

# RAG Exam Generation

Zofia Łągiewka, Jacek Zalewski, Nikita Kozlov, Jakub Świstak

Natural Language Processing

Warsaw University of Technology

November 2024

# Project overview

Steps of the project:

- Explore existing solutions and prepare literature review
- EDA of the datasets that can be used
- Create a RAG system capable of retrieving question-answer pairs given context
- Validate approach
- User-interface

## Datasets Used:

- **SQuAD** (Stanford Question Answering Dataset) – A widely-used benchmark dataset for reading comprehension.
- **NewsQA** – A dataset focused on question answering over news articles.
- **HotpotQA** – A dataset designed for multi-hop question answering.

# Models and Tools Used

## Models:

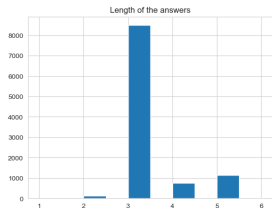
- **GPT-4o**
- **Llama 3.1**

## Tools:

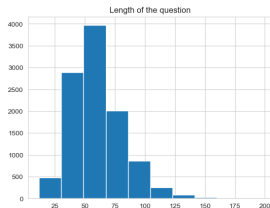
- **LangChain** – A framework designed for building applications that leverage LLMs (Large Language Models) to create pipelines, manage prompts, and interact with external tools.

# EDA of SQUAD dataset

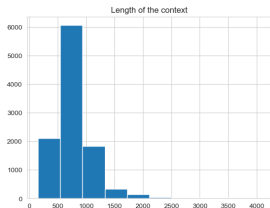
Validation split consists of 10570 question-answer-context triples.



Answer Length  
(characters)



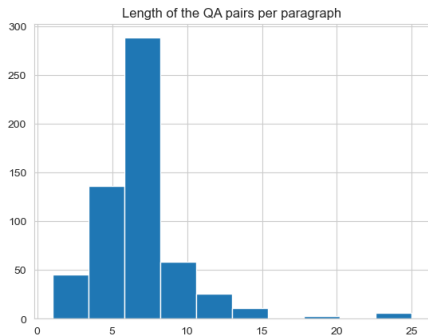
Question Length  
(characters)



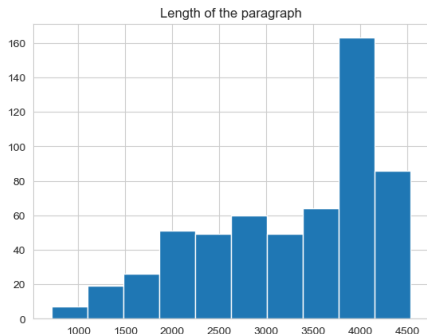
Context Length  
(characters)

# EDA of NewsQA dataset

Validation split consists of 574 question-answer context triples



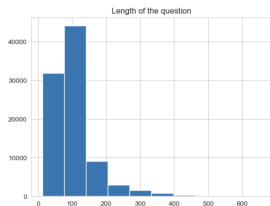
Distribution of the length of the question-answer pairs (number of characters)



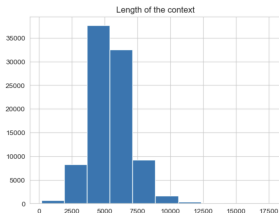
Distribution of the length of the context (number of characters)

# EDA of Hotpot QA dataset

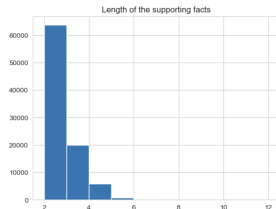
Validation split consists of 90447 question-answer context triples.



Distribution of the length of questions  
(number of characters)



Distribution of the length of contexts  
(number of characters)



Distribution of the length of answers  
(number of characters)

# Generated Questions and Answers by GPT-4

**Question:** What year did Tesla die?

**Generated Answer:** Tesla died in 1943.

**Correct Answers:** ['1943', '1943', '1943']

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



# Generated Questions and Answers by GPT-4

Context: question='Which series were featured on the first Doctor Who soundtrack?' example\_correct\_answers=['the first two series', 'the first two series', 'the first two'] context='Six soundtrack releases have been released since 2005. The first featured tracks from the first two series, the second and third featured music from the third 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 Doctor 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, both by Silva Screen Records.'

Generated Question: What is the release date of the soundtrack for Series 5, and which company released the soundtracks for the 2010 Christmas special and Series 6?

