Implementation and Evaluation of a Retrieval Pipeline

1 Introduction and Approach

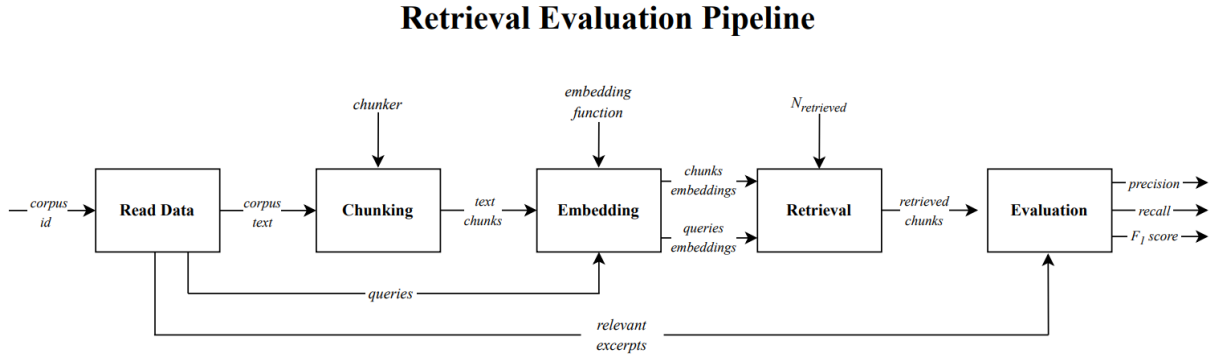
1.1 Project overview

The goal of this project is to implement a retrieval pipeline to retrieve relevant information from a text corpus based on given queries and to evaluate retrieval performance. The dataset and chunking algorithm used, along with the metrics applied for performance evaluation, were taken from the paper: research.trychroma.com/evaluating-chunking. Its codebase can be found at: github.com/brandonstarxel/chunking_evaluation/tree/main/chunking_evaluation/evaluation_framework/general_evaluation_data.

The complete code for this task is available at github.com/ljubenovic/Retrieval-Evaluation-Pipeline.

1.2 Overview of the retrieval pipeline

The steps within the retrieval evaluation pipeline are shown in the diagram below.



The dataset selected for solving this task is *"State of the Union"*, which represents a well-structured and clear transcript of the *2024 State of the Union Address*. This corpus is 10,444 tokens long, measured with Tiktoken for *"cl100k_base"* encoding, the standard used by OpenAI's GPT-4. In addition to the text corpus, the dataset also includes a set of queries relevant to the given document, as well as excerpts from the corpus that act as answers to the queries and will serve as the ground truth.

The *FixedTokenChunker* algorithm was used for chunking the original corpus. It splits the text into chunks of fixed length, with the possibility of overlap between the chunks - meaning that the hyperparameters of this chunker are *chunk_size* (the length of one chunk in tokens) and *chunk_overlap* (the length of overlapping segments between consecutive chunks).

The open-source embedding model used for generating embedding vectors is *"all-MiniLM-L6-v2"* (huggingface.co/sentence-transformers/all-MiniLM-L6-v2), a sentence-transformers model that maps sentences and paragraphs to a 384-dimensional dense vector space. It is commonly used for tasks such as clustering and semantic search.

After constructing the embeddings for each of the text chunks and all queries, retrieval was performed based on the cosine similarity between the chunks and individual queries. Cosine similarity measures the cosine of the angle between two vectors in the embedding space, indicating how similar the two vectors are. It is defined as:

$$cosine\ similarity(\boldsymbol{A}, \boldsymbol{B}) = \frac{\boldsymbol{A} \cdot \boldsymbol{B}}{\|\boldsymbol{A}\|\|\boldsymbol{B}\|}$$

where $\boldsymbol{A}$ and $\boldsymbol{B}$ are the vectors representing the chunk and the query, and $\|\boldsymbol{A}\|$ and $\|\boldsymbol{B}\|$ are their respective magnitudes. For each query, the top N_r chunks with the highest cosine similarity scores were returned, with N_r being a hyperparameter of the model.

