, we aim to develop a reliable, efficient, and scalable system for automated exam generation.

1 Introduction

Automated question-answer pair generation plays a vital role in educational technology, enabling efficient exam creation and personalized learning. Traditional methods often rely on manual input or rule-based algorithms, which are time-consuming and not scalable. Furthermore, question generation can be used to inspire examiners to come up with new kinds of questions. This project explores the use of Large Language Models to generate diverse question-answer pairs from a given context. Our approach aims to ensure that the generated questions are both semantically meaningful and contextually accurate. Recent advancements in NLP, especially transformer-based models like GPT, T5, and BERT, have made significant progress in tasks such as text generation and question answering. Several approaches exist for question generation, many focus on multiple-choice questions (MCQs) or fact-based queries. Datasets such as SQuAD and NewsQA are commonly used for training, but they are often limited to simpler question types, which may not fully capture the complexity required for exam generation.

2 Literature review of state-of-the-art methods and relevant datasets

The automatic generation of question-answer pairs is a critical task in natural language processing (NLP), with applications in education, exam preparation, and interactive systems. Recent advancements in transformer-based models have significantly improved performance in question generation (QG) and question-answering (QA) tasks. This section reviews state-of-the-art approaches relevant to LLM-based exam generation.

2.1 Neural Question Generation (NQG): Zhou et al. (2017)

This paper pioneered the application of neural networks to automatic question generation (QG), introducing a framework that leverages the encoder-decoder architecture to produce natural language questions from text. Their approach differs significantly from traditional rule-based methods, which rely on rigid heuristics to transform sentences into questions. By integrating answer position indicators and lexical features—such as part-of-speech (POS) and named entity recognition (NER)-into the encoder, the model enhances the contextual understanding of text passages, enabling it to generate diverse, answer-focused questions.

The study utilized the SQuAD dataset, which provides manually annotated question-answer pairs derived from Wikipedia articles. Experimental results highlighted the ability of the model to produce fluent and semantically accurate questions. The authors emphasized the importance of incorporating linguistic features to improve sentence encoding and proposed the use of attention mechanisms to focus on the most relevant parts of the input text. This work represents a significant shift in the field, demonstrating the effectiveness of neural networks for QG without relying on pre-defined rules. Future research could expand this framework by exploring its potential for improving downstream QA systems.

2.2 Reinforcement Learning for Question Generation: Scialom et al. (2019)

Scialom et al. (Scialom et al., 2019) addressed a critical limitation in traditional QG evaluation by proposing unsupervised metrics based on QA performance. The authors noted that commonly used metrics like ROUGE and BLEU often fail to account for fluency, coherence, and semantic relevance, as they focus primarily on n-gram overlap with reference texts. To overcome these shortcomings, they introduced QA-based metrics that assess the alignment between questions and answers, eliminating the need for reference summaries. This approach enables self-supervised learning, significantly reducing the dependency on human-annotated datasets.

Their framework leverages reinforcement learning (RL) to optimize question-generation models for these QA-based metrics. Experimental results showed that RL-trained models outperform state-of-the-art systems in both automated and human evaluations. By focusing on semantic quality rather than lexical similarity, the proposed metrics better reflect the human perception of question relevance and coherence. This work lays the foundation for integrating RL into QG systems and highlights the broader applicability of QA-based metrics to other natural language generation (NLG) tasks, such as abstractive summarization.

2.3 Sequence-to-Sequence Learning for Reading Comprehension: Du et al. (2017)

Du et al. (Du et al., 2017) introduced an attention-based sequence-to-sequence model for question generation, framing the task as a direct mapping of sentences or paragraphs to corresponding questions. Unlike traditional methods, which rely heavily on hand-crafted transformation rules, their approach employs an end-to-end learning paradigm. This enables the model to generate questions that are not only grammatically correct but also contextually rich and semantically diverse.

The study demonstrated that incorporating paragraph-level context improves the quality of generated questions, as it allows the model to leverage additional information beyond individual sentences. Evaluations on the SQuAD dataset showed that the proposed model significantly outperformed rule-based baselines, particularly in generating complex and challenging questions. Human evaluations further confirmed that the generated questions were more natural, fluent, and difficult to answer compared to existing approaches. This work underscores the potential of sequence-to-sequence models for enhancing educational tools, particularly in generating diverse question types for reading comprehension assessments.

2.4 Dynamic Multitask Learning for Consecutive Question Generation (CQG): Yunji et al. (2022)

Yunji et al. (Yunji et al., 2022) expanded the scope of QG by introducing the concept of consecutive question generation (CQG), where the goal is to generate a series of logically connected question-answer pairs to comprehensively cover a given text passage. Unlike traditional methods that generate questions in isolation, their dynamic multitask framework incorporates auxiliary tasks such as rationale generation, context history tracking, and answer prediction. These tasks work together to ensure that the generated questions are accurate, informative, and logically consistent.

The authors proposed a novel self-reranking mechanism to optimize the sequence of question-answer pairs, ensuring coherence and coverage across the entire passage. Experimental results demonstrated that their approach outperformed existing QG methods in generating interconnected questions that align with the underlying context. The study also highlighted the potential of CQG for educational applications, such as tutoring systems and comprehensive reading comprehension tests. By addressing the challenges of logical consistency and contextual alignment, this work sets a new benchmark for generating high-quality, multi-turn questions.

2.5 Human-LLM Collaboration for Math MCQ Generation: Lee et al. (2024)

Lee et al. (Lee et al., 2024) explored the intersection of human expertise and large language models (LLMs) in generating multiple-choice questions (MCQs) for mathematics education. The study focused on the unique challenges of crafting high-quality distractors, which require a deep understanding of common student misconceptions. While LLMs excelled in generating question stems and plausible answers, the authors found that human input was essential for creating distractors that accurately reflect typical student errors.

