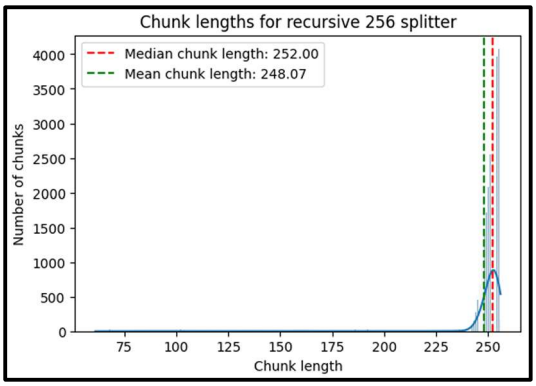
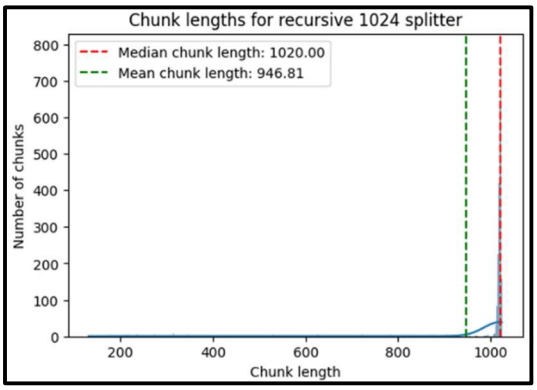
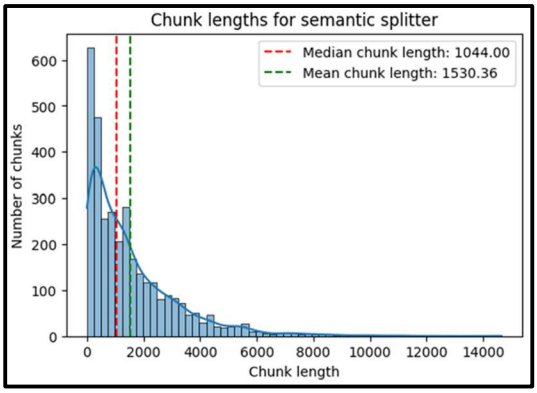
a. Analyzing the Chunks

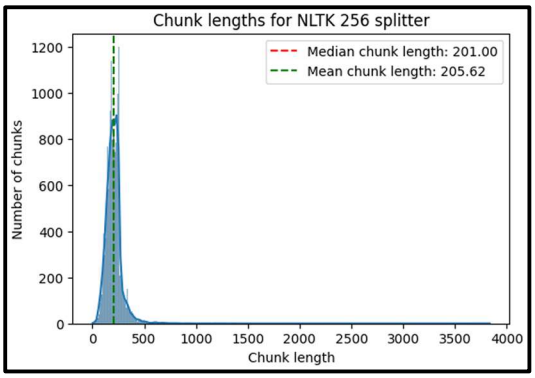


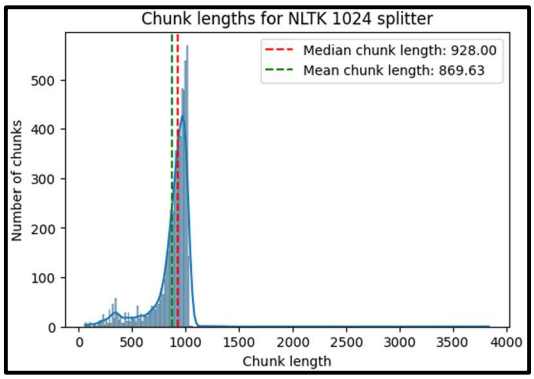


Overall, both graphs are significantly right-skewed, indicating that both configurations effectively yield maximum chunk sizes. The Recursive 256 splitter shows a stronger concentration around the 250 threshold, demonstrating chunk length consistency. On the other hand, the Recursive 1024 splitter has a much wider gap due to longer chunk sizes. The anomaly in the 1024 splitter occurs most likely due to the end of articles, where the splitter takes whatever text is left, resulting in smaller chunks.