The evaluation of the performance of this retrieval pipeline was conducted based on the values of precision and recall metrics, which were averaged across all queries.

However, since precision and recall are often in conflict with each other - meaning that optimizing one can often lead to a decrease in the other - a third metric, the F1 score, was also used. The F1 score represents the harmonic mean of precision and recall, providing a balanced measure that considers both metrics simultaneously.

1.3 Evaluation metrics

In an ideal case, the retrieval system should retrieve only the relevant tokens for each query. In practice, a retrieval unit is most often one or more chunks of text that contain segments considered relevant for the given queries. This means that the retrieved chunk, in addition to relevant tokens, will often also contain irrelevant parts, which can decrease the overall performance of the retrieval system. Other challenges include the possibility of the system returning redundant tokens when chunks overlap, and the potential for some relevant excerpts to be missed in the retrieval process.

Therefore, the precision and recall metrics are defined to assess how well a retrieval system performs. The precision metric (also known as positive predictive value, PPV) represents the proportion of relevant retrieved tokens (true positives, TP) among all tokens retrieved by the

system. On the other hand, the recall metric (also referred to as sensitivity, hit rate, or true positive rate, TPR) indicates the proportion of relevant tokens that were successfully retrieved out of all relevant tokens in the corpus. The formulas used to compute precision and recall for a given query q over the corpus $\boldsymbol{C}$ are presented below. Here, t_e denotes the set of all tokens that are part of the relevant excerpt(s), while t_r denotes the set of all tokens that are retrieved by the system.

$$Precision_q(\boldsymbol{C}) = \frac{|t_e \cap t_r|}{|t_r|}$$

$$Recall_q(\boldsymbol{C}) = \frac{|t_e \cap t_r|}{|t_e|}$$

The F1 score, being the harmonic mean of precision and recall, is calculated using the equation given below.

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} = 2 \times \frac{|t_e \cap t_r|}{|t_r|}$$

1.4 Hyperparameter tuning

As previously mentioned, the hyperparameters of the retrieval pipeline include the chunk size and overlap length used by the *FixedTokenChunker* algorithm, as well as the parameter N_r, which specifies the number of chunks the retrieval system returns for each query.

To optimize these hyperparameters for a specific metric (precision, recall, or F1 score), a grid search was conducted over all possible combinations of parameter values from the following table, resulting in a total of 125 combinations.