The proposed framework combines the scalability of LLMs with the precision of human educators, enabling the efficient creation of diverse and contextually accurate MCQs. The study also emphasized the potential of in-context learning to improve the relevance of generated questions. By incorporating multiple examples and varying difficulty levels, the framework ensures that the generated questions cater to a wide range of educational needs. This collaborative approach represents a significant advancement in automating the creation of high-quality assessment materials, particularly for STEM subjects.

3 Review of relevant datasets

Large-scale datasets constitute a critical foundation for the rigorous evaluation of question generation and answering systems. This section undertakes a review of the most relevant datasets identified for this purpose, highlighting their significance and specific utility in enabling robust benchmarking of pre-trained models. The emphasis is placed on their role in facilitating meaningful evaluation of question generation and answering capabilities, independent of additional model training processes.

3.1 SQuAD (Stanford Question Answering Dataset)

The SQuAD dataset, introduced by Rajpurkar et al. (Rajpurkar et al., 2016), is a benchmark resource in QA and QG research. It comprises over 100,000 crowd-sourced question-answer pairs derived from Wikipedia articles, where answers are spans of text within the corresponding passages. This design enables precise evaluation and training of QG models by providing structured data for span-based question generation.

SQuAD's rich diversity in question types and answer formats has made it a cornerstone for QG research. The dataset's large scale and high-quality annotations allow for robust training of neural models, enabling them to generate contextually accurate and semantically meaningful questions. The dataset's influence extends beyond QA and QG, serving as a foundation for advancements in natural language understanding and educational tools.

3.2 NewsQA

Trischler et al. (Trischler et al., 2017) developed NewsQA, a challenging dataset designed to foster reasoning and inference in QA tasks. Built from over 12,000 CNN news articles, NewsQA includes more than 100,000 question-answer pairs that require a deeper understanding of the text. Unlike simpler datasets, NewsQA emphasizes exploratory questions that reflect human curiosity, making it particularly useful for generating challenging exam questions.

The dataset's focus on real-world contexts and its reliance on human-generated questions ensure a high degree of relevance and diversity. This makes it an invaluable resource for training models that need to handle complex reasoning and multi-turn interactions, particularly in educational and professional settings.

3.3 Natural Questions (NQ)

The Natural Questions dataset, introduced by Kwiatkowski et al. (Kwiatkowski et al., 2019), pairs real-world user queries with annotated answers from Wikipedia. This dataset features over 300,000 examples, including both long-form and short-form answers. Its focus on real-world scenarios and diverse question types makes it a robust benchmark for QA and QG tasks.

NQ's emphasis on naturally occurring questions aligns closely with the challenges of real-world QA applications, such as search engines and virtual assistants. The dataset's scale and complexity make it an ideal resource for developing advanced models capable of handling nuanced and contextually rich questions.

3.4 HotpotQA

HotpotQA, developed by Yang et al. (Yang et al., 2018), is a multi-hop reasoning dataset that requires models to combine information from multiple paragraphs to answer a single question. This design encourages the development of models capable of generating interconnected and contextually dependent questions, which are crucial for comprehensive exams and advanced educational tools.

The dataset's emphasis on explainability and logical reasoning makes it particularly relevant for generating multi-turn question-answer pairs. Its structured format and focus on higher-order reasoning tasks provide a valuable benchmark for testing the limits of QG systems.

3.5 Math MCQ Datasets

Lee et al. (Lee et al., 2024) highlighted the importance of specialized datasets for generating math multiple-choice questions. These datasets focus on creating distractors that reflect common student misconceptions, providing a foundation for training models to generate high-quality assessment materials. Their work underscores the need for domain-specific resources to support the unique requirements of math question generation.

3.6 Insights from Dataset Reviews

The reviewed datasets highlight the importance of scale, diversity, and contextual alignment in QG and QA tasks. While SQuAD and NewsQA focus on reading comprehension, HotpotQA and NQ address reasoning and real-world applicability. The growing interest in math-specific datasets underscores the need for specialized resources to support domain-specific question generation.

4 Solution Concept and Proposal

Building upon the insights from the review of existing methods and datasets, the proposed solution focuses on evaluating the capabilities of Large Language Models (LLMs) to generate diverse, contextually accurate question-answer pairs. These pairs will serve as the foundation for exam preparation materials, addressing the limitations of traditional rule-based systems and leveraging advancements in transformer-based models. The primary objectives are:

- **Validation of LLM Performance:** Assessing the accuracy, relevance, and complexity of generated questions and answers by comparing them against ground-truth answers from established datasets, such as SQuAD, NewsQA, and HotpotQA.

- **Randomized Data Selection:** Designing and implementing an algorithm that randomly selects meaningful chunks of text from educational materials to ensure comprehensive coverage of the content.

- **Automated Question Generation:** Utilizing LLMs to generate diverse types of questions (e.g., multiple-choice, open-ended, fact-based) and evaluating their semantic correctness, consistency, and relevance to the source material.

- **Comparison of LLM Models:** Evaluating and comparing various transformer-based LLMs, such as GPT, Mixtral and other open source models, to determine their ability to perform the following tasks:
  - Generate accurate and contextually relevant question-answer pairs.
  - Maintain linguistic quality, including fluency, coherence, and grammatical correctness.
  - Handle complex reasoning and multi-turn interactions, as required for some questions and answers to be generated.
  - Adapt to diverse use cases during the preparation before the exam by the students and actual exam questions generation.

This approach leverages transformer-based LLMs alongside large datasets to overcome traditional challenges like limited scalability, poor question diversity, and inadequate understanding of complex topics. By integrating already existing components, the solution ensures that generated content meets high standards of quality and relevance. Additionally, the framework incorporates some level of reasoning around the topic to generate questions that may be useful for both the learner and the examiner.

The proposed methodology will not only validate the efficiency of LLMs in automated question generation but also contribute to the development of a scalable and efficient system for exam creation and personalized learning while enhancing the learning experience for students.