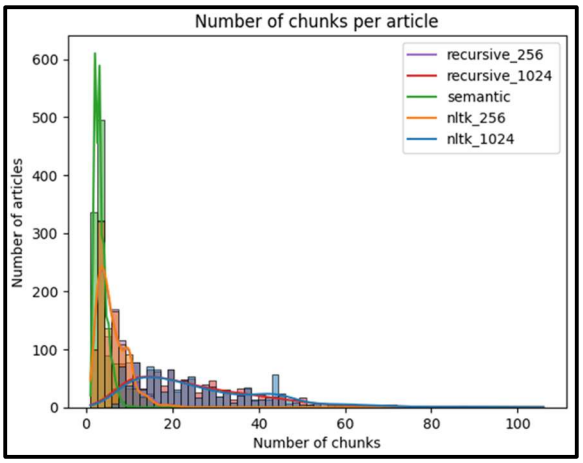


Semantic splitter yields a wider distribution of chunk lengths. This is because semantic splitter prioritizes full sentence/paragraph splits, resulting greater variation in chunk lengths, as there is no cap limit. This exhibits by the difference between mean and median chunk lengths. Using semantic splitter will cause high volatility in chunk lengths, but the trade-off is preserving the full meaning of a text, rather than splitting it halfway through.



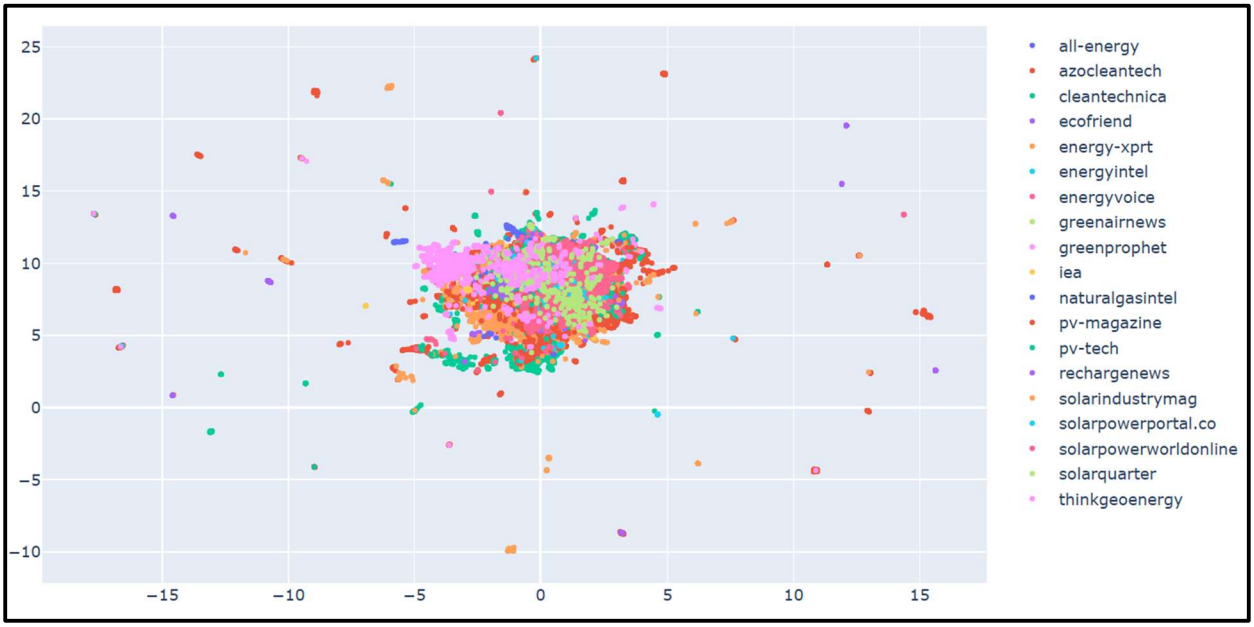


The NLTK splitter behaves like a combination of the recursive and semantic splitters. For both chunk lengths, it achieves a high concentration at the given length split. However, since NLTK splitter also prioritizes full sentence/paragraph splits, hence resulting a wide variation of chunk lengths (up to 4,000). Overall, the NLTK splitter does well at keeping splits close to the given limits while preserving the integrity of the entire text. This represents a more moderate trade-off compared to the semantic splitter.

