# RAG Exam Generation

Zofia Łągiewka, Jacek Zalewski, Nikita Kozlov, Jakub Świstak

Natural Language Processing

Warsaw University of Technology

January 23, 2025

# 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**

# Data Description

**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:**

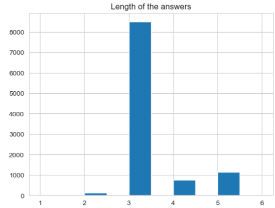
- **GPT-4o**
- **Llama 3.1**

**Tools:**

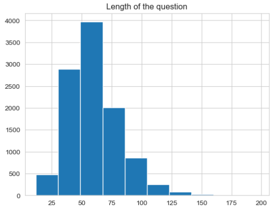
- **LangChain** – A framework designed for building applications that leverage LLMs (Large Language Models) to create pipelines, manage prompts, and interact with external tools.

# EDA of SQUAD dataset

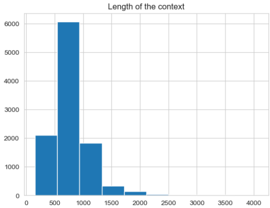
Validation split consists of 10570 question-answer-context triples.



Answer Length (characters)



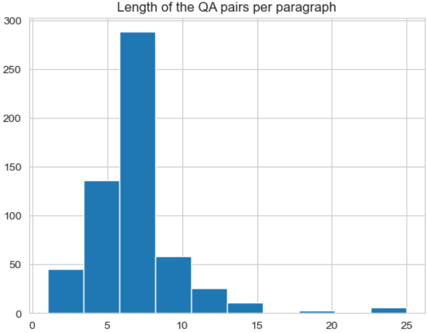
Question Length (characters)



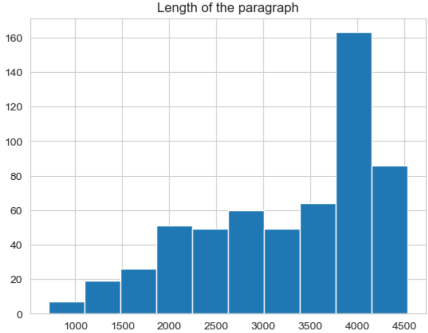
Context Length (characters)

# EDA of NewsQA dataset

Validation split consists of 574 question-answer context triples



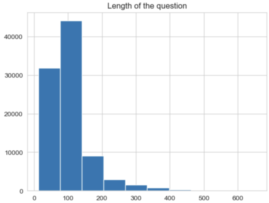
Distribution of the length of the question-answer pairs (number of characters)



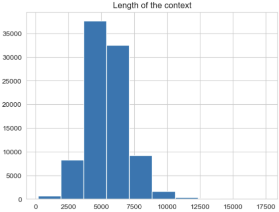
Distribution of the length of the context (number of characters)

# EDA of Hotpot QA dataset

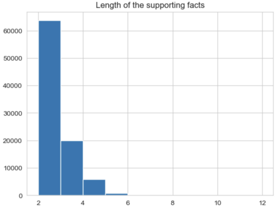
Validation split consists of 90447 question-answer context triples.



Distribution of the length of questions (number of characters)

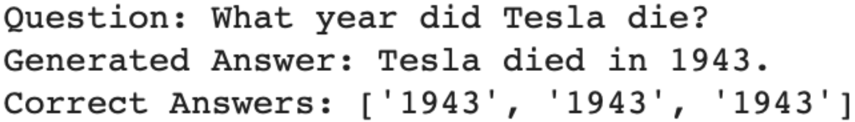


Distribution of the length of contexts (number of characters)



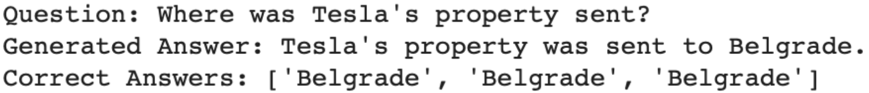
Distribution of the length of answers (number of characters)