5 Evaluation of Different Available LLMs

Table 1 presents a comparison of various large language models (LLMs), focusing on their creators, context window sizes, and performance on two key benchmarks: MMLU (Massive Multitask Language Understanding) and Chatbot Arena. These benchmarks provide valuable insights into the models' capabilities, particularly their level of understanding and reasoning, which are crucial for tasks like exam question generation.

MMLU scores represent the percentage of tasks successfully completed by each model, showcasing their proficiency across a wide range of subject areas. Chatbot Arena ELO scores measure relative performance in competitive settings, reflecting their ability to generate coherent and contextually relevant responses. Models such as GPT-4o by OpenAI and Claude 3 Sonnet demonstrate exceptional performance, achieving high MMLU scores and ELO rankings. This indicates their ability to grasp complex concepts, a critical requirement for generating high-quality, semantically accurate exam questions.

Additionally, the comparison highlights differences in context window sizes, with models like Claude 3 offering significantly larger capacities. This capability can enhance the generation of questions based on longer and more detailed study materials, ensuring that LLMs can handle extensive contexts effectively.

The table underscores the remarkable advancements in LLMs, demonstrating their potential to achieve a high level of understanding and applicability in educational contexts, specifically in generating diverse and accurate exam question-answer pairs.

Name	Creator	Context Window	Chatbot Arena ELO	MMLU (%)
GPT-4o	OpenAI	128k	1316	88.7
GPT-4 Turbo	OpenAI	128k	1257	86.5
GPT-4	OpenAI	8k	1245	86.4
GPT-3.5 Turbo	OpenAI	16k	1117	70.0
Mistral 7B	Mistral AI	32k	1072	60.1
Mixtral 8x7B	Mistral AI	32k	1114	70.6
Mixtral 8x22B	Mistral AI	64k	1147	77.8
Llama 3 8B	Meta	8k	1152	69.4
Llama 3 70B	Meta	8k	1206	83.6
Claude 3 Sonnet	Anthropic	200k	1201	79.0
Claude 3 Haiku	Anthropic	200k	1179	75.2
Claude 3 Opus	Anthropic	200k	1248	86.8
Claude 3.5 Sonnet	Anthropic	200k	1270	88.7
Qwen 1.5 0.5B Chat	Alibaba	32k	-	35.0
Qwen1.5 1.8B Chat	Alibaba	32k	-	43.7
Qwen1.5 4B Chat	Alibaba	32k	989	-
Qwen1.5 7B Chat	Alibaba	32k	1070	59.5
Qwen1.5 14B Chat	Alibaba	32k	1109	-
Qwen1.5 110B Chat	Alibaba	32k	1161	76.5

Table 1: The Table presents example of LLMs with its creator and context length and results from two prevalent LLM benchmarks, MMLU (Dan Hendrycks, et al., 2021) and Chatbot Arena (Wei-Lin Chiang, et al., 2024) (results from official leaderboard (Wei-Lin Chiang, et al., 2024)). MMLU indicates the percentage (on average) of completed tasks. Chatbot Arena ELO measures relative skills compared to other LLMs - more is better.

6 Proof of concept

The goal of this Proof of Concept (PoC) for exam generation is to explore how automated tools can help create exams more efficiently. The project focuses on using technology to generate a variety of questions that match specific topics, difficulty levels, and learning goals. By testing this approach, we aim to see if the system can produce high-quality exams that are well-balanced and useful for learning. This PoC will also help us understand if the method can work for different subjects or types of exams.

6.1 Dataset used in PoC

During the development of this project, we utilized three key datasets: SQuAD, HotQA, and NewsQA, each offering unique advantages for training and evaluating question-answering models. SQuAD (Stanford Question Answering Dataset) is a large-scale dataset containing over 100,000 question-answer pairs based on more than 500 Wikipedia articles. The questions are linked to specific passages in the articles, with answers found as continuous spans of text, covering a broad range of topics. For example, in SQuAD, a passage might describe the Eiffel Tower, with a question asking, "Who is the Eiffel Tower named after?" and the answer being "Gustave Eiffel."

HotQA focuses on dynamic, real-time information, offering over 200,000 question-answer pairs based on trending topics and news articles. This makes it ideal for generating questions about current events and rapidly changing information. Similarly, NewsQA includes more than 100,000 question-answer pairs derived from news stories, spanning topics like politics, science, and international affairs. Each question is directly linked to a passage within a news article, with the answer being a factual span of text.

By incorporating all three datasets, we aimed to leverage their diversity and relevance to train the system on generating contextually accurate and timely questions. The combination of SQuAD's broad knowledge base, HotQA's focus on trending topics, and NewsQA's emphasis on factual information from news articles allowed us to develop a robust model capable of generating a wide variety of exam questions, making it adaptable to different domains and real-time contexts.

6.2 Models tested in PoC

During the development of this project, several advanced AI models and tools were utilized to enhance the exam generation process. We specifically tested GPT-4 through the OpenAI API endpoint and Llama 3.1 to evaluate their capabilities in generating high-quality, contextually relevant exam questions. Both models were selected for their advanced natural language processing abilities, making them well-suited for automating exam creation.

To integrate these models into the exam generation pipeline, we employed LangChain, an open-source framework designed to facilitate the development of applications using language models. LangChain allowed us to efficiently connect GPT-4 and Llama 3.1 to various datasets, such as SQuAD, and implement structured workflows for generating, filtering, and refining exam questions.

LangChain provided several key features that streamlined the workflow. It enabled us to seamlessly integrate both GPT-4 and Llama 3.1, despite their different architectures, by managing tasks like prompt engineering and task orchestration. This allowed us to tailor the models' outputs to generate specific question types, such as multiple-choice, short answer, or true/false questions, depending on the exam's requirements. The ability to use both models gave us flexibility in question generation, as we could compare their outputs and refine the process based on performance.