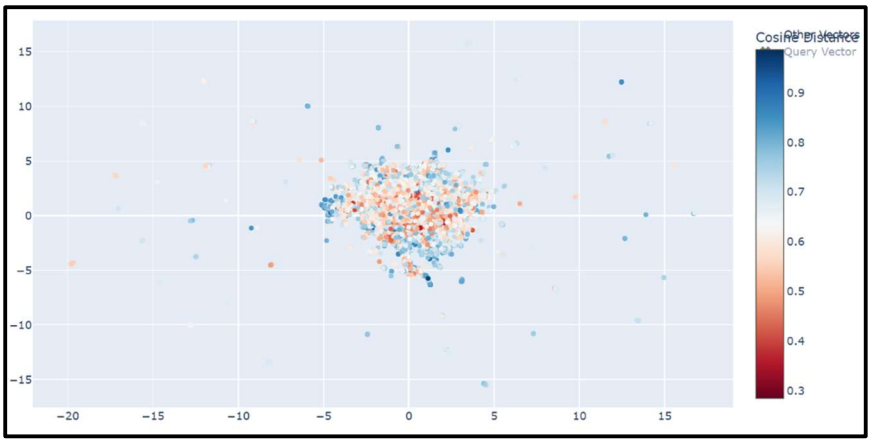


Recursive and NLTK splitter pretty much have similar distribution in the number of chunks per article for their respective chunk lengths. On the other hand, semantic splitter has a denser distribution due to no numerical cap limit, focusing solely on preserving text content. Based on this graph and previous analysis, NLTK splitter is the most promising one. It is well balanced between the objectives of semantic and recursive splitters, while maintaining a moderate distribution after chunking.

b. Analyzing the Embedding Space



Overall, there is significant semantic overlap between the domains, which aligns with the nature of the dataset as it focuses on clean technologies. All articles fall under the same broader umbrella, resulting in a dense central cluster. While there are some notable outliers from domains like ‘solarindustrymag,’ ‘pv-magazine,’ and ‘rechargenews,’ these deviations are not that great. This plot indicates that the dataset is highly homogeneous in terms of content, great overlap; and the chosen embedding model effectively captures this semantic consistency within the dataset.

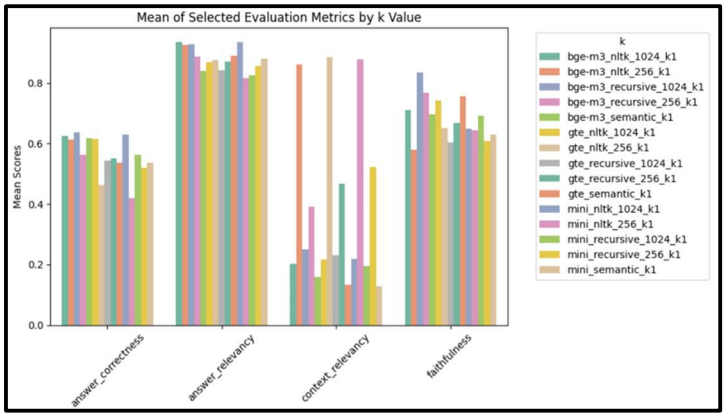


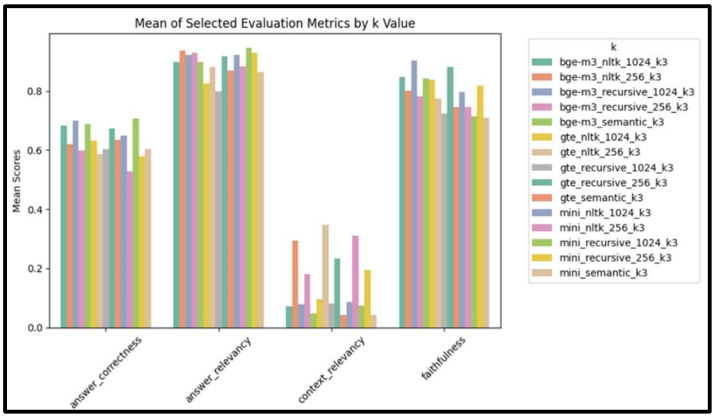
Using the query 'Clean Energy', the plot shows that the theme of this dataset is correctly represented at the center, accurately reflecting that 'Clean Energy' is semantically aligned with the majority of the articles. Additionally, there are 5 articles identified as the closest matches to 'Clean Energy' based on Cosine distance.