chunk_size	overlap_size	Nr
100	10	1
200	20	3
300	30	5
400	40	7
500	50	9

2 Results

2.1 Experiment results table

chunk_size	chunk_overlap	Nr	precision	recall	f1
100	10	1	19.43 ± 19.48	52.59 ± 43.71	26.53 ± 23.74
100	10	3	10.36 ± 7.13	80.82 ± 34.33	17.76 ± 11.02
100	10	5	6.81 ± 5.29	85.74 ± 31.26	12.28 ± 8.64
100	10	7	5.41 ± 3.81	91.93 ± 23.54	10.01 ± 6.51
100	10	9	4.27 ± 2.89	94.60 ± 17.93	8.05 ± 5.09
100	20	1	18.97 ± 19.26	50.16 ± 44.48	25.86 ± 23.97
100	20	3	10.38 ± 8.42	75.07 ± 38.84	17.57 ± 12.75
100	20	5	7.21 ± 5.41	86.50 ± 29.49	12.95 ± 8.76
100	20	7	5.29 ± 3.89	89.00 ± 27.32	9.78 ± 6.65
100	20	9	4.25 ± 3.02	92.38 ± 24.25	8.00 ± 5.33
100	30	1	22.79 ± 20.86	60.16 ± 43.30	30.85 ± 24.68
100	30	3	10.49 ± 8.89	79.33 ± 38.20	17.79 ± 13.34
100	30	5	7.00 ± 5.19	87.70 ± 30.28	12.62 ± 8.50
100	30	7	5.31 ± 3.73	91.19 ± 26.07	9.83 ± 6.41
100	30	9	4.34 ± 3.07	93.90 ± 22.80	8.15 ± 5.42
100	40	1	20.24 ± 18.54	54.70 ± 42.93	27.85 ± 23.16
100	40	3	10.31 ± 7.86	78.54 ± 34.61	17.57 ± 11.81
100	40	5	6.99 ± 5.23	87.49 ± 28.03	12.59 ± 8.48
100	40	7	5.50 ± 3.79	94.13 ± 18.26	10.17 ± 6.44
100	40	9	4.34 ± 2.97	95.11 ± 17.52	8.17 ± 5.22
100	50	1	22.25 ± 18.99	62.56 ± 42.46	30.88 ± 23.34
100	50	3	10.26 ± 7.60	79.84 ± 34.81	17.52 ± 11.58
100	50	5	6.84 ± 5.44	85.66 ± 31.77	12.32 ± 8.88
100	50	7	5.38 ± 3.71	93.66 ± 20.42	9.97 ± 6.32
100	50	9	4.38 ± 2.97	96.21 ± 15.73	8.24 ± 5.23
200	20	1	11.10 ± 10.35	56.72 ± 44.92	17.78 ± 15.63
200	20	3	5.99 ± 4.64	87.67 ± 31.11	10.94 ± 7.81
200	20	5	3.84 ± 2.74	92.82 ± 23.28	7.26 ± 4.89
200	20	7	2.84 ± 1.90	96.25 ± 16.97	5.45 ± 3.50
200	20	9	2.23 ± 1.46	97.83 ± 11.93	4.33 ± 2.74
200	40	1	11.83 ± 9.89	63.15 ± 44.39	19.21 ± 15.30
200	40	3	5.99 ± 4.70	86.61 ± 31.78	10.94 ± 7.90
200	40	5	3.89 ± 2.74	94.26 ± 21.16	7.35 ± 4.90
200	40	7	2.87 ± 1.89	97.91 ± 12.20	5.51 ± 3.48
200	40	9	2.25 ± 1.47	98.68 ± 11.47	4.36 ± 2.76
200	60	1	12.61 ± 11.91	63.39 ± 44.22	19.97 ± 16.93
200	60	3	5.73 ± 4.68	83.61 ± 34.26	10.46 ± 7.88
200	60	5	3.70 ± 2.67	92.39 ± 22.96	7.01 ± 4.76
200	60	7	2.81 ± 1.92	96.40 ± 17.02	5.39 ± 3.52
200	60	9	2.22 ± 1.46	98.25 ± 12.02	4.31 ± 2.74
200	80	1	13.18 ± 12.37	65.84 ± 44.77	20.93 ± 17.50