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

# Generated Questions and Answers by GPT-4

```
Context: context="The Black Death is thought to have originated in the arid plains of Central Asia, where it then travelled along the Silk Road, reaching Crimea by 1343. 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 Mediterranean and Europe, the Black Death is estimated to have killed 30–60% of Europe's total 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 whole did not recover to pre-plague levels until the 17th century. The plague recurred 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: GPT-4 - generated closed-ended question based on the context provided

# Generated Questions and Answers by Llama 3.1

Question: Which sitcom ended its broadcast run for ABC in 1983?

Generated Answer: Laverne & Shirley.

Correct Answers: ['Laverne & Shirley', 'Laverne & Shirley', 'Laverne & Shirley']

**Figure:** 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 degree Bachelor's Degree", "post-secondary degree Bachelor's Degree"]

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

# Generated Questions and Answers by Llama 3.1

Context: Later in life, Tesla made claims concerning a "teleforce" weapon after studying the Van de Graaff generator. The press variably referred to it as a "peace ray" 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: Llama 3.1 - generated question based on the context provided

# Questions and answers by LLMs

- Can LLM generate the answer to the question based on the context provided?
- Can LLM grade the answer to the question based on the context provided and the example answer?
- Does LLM actually grade an answer or hallucinate it?

# Process of asking and grading a question

- $\text{Answer} = \text{Context} + \text{Question} [+ \text{Choices}]$
- $\text{Grade} = \text{Context} + \text{Question} [+ \text{Choices}] + \text{Answer}$
- $\text{Grade} \in [0, 5]$

# GPT 4o results

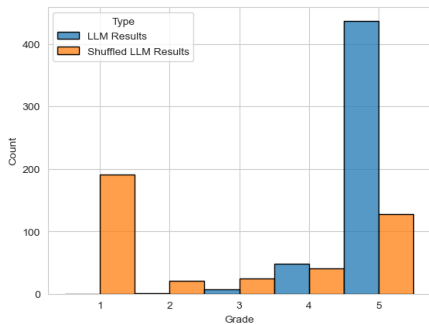


Figure: GPT-4o SQUAD: Grading answers to questions vs grading shuffled answers to questions

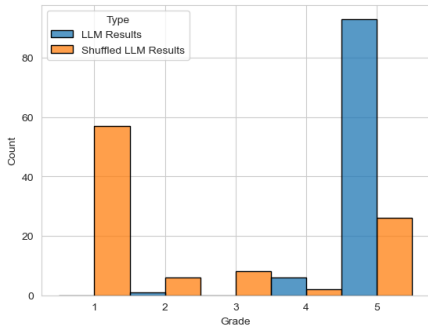
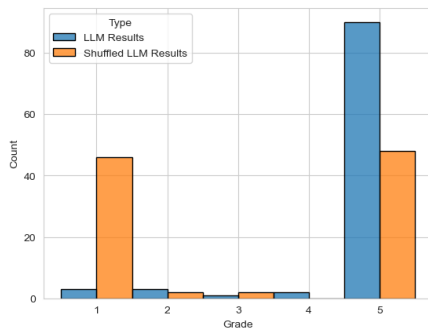


Figure: GPT-4o NewsQA: Grading answers to questions vs grading shuffled answers to questions

# GPT 4o results



**Figure:** GPT-4o HotpotQA: Grading answers to questions vs grading shuffled answers to questions



# LLama 3.1 results

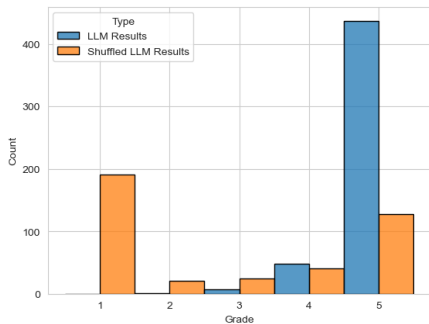


Figure: LLaMa 3.1 SQUAD: Grading answers to questions vs grading shuffled answers to questions

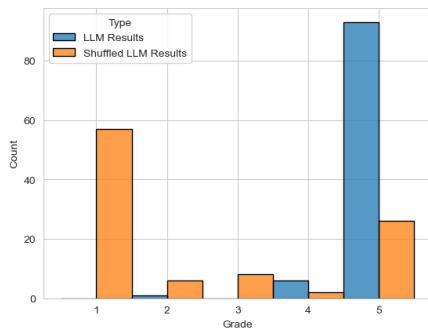
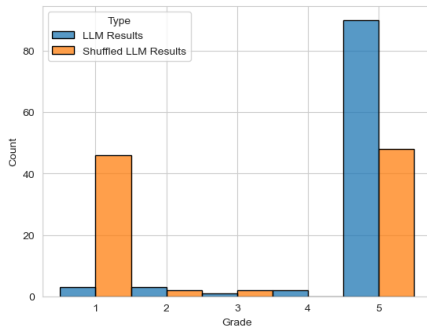


Figure: LLaMa 3.1 NewsQA: Grading answers to questions vs grading shuffled answers to questions

# LLama 3.1 results



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

# Future work

- Implement Retrieval Augmented Generation using PDF parsers to build a context for questions
- Fine-tune prompts to include chain of thoughts for grading the answers
- Create a user-friendly application using Streamlit
- Build our own dataset for some of our current studies

# RAG Exam Generation

Zofia Łągiewka, Jacek Zalewski, Nikita Kozlov, Jakub Świstak

Natural Language Processing

Warsaw University of Technology

November 2024