SQuAD v2 Dataset Analysis Report

(Exploratory Data Analysis Summary)

1. Overview

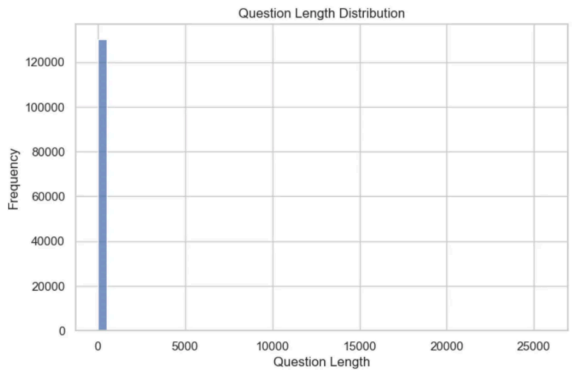
The **Stanford Question Answering Dataset v2 (SQuAD v2)** is a widely used benchmark dataset designed for training and evaluating **question-answering (QA)** models.
It extends the original SQuAD v1 dataset by introducing **unanswerable questions**, making it more challenging and closer to real-world scenarios.

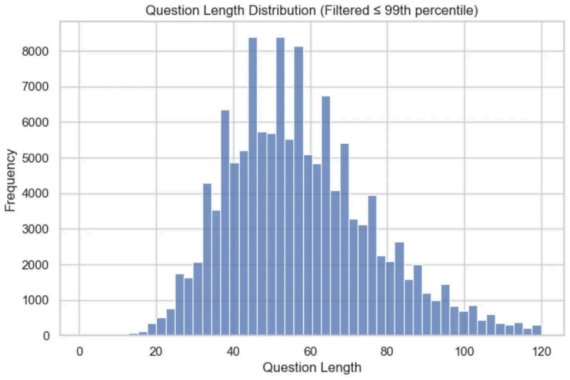
- **Training Samples:** 130,319
- **Validation Samples:** 11,873

2. Dataset Characteristics

2.1. Question Length Statistics

The question length distribution was analyzed to understand the variability and potential outliers in the dataset.
Two histograms below illustrate the differences between the raw distribution and the filtered distribution (excluding the top 1% of extreme cases).





Distribution Analysis

- **Left (Raw Distribution)**
  The raw question length distribution shows a **long right tail** caused by extreme outliers.
  While the majority of questions are short, some exceed **25,000 tokens**, which heavily skews the overall distribution and can affect model training efficiency.

- **Right (Filtered ≤ 99th Percentile)**
  After removing the top 1% of extreme cases, the distribution becomes **much more concentrated**.
  Most questions fall within the **40 to 70 token** range, with a **median of 55 tokens** and a **mean of approximately 58 tokens**.
  This filtered view provides a more realistic representation of the dataset for downstream modeling.

Metric	Value
Total Samples	130,319
Mean	~58 tokens
Median	55 tokens
75th Percentile	69 tokens
Maximum	25,651 tokens *(extreme outlier)*