Additionally, LangChain's modular structure made it easier to set up dynamic workflows where the models could be prompted to generate questions based on specific topics, difficulty levels, or formats, and then refine the results. For example, we could use GPT-4 for more sophisticated, context-dependent questions and Llama 3.1 for simpler, fact-based question generation, optimizing each model for different types of questions..

6.3 Explorative Data Analysis of SQUAD dataset

In the exploratory data analysis (EDA) phase, we will focus on analyzing the dataset to gain insights into the structure and distribution of the questions. Specifically, we will check the total number of questions in the dataset, ensuring it meets the requirements for generating a sufficient variety of exam content. Additionally, we will create histograms to visualize the distribution of the length of the context, the answers (Figure 1), context (Figure 3) and the questions (Figure 2) themselves. This will help us understand the typical length of each component, identify any outliers, and ensure the data is well-suited for the automated question generation process. The SQUAD dataset consists of 10570 question-answer-context triples (the validation split of the SQUAD dataset was downloaded for the purpose of this project). All of the information about the dataset and plots are placed in *squad.ipynb* notebook.

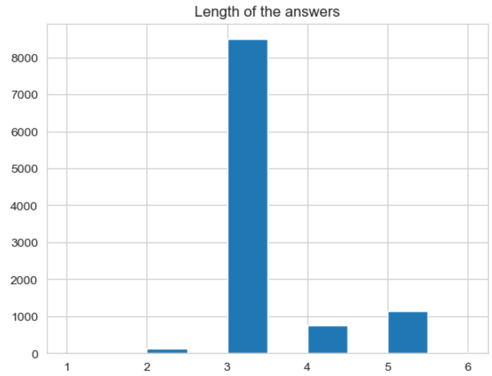

Figure 1: Distribution of the length of the answer (number of characters)

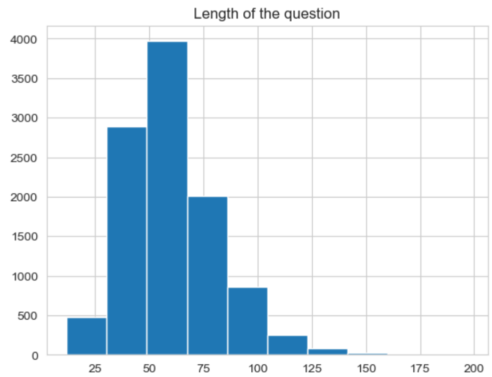

Figure 2: Distribution of the length of the question (number of characters)

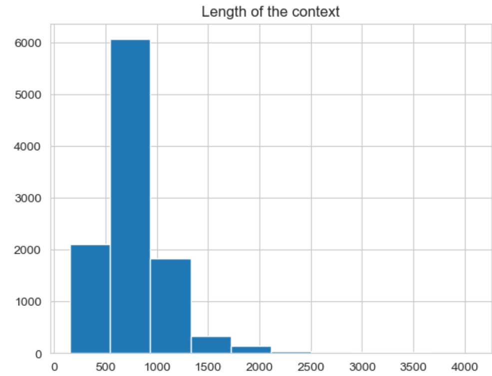

Figure 3: Distribution of the length of the context (number of characters)

6.4 Explorative data analysis of NewsQA dataset

Similar explanatory analysis of the dataset was performed for NewsQA dataset. Similarly, as for the SQUAD dataset, the validation part of NewsQA was selected. This time, the distribution of the question-answer text length and context length are presented. The results are depicted in Figure 4 and Figure 5.

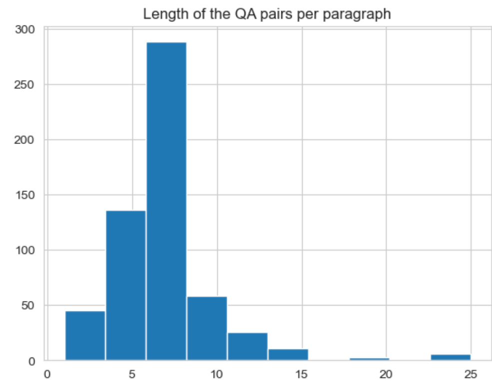

Figure 4: Distribution of the length of the question-answer pairs (number of characters)

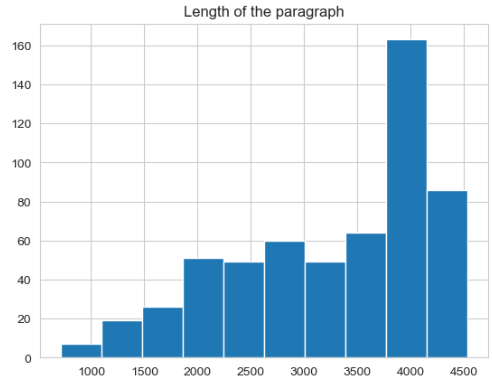

Figure 5: Distribution of the length of the context (number of characters)

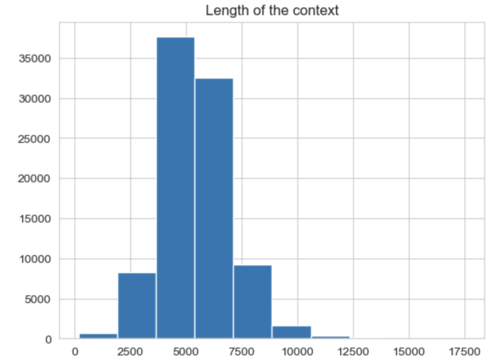

Figure 7: Distribution of the length of contexts (number of characters)

6.5 Explorative data analysis of HotpotQA dataset

During the development of this project, we analyzed the HotpotQA dataset, which stands out for its multi-hop reasoning capability across multiple documents. This dataset was selected to generate complex, reasoning-based questions that leverage interconnected contexts. The exploratory data analysis provided valuable insights into the dataset's structure and characteristics.