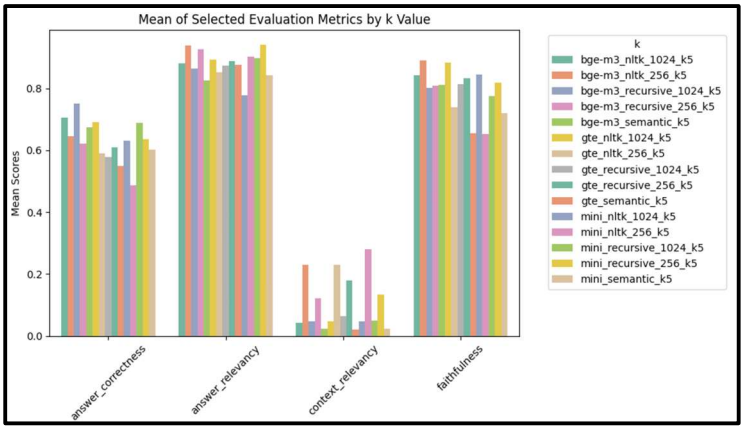
Plot of the Cosine distance between the query 'Clean Energy' and all the articles reveals that the majority of articles located in the center of the cluster, are semantically similar and well-aligned with the query. The darker colors represent articles that are less similar to the query, but these are sparse and not significant at all, indicating the articles are semantically cohesive.

c. In the RAG pipeline, experiment number of chunks k = [1, 3, 5]

Mean RAGAS



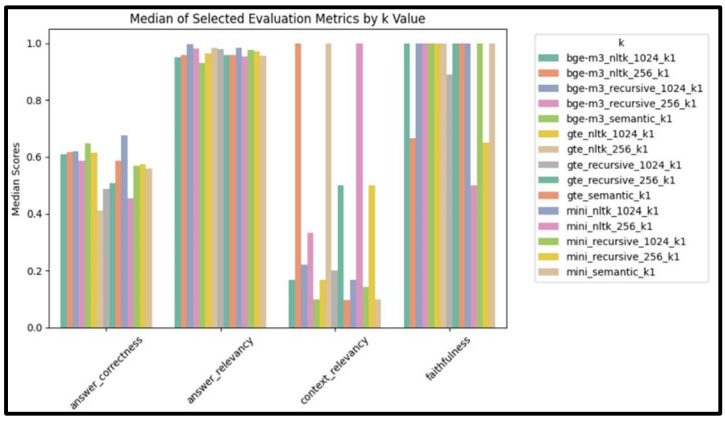


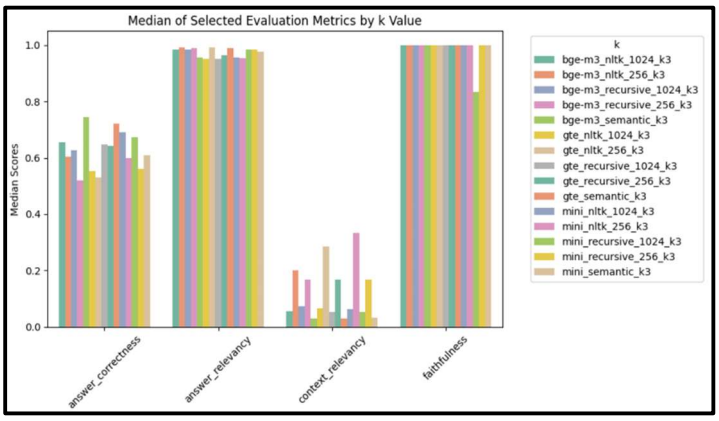


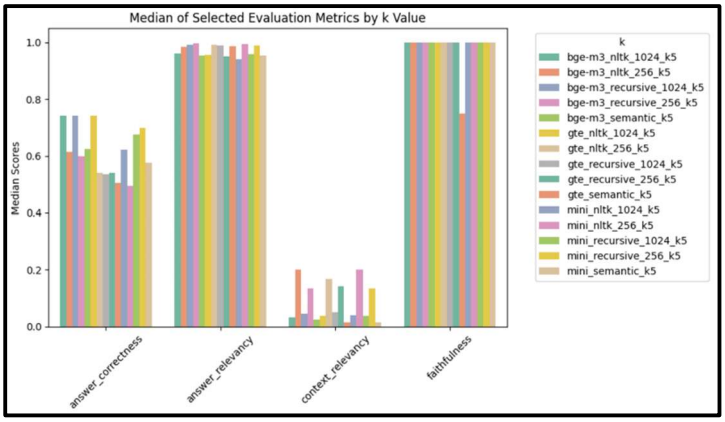
In terms of mean, the metrics answer correctness, answer relevancy, and faithfulness are consistently high across all configurations, showing stable performance across the models. However, context relevancy scores are lower and more variable, especially for semantic splitters