# Generated Questions and Answers by GPT-4



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

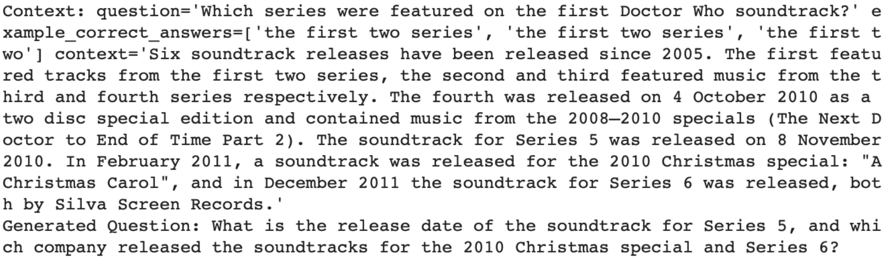
Figure: GPT-4 - generated answer to the question based on the context provided



```
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

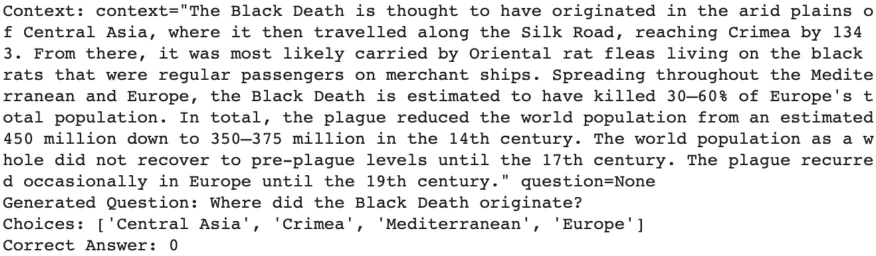
# Generated Questions and Answers by GPT-4



```
Context: question='Which series were featured on the first Doctor Who soundtrack?' 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: GPT-4 - generated question based on the context provided

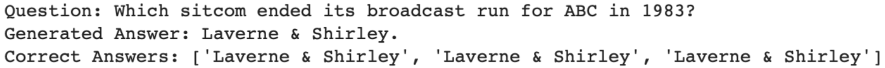
# Generated Questions and Answers by GPT-4



```
Context: context="The Black Death is thought to have originated in the arid plains 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
```

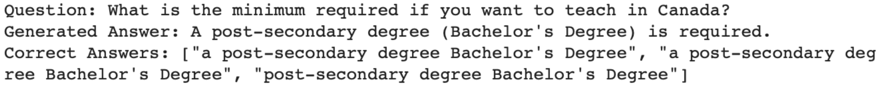
Figure: GPT-4 - generated closed-ended question based on the context provided

# 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']
```

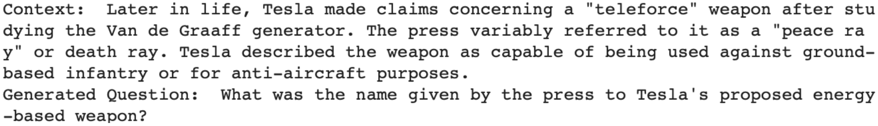
Figure: Llama 3.1 - generated answer to the question based on the context provided



```
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 deg
ree 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 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: 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

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

# GPT 4o results

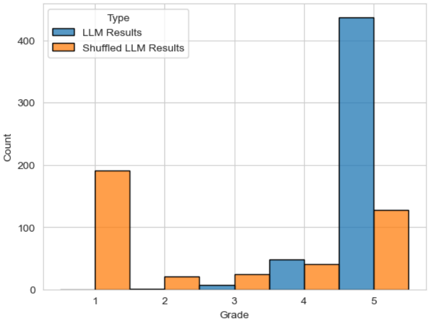


Figure: GPT-4o SQUAD: Grading answers to questions vs grading shuffled answers to questions