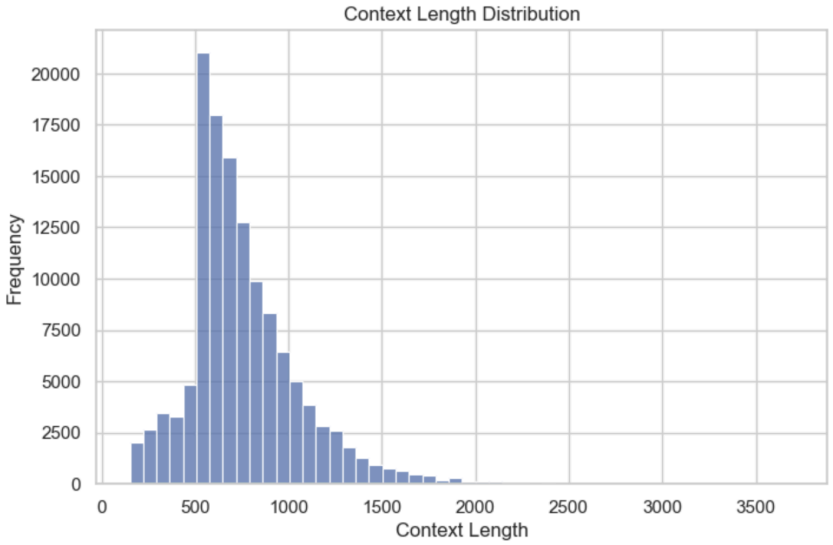
Key Insights

- Most questions are **short and concise**, typically between **40 and 70 tokens**.
- A small portion of questions are **extremely long** and should be treated as **outliers**.
- Filtering these outliers is recommended to:
  - Improve **model stability**
  - Reduce **training overhead**
  - Set an appropriate **maximum input length** for question-answering models.

2.2. Context Length Statistics

The distribution of context passage lengths shows that most contexts are relatively long, with a significant number falling between 500 and 1,000 tokens. However, there is a noticeable long-tail distribution, where a smaller set of contexts exceed 2,000 tokens, and a few extreme outliers reach up to 3,706 tokens.

This indicates that while the majority of contexts are moderately sized, models should be prepared to handle very long passages to ensure optimal performance.



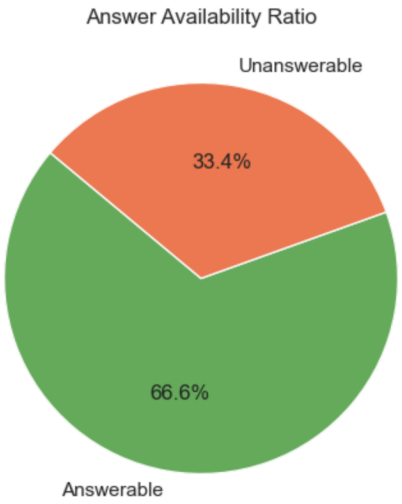
Metric	Value
Mean	~754 tokens
Median	692 tokens
75th Percentile	891 tokens
Maximum	3,706 tokens

Key Insights:

- Most contexts range between **500 and 1,000 tokens**, making them relatively long compared to typical QA datasets.
- There is a **long-tail distribution** where a smaller subset of contexts are significantly longer, exceeding **2,000 tokens**.
- To effectively process the dataset, models should support at least **1,000 tokens of context length** and implement efficient truncation or sliding window strategies for handling extremely long passages.

2.3. Answer Availability

The dataset contains both answerable and unanswerable questions, making it more challenging and realistic compared to traditional QA datasets. Approximately one-third of the dataset consists of unanswerable questions, requiring models to learn not only how to extract answers but also when no valid answer exists.



Category	Count (approx.)	Percentage
Answerable	~86,700	**66.6%**
Unanswerable	~43,600	**33.4%**
Total	130,319	100%

Key Insights:

- About **66.6%** of the dataset contains questions with valid answers, while **33.4%** are intentionally unanswerable.
- This characteristic makes **SQuAD v2** fundamentally different from v1 and requires models to handle **"no answer"** predictions effectively.
- Properly predicting unanswerable cases significantly impacts **F1** and **EM (Exact Match)** scores during evaluation.
- Implementing a **threshold-based confidence mechanism** can help improve model performance when distinguishing between answerable and unanswerable questions.

3. Conclusion

This **exploratory analysis** of the **SQuAD v2** dataset provides detailed insights into its **structural characteristics** and **modeling challenges**.
The dataset is highly suitable for training advanced **Question Answering (QA)** models and is particularly relevant for **Retrieval-Augmented Generation (RAG)** pipelines due to its inclusion of **unanswerable cases**.

By understanding these characteristics, we can make better decisions when:

- Selecting embedding models
- Designing context retrieval strategies
- Optimizing answer extraction pipelines