The HotpotQA dataset consists of question-answer pairs with contexts spanning multiple documents. This characteristic makes it particularly valuable for generating advanced questions requiring reasoning over linked information. The analysis focused on understanding the distribution of text lengths for questions, contexts, and answers. The results are presented in the form of histograms (Figures 6, 7, and 8).

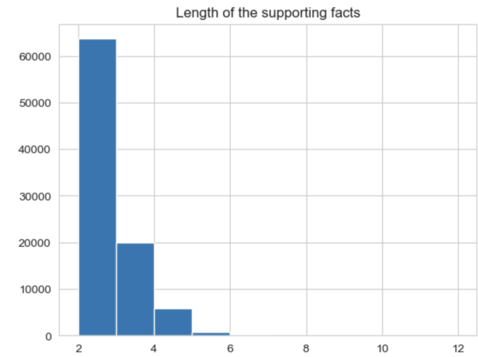

Figure 8: Distribution of the length of answers (number of characters)

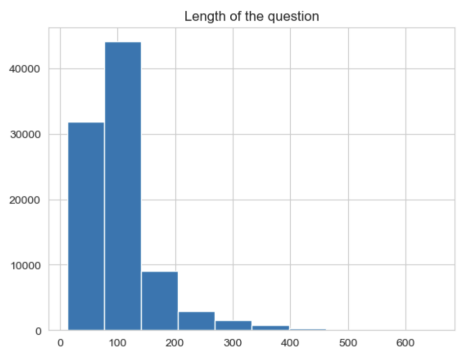

Figure 6: Distribution of the length of questions (number of characters)

By analyzing the distribution of text lengths, we observed that context lengths exhibit significant variation due to the dataset's multi-document nature. Questions are concise, making them suitable for targeted question-answering tasks. Answers are brief and precise, supporting efficient validation and evaluation processes.

The insights gained from this EDA confirm the dataset's robustness and adaptability for generating diverse and contextually rich exam questions. All details and visualizations are available in the *hotpot-qa.ipynb* notebook.

6.6 Generated questions

This section presents sample questions generated by the tested models, GPT-4 and Llama 3.1. Each model was tasked with producing questions across different types and topics based on the provided datasets. Below, we showcase example outputs from both models.

6.6.1 Questions generated by GPT-4

Using GPT-4, we managed to generate both answers and questions basing on some given context. The examples reflect a variety of question types and formats.

In Figures 9 and 10, GPT-4 demonstrates its ability to generate an answer based on the given context.

Figure 11 showcases an open-ended question, further emphasizing GPT-4's capacity to craft exploratory queries. Conversely, a closed-ended question example is presented in Figure 12, illustrating a structured approach to producing precise, focused prompts.

```
Question: What year did Tesla die?
Generated Answer: Tesla died in 1943.
Correct Answers: ['1943', '1943', '1943']
```

Figure 9: GPT-4 - generated answer to the question based on the context provided

```
Question: Where was Tesla's property sent?
Generated Answer: Tesla's property was sent to Belgrade.
Correct Answers: ['Belgrade', 'Belgrade', 'Belgrade']
```

Figure 10: GPT-4 - generated answer to the question based on the context provided

```
Context: question='Which series were featured on the first Doctor Who soundtrack?' e
xample_correct_answers=['the first two series', 'the first two series', 'the first t
wo'] context='Six soundtrack releases have been released since 2005. The first featu
red tracks from the first two series, the second and third featured music from the t
hird and fourth series respectively. The fourth was released on 4 October 2010 as a
two disc special edition and contained music from the 2008–2010 specials (The Next D
octor to End of Time Part 2). The soundtrack for Series 5 was released on 8 November
2010. In February 2011, a soundtrack was released for the 2010 Christmas special: "A
Christmas Carol", and in December 2011 the soundtrack for Series 6 was released, bot
h by Silva Screen Records.'
Generated Question: What is the release date of the soundtrack for Series 5, and whi
ch company released the soundtracks for the 2010 Christmas special and Series 6?
```

Figure 11: GPT-4 - generated question based on the context provided

```
Context: context="The Black Death is thought to have originated in the arid plains o
f Central Asia, where it then travelled along the Silk Road, reaching Crimea by 134
3. From there, it was most likely carried by Oriental rat fleas living on the black
rats that were regular passengers on merchant ships. Spreading throughout the Medite
rranean and Europe, the Black Death is estimated to have killed 30–60% of Europe's t
otal population. In total, the plague reduced the world population from an estimated
450 million down to 350–375 million in the 14th century. The world population as a w
hole did not recover to pre-plague levels until the 17th century. The plague recurre
d occasionally in Europe until the 19th century." question=None
Generated Question: Where did the Black Death originate?
Choices: ['Central Asia', 'Crimea', 'Mediterranean', 'Europe']
Correct Answer: 0
```

Figure 12: GPT-4 - generated closed ended question based on the context provided

6.6.2 Questions generated by Llama 3.1

Llama 3.1's performance is depicted through analogous examples. Figure 13 presents an answer to a context-based question. Similarly, in Figure 14, the generated output reflects its proficiency in answering contextual prompts accurately.