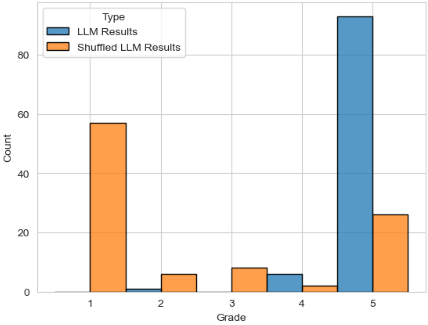


Figure: GPT-4o NewsQA: Grading answers to questions vs grading shuffled answers to questions

# GPT 4o results

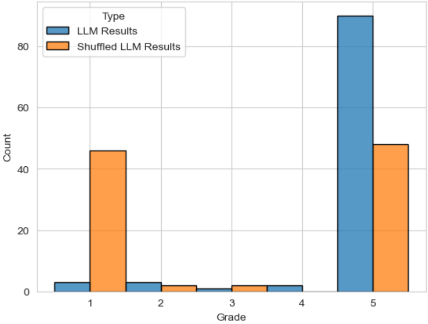


Figure: GPT-4o HotpotQA: Grading answers to questions vs grading shuffled answers to questions

# LLama 3.1 results

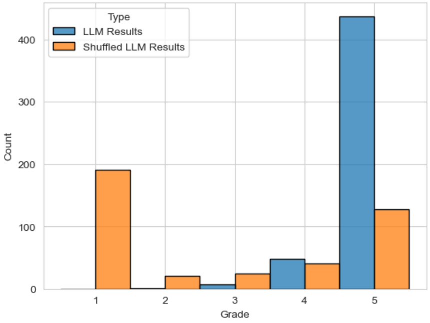


Figure: LLaMa 3.1 SQUAD: Grading answers to questions vs grading shuffled answers to questions

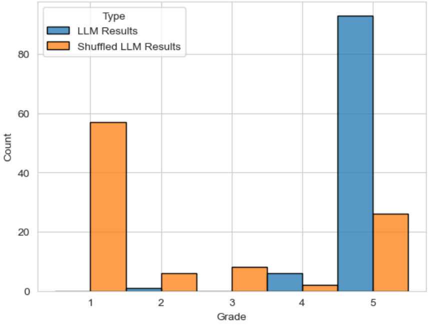


Figure: LLaMa 3.1 NewsQA: Grading answers to questions vs grading shuffled answers to questions

# LLama 3.1 results

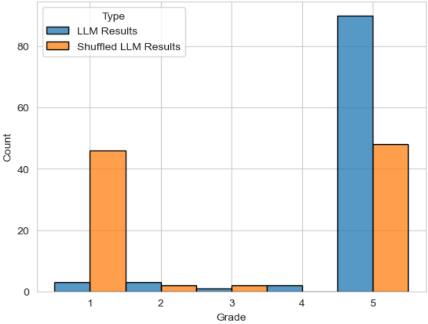


Figure: LLaMa 3.1 HotpotQA: Grading answers to questions vs grading shuffled answers to questions

# Content Uniqueness (CU) - Definition

**What it measures:**

- CU quantifies the uniqueness of the answer bigrams in comparison to:
  - Repeated bigrams within the answer itself (*self-redundancy*).
  - Bigrams shared between the answer and the question.

**How it works:**

- The metric computes the proportion of non-redundant bigrams in the answers after accounting for bigrams that:
  - Appear more than once in the answers.
  - Are present in the questions.

# Content Uniqueness (CU) - Definition

**What it measures:**

- CU quantifies the uniqueness of the answer bigrams in comparison to:
  - Repeated bigrams within the answer itself (*self-redundancy*).
  - Bigrams shared between the answer and the question.

**How it works:**

- The metric computes the proportion of non-redundant bigrams in the answers after accounting for bigrams that:
  - Appear more than once in the answers.
  - Are present in the questions.

# Content Uniqueness (CU) - Uniqueness

**Interpretation:**

- A high CU value (close to 1) suggests that the answer content is highly unique:
  - Limited overlap with the question.
  - Minimal repetition within the answer itself.
- A low CU value (close to 0) suggests that the answer content is either:
  - Highly redundant (repeats bigrams frequently).
  - Shares significant overlap with the question.