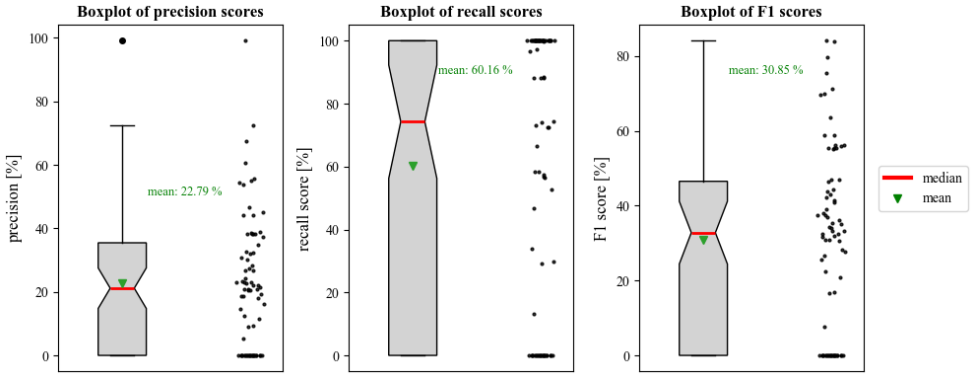
chunk_size	chunk_overlap	Nr	precision	recall	f1
200	80	3	6.16 ± 4.79	87.36 ± 31.75	11.24 ± 8.08
200	80	5	3.91 ± 2.74	94.59 ± 21.05	7.39 ± 4.89
200	80	7	2.84 ± 1.93	96.31 ± 18.42	5.46 ± 3.56
200	80	9	2.22 ± 1.51	96.31 ± 18.42	4.29 ± 2.82
200	100	1	14.42 ± 12.87	71.70 ± 40.64	22.79 ± 17.48
200	100	3	6.17 ± 4.44	90.44 ± 25.59	11.29 ± 7.44
200	100	5	3.96 ± 2.60	97.32 ± 14.81	7.51 ± 4.64
200	100	7	2.86 ± 1.91	97.63 ± 14.61	5.49 ± 3.50
200	100	9	2.28 ± 1.45	100.00 ± 0.00	4.42 ± 2.71
300	30	1	9.01 ± 8.95	67.40 ± 44.80	15.26 ± 13.63
300	30	3	4.09 ± 3.10	88.51 ± 30.02	7.69 ± 5.49
300	30	5	2.61 ± 1.78	95.52 ± 18.25	5.02 ± 3.29
300	30	7	1.91 ± 1.27	96.84 ± 14.49	3.71 ± 2.39
300	30	9	1.49 ± 0.97	98.16 ± 9.13	2.92 ± 1.86
300	60	1	8.81 ± 8.84	64.76 ± 46.74	14.90 ± 13.71
300	60	3	4.10 ± 3.10	90.55 ± 27.81	7.70 ± 5.48
300	60	5	2.65 ± 1.77	96.73 ± 16.94	5.10 ± 3.28
300	60	7	1.90 ± 1.26	97.37 ± 16.11	3.70 ± 2.39
300	60	9	1.52 ± 0.96	100.00 ± 0.00	2.98 ± 1.84
300	90	1	9.24 ± 7.43	74.72 ± 42.43	15.95 ± 11.81
300	90	3	4.12 ± 3.10	89.74 ± 29.19	7.73 ± 5.50
300	90	5	2.69 ± 1.77	98.01 ± 12.00	5.19 ± 3.27
300	90	7	1.95 ± 1.25	99.65 ± 2.50	3.80 ± 2.36
300	90	9	1.53 ± 0.97	99.93 ± 0.58	2.99 ± 1.86
300	120	1	9.84 ± 8.66	73.90 ± 40.84	16.69 ± 13.03
300	120	3	4.24 ± 3.08	93.93 ± 22.83	7.98 ± 5.44
300	120	5	2.66 ± 1.78	97.07 ± 14.13	5.12 ± 3.28
300	120	7	1.94 ± 1.26	98.83 ± 7.67	3.77 ± 2.37
300	120	9	1.51 ± 0.98	98.83 ± 7.67	2.95 ± 1.87
300	150	1	11.52 ± 9.54	82.32 ± 35.97	19.40 ± 14.01
300	150	3	4.36 ± 2.97	95.64 ± 18.05	8.20 ± 5.23
300	150	5	2.63 ± 1.81	95.95 ± 17.92	5.07 ± 3.34
300	150	7	1.93 ± 1.26	98.65 ± 8.29	3.76 ± 2.37
300	150	9	1.50 ± 0.98	98.65 ± 8.29	2.94 ± 1.87
400	40	1	6.83 ± 7.48	66.67 ± 47.06	11.94 ± 11.94
400	40	3	2.80 ± 2.40	83.50 ± 36.82	5.35 ± 4.39
400	40	5	1.76 ± 1.40	87.45 ± 32.74	3.43 ± 2.65
400	40	7	1.33 ± 0.98	92.12 ± 26.99	2.60 ± 1.87
400	40	9	1.05 ± 0.75	93.43 ± 24.79	2.06 ± 1.45
400	80	1	5.81 ± 6.96	61.62 ± 47.56	10.21 ± 11.05
400	80	3	2.78 ± 2.40	82.03 ± 37.04	5.31 ± 4.39