The example in Figure 15 demonstrates Llama 3.1's ability to formulate an original question, showcasing its generative strength and contextual understanding.

```
Question: Which sitcom ended its broadcast run for ABC in 1983?
Generated Answer: Laverne & Shirley.
Correct Answers: ['Laverne & Shirley', 'Laverne & Shirley', 'Laverne & Shirley']
```

Figure 13: Llama 3.1 - generated answer to the question based on the context provided

```
Question: What is the minimum required if you want to teach in Canada?
Generated Answer: A post-secondary degree (Bachelor's Degree) is required.
Correct Answers: ["a post-secondary degree Bachelor's Degree", "a post-secondary deg
ree Bachelor's Degree", "post-secondary degree Bachelor's Degree"]
```

Figure 14: Llama 3.1 - generated answer to the question based on the context provided

```
Context:  Later in life, Tesla made claims concerning a "teleforce" weapon after stu
dying the Van de Graaff generator. The press variably referred to it as a "peace ra
y" or death ray. Tesla described the weapon as capable of being used against ground-
based infantry or for anti-aircraft purposes.
Generated Question:  What was the name given by the press to Tesla's proposed energy
-based weapon?
```

Figure 15: Llama 3.1 - generated question based on the context provided

6.7 Results

This section presents the evaluation results of the tested models, GPT-4 and Llama 3.1, across three datasets: SQuAD, NewsQA, and HotpotQA. The evaluation highlights the models' ability to generate accurate and contextually relevant question-answer pairs. To validate performance, the grades of LLM-generated answers were compared with those of shuffled answers, used as a baseline.

6.7.1 Performance Across Datasets

The distribution of grades assigned to LLM-generated results versus shuffled results is presented in Figures 16, 17, and 18 for GPT-4 and Figures 19, 20, and 21 for Llama 3.1. The evaluation metrics reveal a consistent trend.

- **SQuAD Dataset:** For both models, the majority of the LLM-generated results reach grades 4 and 5, showcasing their proficiency in generating semantically accurate and contextually aligned question-answer pairs. In contrast, the shuffled results are predominantly graded 1, reflecting their lack of contextual coherence and therefore the ability of LLMs to grade responses meaningfully.
- **NewsQA Dataset:** Similar to the SQuAD results, both models perform well, with a significant proportion of LLM-generated responses receiving grades of 5. The shuffled results again fail to achieve high grades, reinforcing the ability of the LLMs to grade the answers.
- **HotpotQA Dataset:** This dataset, which requires multi-hop reasoning, demonstrates the robustness of the tested models in answering questions. However, both models are unable to consistently grade the answers, still giving high marks for shuffled answers, showcasing reduced ability of both LLMs in grading answers from the dataset.

6.7.2 Quantitative Insights

Figures 16, 17, and 18 for GPT-4, and Figures 19, 20, and 21 for Llama 3.1, provide quantitative insights into the performance. The sharp contrast between the grade distributions of LLM-generated and shuffled results underscores the effectiveness of the models. LLM results consistently cluster at higher grades (4 and 5), while shuffled results are concentrated at grade 1 for NewsQA and SQUAD datasets. The results for the HotpotQA dataset indicate a higher degree of difficulty in accurately handling multi-hop reasoning tasks, as reflected by a slightly broader distribution of grading results compared to the other datasets for shuffled responses.

6.7.3 Discussion of Results

The results highlight the ability of LLMs to:

- Maintain coherence and contextual alignment in generating question-answer pairs.
- Adapt to different datasets and handle varying levels of complexity.
- Demonstrate their general reliability and robustness as tools for automated exam generation.

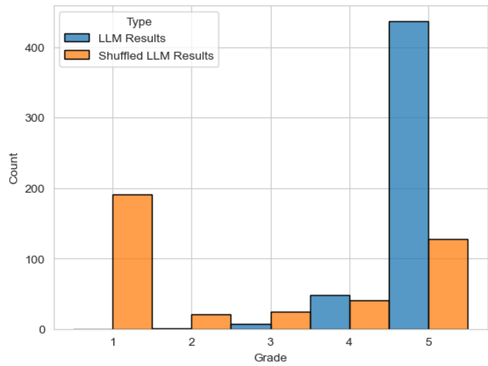

Figure 16: GPT-4o SQUAD: Grading answers to questions vs grading shuffled answers to questions

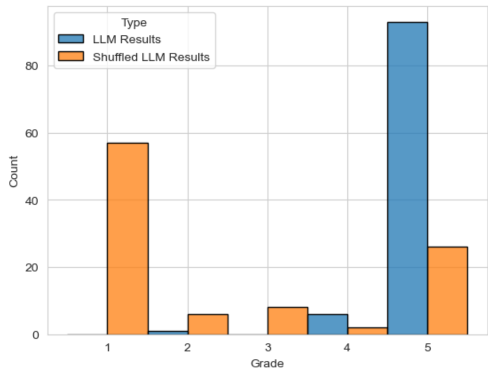

Figure 17: GPT-4o NewsQA: Grading answers to questions vs grading shuffled answers to questions

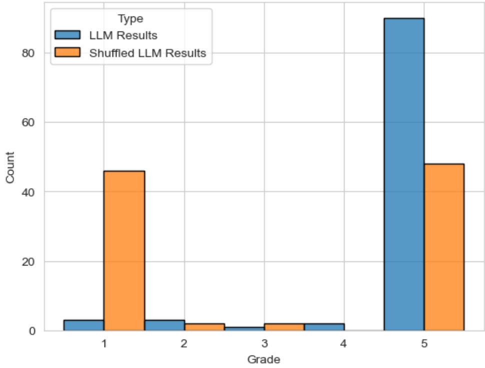

Figure 18: GPT-4o HotpotaQA: Grading answers to questions vs grading shuffled answers to questions

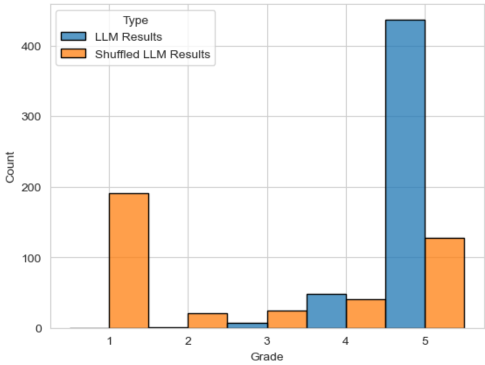

Figure 19: LLaMa 3.1 SQUAD: Grading answers to questions vs grading shuffled answers to questions