# What Does DSI Measure?

**Definition:**

- DSI quantifies the average pairwise cosine dissimilarity (1 − cosine similarity) between token embeddings within a text (e.g., sentences or answers).

**Interpretation:**

- **High DSI:**
  - Indicates a lower degree of semantic cohesion.
  - Words (tokens) are semantically less similar, reflecting diverse or less contextually focused content.
- **Low DSI:**
  - Indicates a higher degree of semantic cohesion.
  - Words (tokens) are semantically more similar, suggesting repetitive or highly focused content.

# How Does DSI Work?

**Steps:**

- **Tokenization and Hidden States Extraction:**
  - The text is segmented into sentences using a tokenizer.
  - Each sentence is tokenized into words and converted into embeddings using hidden states from layers 6 and 7 of a pre-trained BERT model (`bert-large-uncased`).
- **Cosine Dissimilarity Calculation:**
  - For each pair of embeddings within the text (ignoring diagonal pairs and duplicates), the cosine dissimilarity is computed.
- **Aggregation:**
  - The average dissimilarity across all token pairs forms the DSI score for the text.

# What Does the DSI Metric Say?

**High DSI Values:**

- Texts are more diverse in semantic content.
- Indicates rich, varied language usage or a lack of focus in structure.
- Example: Creative, varied responses in educational datasets.

**Low DSI Values:**

- Texts are more homogeneous or repetitive.
- May indicate repetitive phrasing, structured responses, or template-like answers.
- Example: Responses following predefined patterns.

# D Metric: Measuring Redundancy in Text

**What it measures:**

- The D Metric calculates the redundancy in a string by measuring how often words repeat relative to the total number of words.

**Interpretation:**

- **High D values:** Significant repetition of words, indicating low diversity.
- **Low D values:** More unique or diverse word usage.

# Metrics in Sentence Analysis: Linguistic Features

**General Linguistic Features:**

- **CharacterNumber and WordNumber:** Total number of characters and words; basic indicators of text length.
- **CommonWordNumber:** Proportion of common words in text; reflects simplicity of language.
- **UniqueWordNumber:** Count of unique words; indicates vocabulary diversity.

**Vocabulary Metrics:**

- **TTR (Type-Token Ratio):** Ratio of unique words to total words; shows lexical variety but is sensitive to text length.
- **CTTR (Corrected TTR):** A normalized version of TTR that reduces sensitivity to text length:

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

# Metrics in Sentence Analysis: Linguistic Features

**Redundancy Metric:**

- **DMetric:** Measures redundancy in the text by calculating the frequency of word repetition relative to the total word count.

**Readability Features:**

- **SyllableNumber:** Approximates syllable count for readability calculations.
- **SentenceNumber:** Total number of sentences in the text.
- **MeanSentenceLength:** Average characters per sentence; indicates verbosity and sentence complexity.

# Metrics in Sentence Analysis: Readability Features

**Advanced Readability Metrics:**

- **FRES (Flesch Reading Ease Score):** Measures how easy the text is to read Higher scores mean easier reading (e.g., 90–100: 5th grade, 0–30: College level).
- **FKGL (Flesch-Kincaid Grade Level):** Determines the grade level required to understand text
- **ARI (Automated Readability Index):** Uses characters instead of syllables to assess readability Correlates with grade levels for reading comprehension.

# Metrics Overview: Name and Range

| **Name of Metric** | **Range** |
|---|---|
| Content Uniqueness (CU) | [0, 1] |
| Dissimilarity Index (DSI) | [0, 1] |
| Redundancy Metric (D) | [0, 1] |
| Character Number | Positive integers |
| Word Number | Positive integers |
| Common Word Number | Positive integers |
| Unique Word Number | Positive integers |
| Type-Token Ratio (TTR) | [0, 1] |
| Corrected TTR (CTTR) | [0, 1] |
| Flesch Reading Ease Score (FRES) | 0 to 100 |
| Flesch-Kincaid Grade Level (FKGL) | 0 and above |
| Automated Readability Index (ARI) | 0 and above |
| Mean Sentence Length | Positive integers |
| Syllable Number | Positive integers |
| Sentence Number | Positive integers |