chunk_size	chunk_overlap	Nr	precision	recall	f1
400	80	5	1.76 ± 1.40	85.90 ± 33.23	3.43 ± 2.66
400	80	7	1.31 ± 1.00	88.89 ± 30.07	2.57 ± 1.92
400	80	9	1.03 ± 0.77	90.21 ± 28.26	2.03 ± 1.49
400	120	1	6.81 ± 6.62	66.86 ± 44.63	12.00 ± 10.76
400	120	3	3.00 ± 2.39	86.92 ± 32.24	5.72 ± 4.36
400	120	5	2.02 ± 1.31	98.33 ± 9.10	3.92 ± 2.47
400	120	7	1.44 ± 0.94	98.33 ± 9.10	2.83 ± 1.79
400	120	9	1.13 ± 0.73	98.48 ± 9.02	2.22 ± 1.41
400	160	1	6.87 ± 6.82	67.50 ± 45.61	12.05 ± 11.07
400	160	3	3.10 ± 2.31	91.27 ± 27.19	5.91 ± 4.20
400	160	5	2.01 ± 1.31	97.88 ± 13.37	3.90 ± 2.47
400	160	7	1.43 ± 0.94	97.88 ± 13.37	2.81 ± 1.80
400	160	9	1.14 ± 0.72	100.00 ± 0.00	2.24 ± 1.40
400	200	1	7.86 ± 7.08	78.09 ± 40.27	13.79 ± 11.20
400	200	3	3.25 ± 2.30	94.26 ± 21.26	6.21 ± 4.18
400	200	5	2.05 ± 1.31	100.00 ± 0.00	3.98 ± 2.47
400	200	7	1.47 ± 0.94	100.00 ± 0.00	2.88 ± 1.80
400	200	9	1.14 ± 0.73	100.00 ± 0.00	2.24 ± 1.40
500	50	1	5.34 ± 5.35	63.53 ± 46.79	9.60 ± 9.15
500	50	3	2.05 ± 1.71	75.95 ± 41.14	3.95 ± 3.23
500	50	5	1.32 ± 1.09	81.25 ± 38.31	2.58 ± 2.10
500	50	7	0.99 ± 0.76	85.20 ± 34.67	1.94 ± 1.48
500	50	9	0.78 ± 0.58	87.80 ± 31.82	1.54 ± 1.14
500	100	1	4.76 ± 5.57	58.79 ± 48.49	8.57 ± 9.39
500	100	3	2.33 ± 2.03	79.60 ± 40.11	4.49 ± 3.79
500	100	5	1.43 ± 1.20	82.23 ± 38.00	2.79 ± 2.30
500	100	7	1.06 ± 0.83	87.50 ± 32.74	2.08 ± 1.61
500	100	9	0.85 ± 0.62	92.73 ± 25.49	1.68 ± 1.21
500	150	1	4.75 ± 5.12	57.73 ± 46.77	8.55 ± 8.68
500	150	3	2.13 ± 1.77	79.77 ± 37.09	4.11 ± 3.30
500	150	5	1.37 ± 1.06	85.96 ± 32.04	2.68 ± 2.03
500	150	7	1.01 ± 0.75	88.08 ± 29.99	1.99 ± 1.46
500	150	9	0.84 ± 0.57	92.95 ± 22.97	1.65 ± 1.11
500	200	1	5.96 ± 5.40	73.03 ± 41.48	10.71 ± 9.00
500	200	3	2.36 ± 1.81	85.42 ± 31.90	4.54 ± 3.38
500	200	5	1.54 ± 1.12	91.98 ± 24.12	3.01 ± 2.13
500	200	7	1.13 ± 0.78	94.96 ± 20.45	2.22 ± 1.51
500	200	9	0.88 ± 0.61	94.96 ± 20.45	1.73 ± 1.18
500	250	1	5.97 ± 4.95	75.55 ± 41.40	10.80 ± 8.39
500	250	3	2.58 ± 1.82	93.44 ± 22.88	4.97 ± 3.36
500	250	5	1.59 ± 1.06	97.68 ± 12.45	3.12 ± 2.02
500	250	7	1.16 ± 0.75	99.44 ± 3.45	2.29 ± 1.45
500	250	9	0.90 ± 0.58	99.44 ± 3.45	1.78 ± 1.13