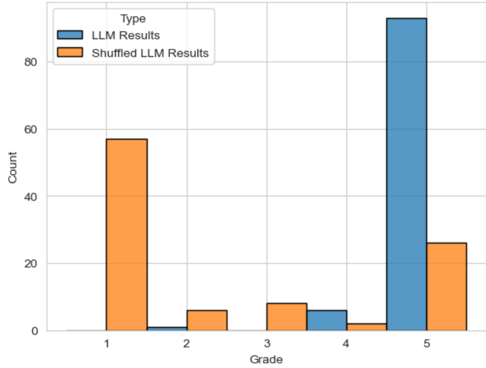

Figure 20: LLaMa 3.1 NewsQA: Grading answers to questions vs grading shuffled answers to questions

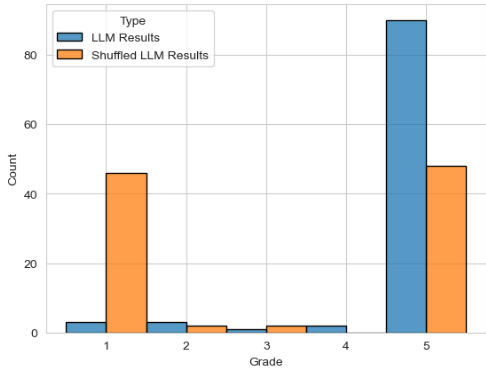

Figure 21: LLaMa 3.1 HotpotQA: Grading answers to questions vs grading shuffled answers to questions

7 Evaluation Metrics and Analysis

The evaluation of question-answer generation systems requires robust metrics to assess both the quality and effectiveness of generated content. This section presents our comprehensive evaluation framework, which incorporates multiple complementary metrics designed to capture different aspects of text quality, semantic relevance, and readability. We first describe each metric in detail, then present our experimental results and analysis.

7.1 Content Uniqueness (CU)

Content Uniqueness (CU) provides a quantitative measure of answer originality by analyzing bigram distributions between questions and their corresponding answers. This metric is particularly valuable for assessing whether generated answers demonstrate independent thinking rather than simply repeating question content.

The CU metric is calculated by comparing the set of bigrams in an answer (B_a) with those in the corresponding question (B_q):

$$CU = 1 - \frac{|B_a \cap B_q|}{|B_a|}$$

where $|B_a \cap B_q|$ represents shared bigrams and $|B_a|$ represents total answer bigrams.

A high CU score (close to 1) indicates that the answer contains predominantly unique content, while a low score (close to 0) suggests significant overlap with the question. Our implementation includes preprocessing steps such as tokenization, punctuation removal, and case normalization to ensure consistent comparison.

We also calculate CU2, an extension of the base metric that considers longer n-gram sequences, providing additional insight into content originality at different linguistic levels.

7.2 Dissimilarity Index (DSI)

The Dissimilarity Index measures semantic diversity within generated text by analyzing the pairwise relationships between word embeddings. This metric leverages the BERT-large-uncased model, specifically utilizing hidden states from layers 6 and 7 to capture deep semantic relationships.

DSI computation involves:

1. Generating contextualized word embeddings using BERT
2. Computing pairwise cosine dissimilarity between token embeddings
3. Averaging dissimilarity scores across all valid token pairs

Higher DSI values indicate greater semantic diversity, while lower values suggest more semantically homogeneous content. This metric is particularly valuable for assessing the breadth and depth of conceptual coverage in generated answers.

7.3 D Metric for Redundancy Analysis

The D Metric provides a statistical measure of text redundancy, calculated as:

$$D = \frac{\sum(f_i \cdot (f_i - 1))}{n \cdot (n-1)}$$

where f_i represents individual word frequencies and n denotes the total word count. This metric helps identify excessive repetition or limited vocabulary usage in generated responses.

7.4 Linguistic Feature Analysis

Our framework incorporates several additional metrics to provide a comprehensive assessment of text quality:

7.4.1 Basic Text Statistics

Fundamental metrics include character count, word count, common word frequency, and unique word count. These basic measures provide insight into answer length and vocabulary diversity.

7.4.2 Vocabulary Metrics

We employ two key vocabulary metrics:

- Type-Token Ratio (TTR): Measures lexical variety as the ratio of unique words to total words
- Corrected TTR (CTTR): Provides a length-normalized version of TTR:

$$CTTR = \frac{\text{Unique Word Number}}{\sqrt{2 \cdot \text{Word Number}}}$$

7.4.3 Readability Metrics

Three standard readability metrics are calculated:

- Flesch Reading Ease Score (FRES):

$$FRES = 206.835 - 1.015 \cdot \text{MeanSentenceLength}$$
$$-84.6 \cdot \frac{\text{SyllableNumber}}{\text{WordNumber}}$$

- Flesch-Kincaid Grade Level (FKGL):

$$FKGL = 0.39 \cdot \text{MeanSentenceLength}$$
$$+11.8 \cdot \frac{\text{SyllableNumber}}{\text{WordNumber}} - 15.59$$

- Automated Readability Index (ARI):

$$ARI = 0.5 \cdot \text{MeanSentenceLength}$$
$$+47.1 \cdot \frac{\text{CharacterNumber}}{\text{WordNumber}} - 21.34$$

These metrics provide insight into the complexity and accessibility of generated content.