and larger chunk sizes. This suggests difficulties in maintaining detailed context when retrieving over broader scopes. Models using “bge-m3” and “gte” embeddings perform better than “mini”, which is expected due to their ability to handle more complexity. Based on these three plots, and focusing on context relevancy, the NLTK splitter with a 256-chunk size performs best.

Median RAGAS





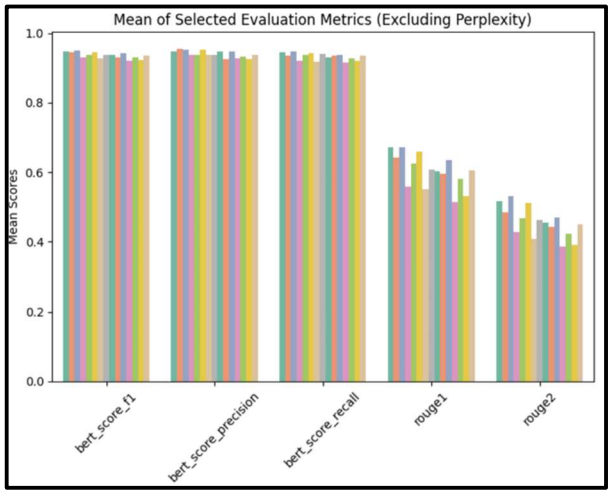


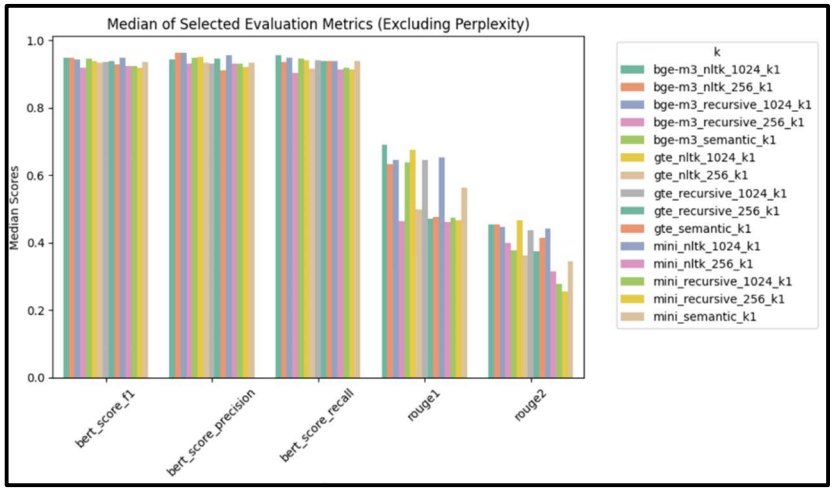
In terms of median, the distribution pattern is similar to that of the mean. Context relevancy shows more variation, particularly with k=1, indicating that while the majority of retrievals perform well, there are some outliers or edge cases where maintaining detailed context becomes more challenging. This is particularly evident when using NLTK or semantic splitters, which result in different chunk lengths. Overall, NLTK with 256 chunks still provides the most reliable performance regardless of the embedding model used.