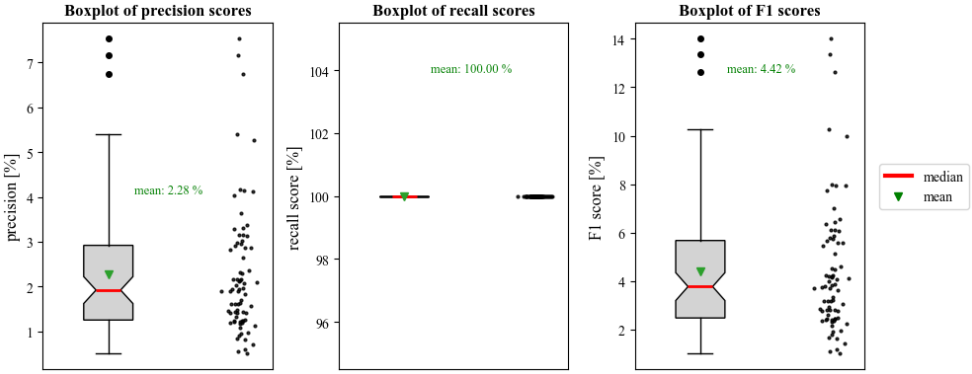
The highlighted rows represent the hyperparameter combinations that yielded the maximum value for one of the evaluated metrics.

2.2 Boxplots for best hyperparameter combinations

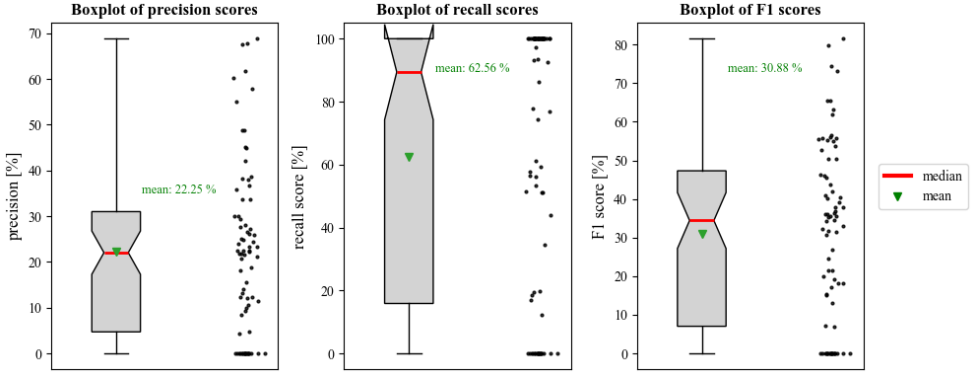
The following plots display the distribution of precision, recall, and F1 scores corresponding to individual queries for the hyperparameter combination that resulted in the highest precision score (*chunk_size* = 100, *chunk_overlap* = 30 %, N_r = 1).



The same plots were also generated for the hyperparameter combination that resulted in the highest recall (*chunk_size* = 200, *chunk_overlap* = 50 %, N_r = 9), and they are shown below.



For the parameter combination that maximizes the F1 score (*chunk_size* = 100, *chunk_overlap* = 50 %, N_r = 1), the following boxplots were obtained.



3 Discussion

3.1 Results analysis

As expected, the precision and recall metrics cannot be simultaneously maximized, which has been confirmed by the tabular evaluation results shown earlier. As the precision score increases, the recall decreases, and vice versa. This is the primary reason why the F1 score metric was included in the analysis – as the harmonic mean of these metrics, the F1 score can serve to select a good compromise solution that results in both a reasonably high precision score and a reasonably high recall score.