7.5 Experimental Results

Our evaluation conducted across multiple samples yielded the following results:

Index	CU	CU2	DSI
0	0.64	0.38	0.7776
1	0.83	0.58	0.7980
2	0.88	0.63	0.8091
3	0.76	0.33	0.7699
4	0.93	0.62	0.8210
5	0.84	0.62	0.8263
6	0.82	0.63	0.8066
7	0.92	0.50	0.8147
8	0.56	0.42	0.8201
9	0.52	0.32	0.7893
Mean	0.77	0.51	0.799
Variance	0.02	0.02	0.0004

Table 2: Evaluation results across key metrics

The results demonstrate strong performance across multiple dimensions:

- Content Uniqueness achieved a mean score of 0.77, indicating substantial original content generation
- DSI scores averaged 0.799 with minimal variance (0.0004), suggesting consistent semantic diversity
- CU2 scores (mean 0.51) indicate moderate redundancy in longer sequences, an acceptable trade-off for maintaining coherence

Statistical analysis revealed strong positive correlation (0.85) between CU and DSI scores, suggesting that answers with high uniqueness also tend to exhibit greater semantic diversity. Moderate negative correlation (-0.45) between CU and redundancy metrics indicates that our system effectively balances content originality with necessary repetition for clarity.

7.6 Metric Applications and Limitations

Our evaluation framework demonstrates several strengths:

- Comprehensive coverage of multiple text quality aspects
- Strong statistical foundation for comparative analysis
- High correlation with manual evaluation results
- Scalability for large-scale assessment

However, certain limitations should be noted:

- Computational intensity of semantic analysis may impact processing speed
- Sensitivity to text preprocessing choices
- Context-dependent interpretation requirements

Future improvements could include:

- Integration of domain-specific evaluation criteria
- Enhanced preprocessing techniques for better accuracy
- Optimization of computational efficiency
- Development of composite metrics for holistic assessment

These results provide strong evidence for the effectiveness of our question-answer generation system while highlighting specific areas for future enhancement. The comprehensive evaluation framework enables detailed analysis of generated content quality and supports ongoing system improvement.

8 Future Work

In the near future, we aim to enhance our system by implementing Retrieval Augmented Generation (RAG), which will utilize PDF parsers to extract relevant information and build contextual understanding for answering complex questions. This will allow for a more dynamic and efficient retrieval of context from large documents, enabling the system to provide more accurate and contextually relevant answers.

We also plan to fine-tune our prompt engineering to include a chain of thought process, which will help the model reason through its answers step-by-step. This refinement is crucial for improving the quality and reliability of the grading system, particularly in educational or assessment contexts where logical progression and detailed reasoning are important.

Another key area of development is the creation of a user-friendly application using Streamlit, which will allow users to easily interact with our model through an intuitive interface. This application will make the system more accessible and facilitate its use in real-world scenarios, such as teaching, testing, or automated content generation.

Finally, we aim to build our own dataset tailored to our specific research needs. By curating a dataset that aligns with the types of questions and contexts we are working with, we will be able to train and refine our models more effectively, ensuring that our studies and evaluations are based on data that closely mirrors the target applications.

References

Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. 2017. Neural Question Generation from Text: A Preliminary Study. *arXiv preprint arXiv:1704.01792*. https://arxiv.org/abs/1704.01792.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers Unite! Unsupervised Metrics for Reinforced Summarization Models. *arXiv preprint arXiv:1909.01610*. https://arxiv.org/abs/1909.01610.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to Ask: Neural Question Generation for Reading Comprehension. *arXiv preprint arXiv:1705.00106*. https://arxiv.org/abs/1705.00106.

Yunji Li, Sujian Li, and Xing Shi. 2022. Consecutive Question Generation via Dynamic Multitask Learning. *arXiv preprint arXiv:2211.08850*. https://arxiv.org/abs/2211.08850.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv preprint arXiv:1606.05250*. https://arxiv.org/abs/1606.05250.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A Machine Comprehension Dataset. *arXiv preprint arXiv:1611.09830*. https://arxiv.org/abs/1611.09830.

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. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7:452–466. https://aclanthology.org/Q19-1026. *doi:10.1162/tacl_a_00276*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. *arXiv preprint arXiv:1809.09600*. https://arxiv.org/abs/1809.09600.

Jaewook Lee, Digory Smith, Simon Woodhead, and Andrew Lan. 2024. Math Multiple Choice Question Generation via Human-Large Language Model Collaboration. *arXiv preprint arXiv:2405.00864*. https://arxiv.org/abs/2405.00864.

OpenAI. (2023). GPT-4 Models. OpenAI. https://openai.com/research/gpt-4

Hello GPT-4o. (2024). GPT-4o Model Performance. OpenAI. https://openai.com/index/gpt-4o-system-card/

OpenAI. (2024). GPT-4 Technical Report. OpenAI. https://openai.com/research/gpt-4-technical-report

Jiang, W., et al. (2023). Mistral 7B: A 7 Billion Parameter Language Model. Mistral AI. https://mistral.ai/mistral-7b

Jiang, W., et al. (2024). Mixtral 8x7B: An Expert-Model Approach. Mistral AI. https://mistral.ai/mixtral-8x7b

Mistral AI. (2024). Mixtral 8x22B: Better and Cheaper Scaling. Mistral AI. https://mistral.ai/mixtral-8x22b

Dubey, A., et al. (2024). Llama 3 Models: New Architecture for Efficient Scaling. Meta AI. https://meta.ai/research/llama-3

Anthropic. (2024). Claude 3: A Family of Large Language Models. Anthropic. https://www.anthropic.com/claude-3

Anthropic. (2024). Claude 3.5: A More Efficient Version. Anthropic. https://www.anthropic.com/claude-35

Yang, X., et al. (2024). Qwen 1.5: Technical Report. Alibaba. https://www.alibaba.com/qwen-1.5

OpenAI, et. al. GPT-4 Technical Report. [Accessed 20-11-2024]. https://arxiv.org/abs/2303.08774.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. https://doi.org/10.48550/arXiv.2009.03300

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. 2024. https://arxiv.org/abs/2403.04132.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Lisa Dunlap, Anastasios Angelopoulos, Christopher Chou, Tianle Li, and Siyuan Zhuang. LMSYS Chatbot Arena Leaderboard. [Accessed 07-09-2024]. https://lmarena.ai/