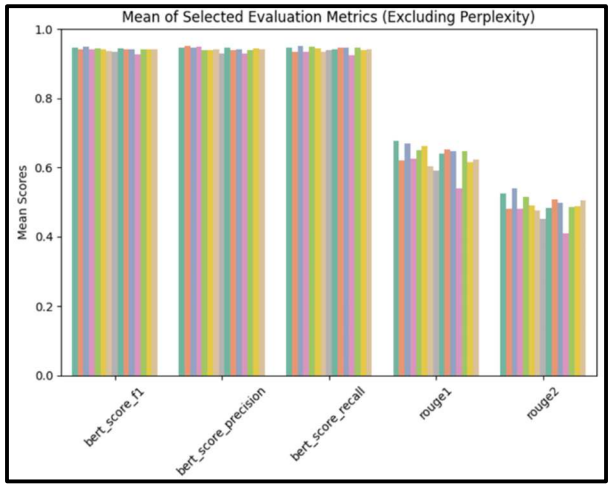
Overall, the four metrics evaluated are answer correctness, which measures if the retrieved answers are accurate; answer relevancy, which checks how relevant the answers are to the query; context relevancy, which assesses if the retrieved context matches the query well; and faithfulness, which evaluates how faithfully the answer aligns with the provided context. Comparing the mean and median across the six plots, both show similar patterns for answer correctness, answer relevancy, and faithfulness, with consistently high scores. However, context relevancy shows more variability, particularly in the mean graphs, which are influenced by outliers. This highlights that while most retrievals perform well, there are occasional edge cases, particularly with larger chunk sizes or semantic splitters, where maintaining detailed context is more challenging.

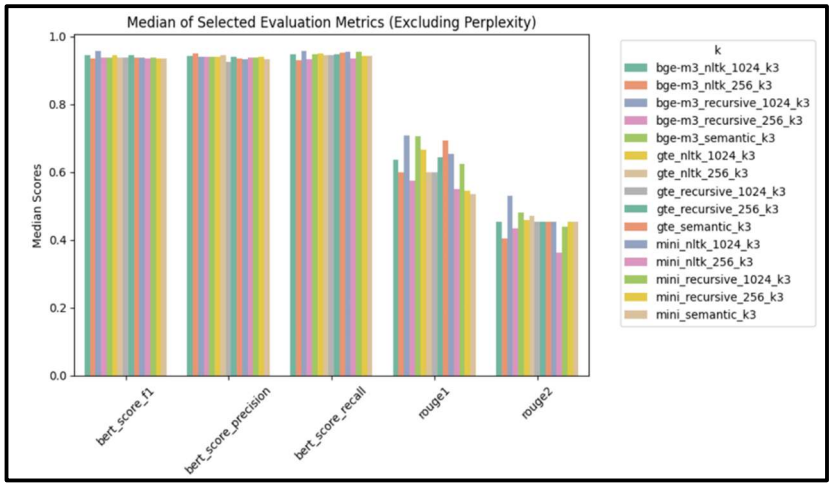
d. In addition to RAGAS, apply additional metrics: Rouge, Perplexity and Bert score.