# Metrics Overview: Good Values

| **Name of Metric** | **Good Value** |
|---|---|
| Content Uniqueness (CU) | Close to 1: Highly unique |
| Dissimilarity Index (DSI) | High: Diverse and semantically varied |
| Redundancy Metric (D) | Low: Minimal word repetition |
| Character Number | Depends on the use case: Neither too short/long |
| Word Number | Depends on the use case: Appropriate length |
| Common Word Number | Moderate: Simplifies text for readability |
| Unique Word Number | High: Reflects vocabulary diversity |
| Type-Token Ratio (TTR) | High: High lexical variety |
| Corrected TTR (CTTR) | High: Balanced lexical diversity |
| Flesch Reading Ease Score (FRES) | High (e.g., 60–100): Easy to read |
| Flesch-Kincaid Grade Level (FKGL) | Low (e.g., below 8): Simple, accessible text |
| Automated Readability Index (ARI) | Low (e.g., below 10): Easy to understand |
| Mean Sentence Length | Moderate: Indicates concise, clear sentences |
| Syllable Number | Moderate: Balanced text complexity |
| Sentence Number | Appropriate to content: Balanced paragraphing |

# Metrics Overview: Bad Values

| **Name of Metric** | **Bad Value** |
|---|---|
| Content Uniqueness (CU) | Close to 0: Highly redundant or overlapping |
| Dissimilarity Index (DSI) | Low: Repetitive or homogeneous |
| Redundancy Metric (D) | High: Excessive word repetition |
| Character Number | Extremely short or excessively verbose |
| Word Number | Very short or excessively wordy |
| Common Word Number | Too high: May indicate lack of variety |
| Unique Word Number | Low: May indicate overuse of certain words |
| Type-Token Ratio (TTR) | Low: Limited vocabulary |
| Corrected TTR (CTTR) | Low: Restricted lexical variety |
| Flesch Reading Ease Score (FRES) | Low (e.g., 0–30): Difficult to read |
| Flesch-Kincaid Grade Level (FKGL) | High (e.g., 12+): Complex, advanced text |
| Automated Readability Index (ARI) | High (e.g., 15+): Challenging text |
| Mean Sentence Length | High: Verbose or overly complex sentences |
| Syllable Number | High: Overly complex words, harder to read |
| Sentence Number | Too low or too high: May suggest poor structure |

# Questions 1 to 3

What is the primary difference between syntactic analysis and semantic analysis in the context of natural language processing?

What are the major tasks involved in natural language processing, and how is data typically collected for these tasks?

What is lemmatization and how does it differ from other techniques for reducing words to their normalized form?

# Questions 4 and 5

What is the main difference between stemming and lemmatization in terms of their approach to reducing words to their base form?

What is the difference between anaphora resolution and the more general task of coreference resolution, and how does the concept of ”bridging relationships” fit into coreference resolution?

# CU, CU2, and DSI metrics

| 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: First 10 rows of the data with mean and variance.

# Metrics for NLP Questions

| **Metric** | Q1 | Q2 | Q3 | Q4 | Q5 |
|---|---|---|---|---|---|
| Character Number | 126 | 118 | 111 | 127 | 189 |
| Word Number | 18 | 19 | 18 | 21 | 27 |
| Common Word Number | 0.5 | 0.5789 | 0.6111 | 0.5714 | 0.5185 |
| Unique Word Number | 16 | 19 | 18 | 19 | 22 |
| TTR | 0.8889 | 1.0 | 1.0 | 0.9048 | 0.8148 |
| CTTR | 2.6667 | 3.0822 | 3.0 | 2.9318 | 2.9938 |
| DMetric | 0.0131 | 0.0 | 0.0 | 0.0095 | 0.0171 |
| Syllable Number | 41 | 36 | 31 | 34 | 60 |
| Sentence Number | 1 | 1 | 1 | 1 | 1 |
| Mean Sentence Length | 126.0 | 118.0 | 111.0 | 127.0 | 189.0 |

**Thank you for your attention!**