In this specific case, for the analyzed dataset and given queries, the combination of hyperparameters that resulted in the highest F1 score (*chunk_size* = 100, *chunk_overlap* = 50 %, N_r = 1) could be regarded as optimal when compared to all the combinations considered. However, in practice, the parameters considered optimal depend on which metric is more important to maximize (which carries more weight), and this depends on the specific use of the system.

The reason why one metric (precision) increases while the other (recall) decreases in this specific problem can be explained as follows. Since recall represents the proportion of relevant tokens that are retrieved, and because the *FixedTokenChunker* algorithm was used for chunking (which does not take into account the semantic meaning of the text but rather divides it based on positions within the corpus – it's not "intelligent"), a higher likelihood of retrieving more relevant tokens can be achieved by increasing the number of returned chunks. However, the more chunks that are returned, the more retrieved tokens are included, which leads to a decrease in precision.

This has been confirmed by the results shown in the table. For the same values of *chunk_size* and *chunk_overlap*, recall increases as the number of returned chunks grows, and it reaches its maximum for the highest N_r analyzed. On the other hand, for the same *chunk_size* and *chunk_overlap* values, the precision score consistently decreases as N_r increases, being highest when N_r is lowest.

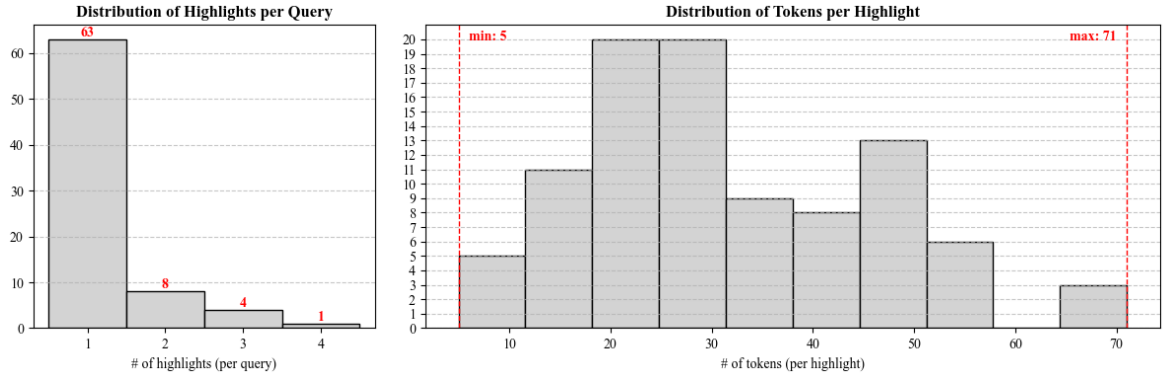
Certainly, the number of returned tokens depends not only on the number of returned chunks but also on the length of each chunk. Therefore, the precision score significantly decreases as the product *chunk_size* x N_r increases. The precision score is highest when these parameters take their minimum considered values, which are *chunk_size* = 100 and N_r = 1.

3.2 Analysis of ground truth excerpts

The reason chunk lengths shorter than 100 tokens were not analyzed is that when setting the chunk length, it must be ensured that the chunk is large enough to cover the entire relevant excerpt (especially when the number of retrieved chunks, N_r, is small).

Additionally, it's important to consider the possibility that for a single query, multiple relevant excerpts could be returned, which may be spatially distant within the text.

For this reason, an analysis of the ground truth excerpts was conducted, and the resulting data is shown in the following histograms.



For each query, the number of relevant excerpts (highlights) returned was determined. As it can be seen, this number is not always 1, which is why the *chunk_size* parameter should not be too small.

Additionally, the length of each relevant excerpt in tokens was determined. Based on the results shown in the histogram above, it can be concluded that there are relevant excerpts that are significantly longer than others (the longest containing 71 tokens). Therefore, it makes sense to assume that the *chunk_size* parameter should not have values significantly smaller than 100 tokens, especially if N_r is small.