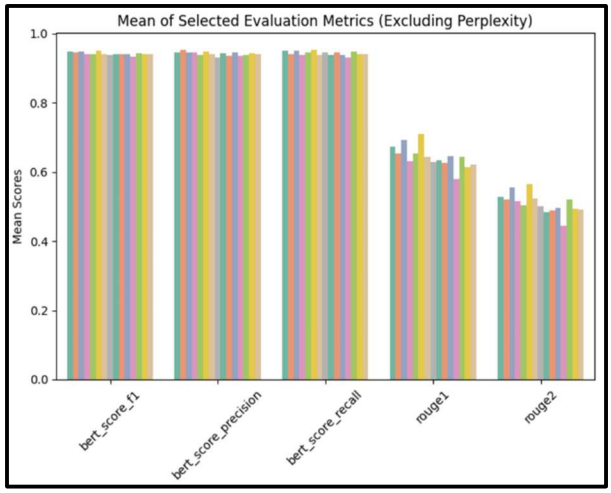
Rouge and Bert scores

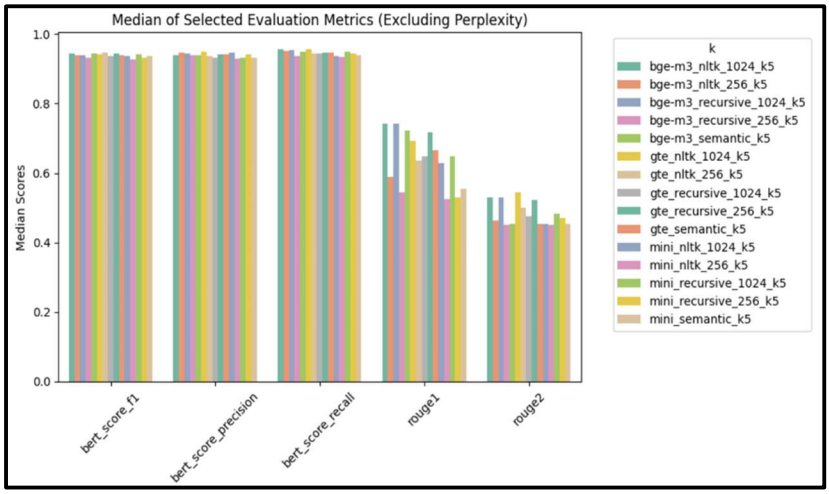








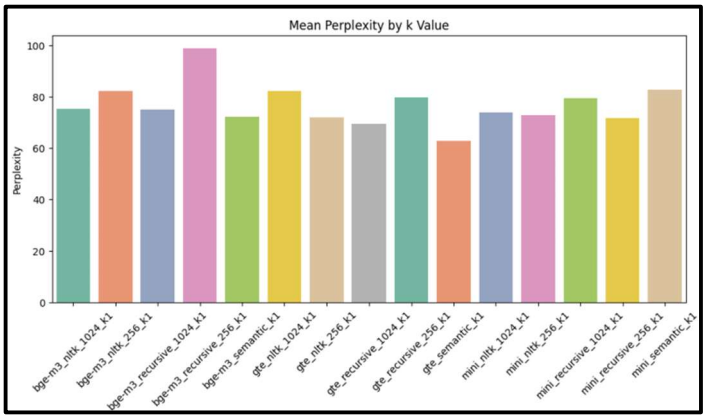


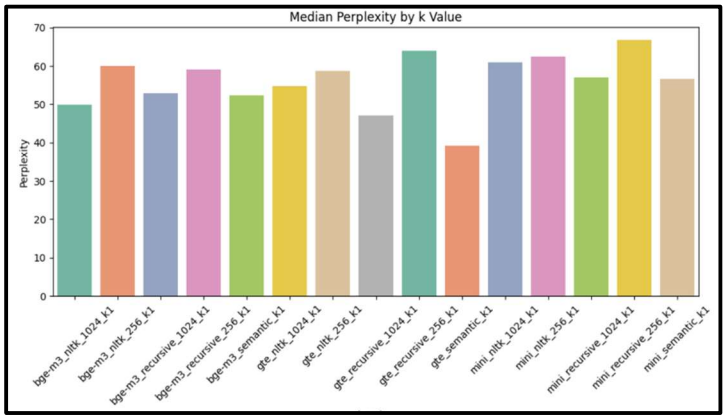


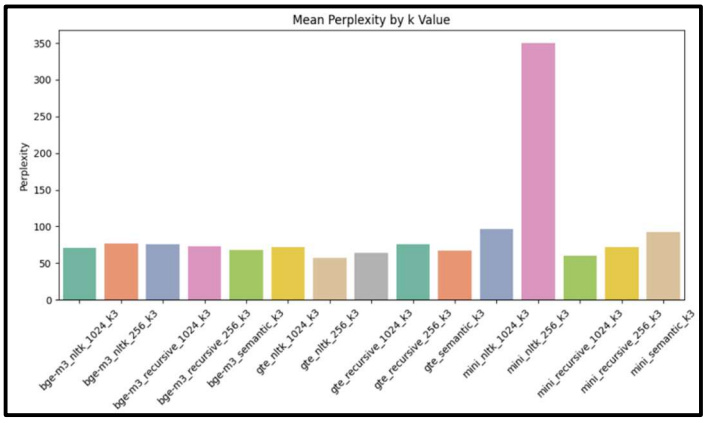
The mean and median plots for Bert scores (F1, precision, recall) and Rouge scores (Rouge1, Rouge2) show consistently high values for the Bert metrics, indicating the strong semantic alignment and quality of generated answers across all configurations. The Bert scores remain close to 1.0 for all k values, highlighting that the embeddings and chunking strategies effectively retrieve semantically accurate and relevant information. However, Rouge scores show more variation, particularly with Rouge2, which measures bigram overlap. This indicates that while semantic accuracy is high, the exact word or phrase matches measured by Rouge are more challenging, especially for larger chunk sizes and semantic splitters.

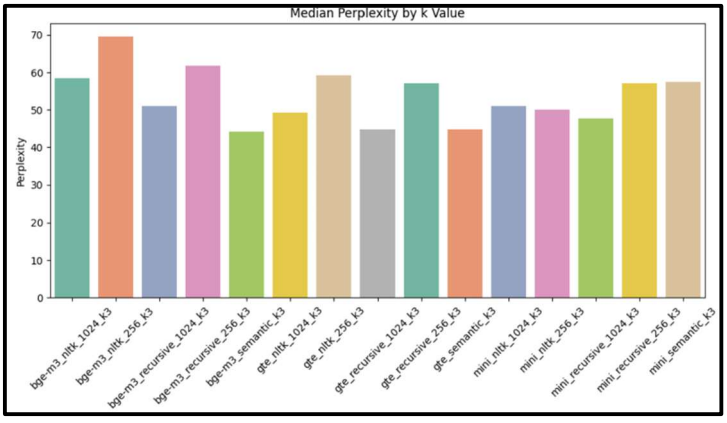
Comparing the mean and median plots, the trends are consistent, but the median plots highlight more stability across metrics, as they are less influenced by outliers. For Bert scores, the median values are nearly identical to the mean, reinforcing the reliability of semantic relevance across the board. For Rouge scores, the median plots show slightly higher consistency, particularly in Rouge1, as the mean values are more affected by occasional lower-scoring outliers. This suggests that while the models generally perform well, there are a few cases where the lexical overlap deviates more significantly from the semantic alignment. This emphasizes the importance of using both types of metrics to evaluate performance comprehensively. Based on these score, NLTK splitter with 256 chunks still perform best, slightly lower score compared to 1024 chunks but not that significant. Specifically, pairing with “bge-m3” embeddings.

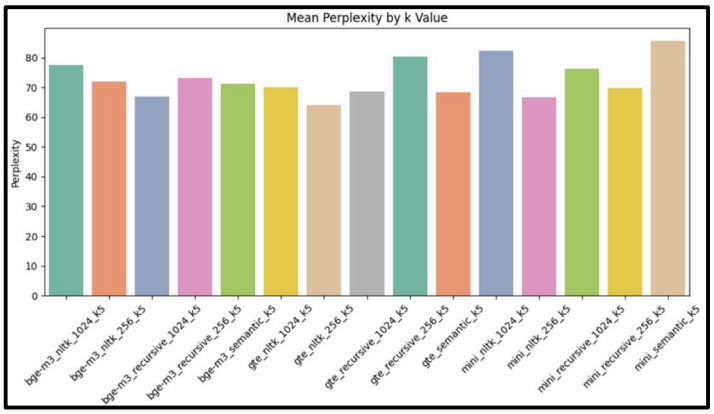
Perplexity

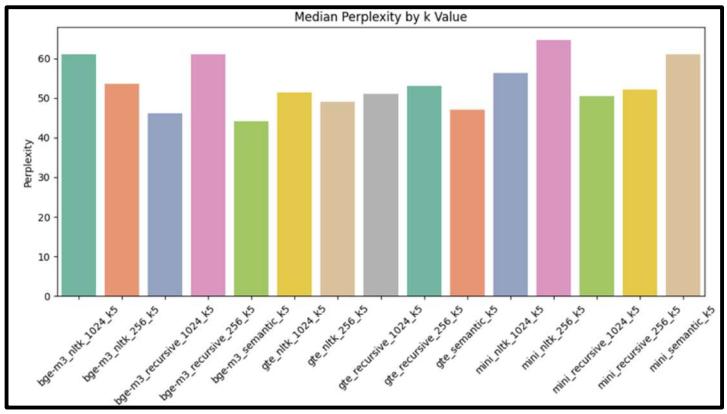












Perplexity is a measure of how well a language model predicts a sample, with lower values indicating better performance. Across the three plots, embedding models “bge-m3” and “gte” consistently exhibit lower perplexity for both mean and median, suggesting stronger predictive capabilities compared to “mini.” Semantic splitters generally perform well across all configurations, showing consistent results. However, for k=3, some configurations exhibit poor performance, with notable spikes in perplexity. This highlights the reliability of the NLTK splitter with a 256-chunk size, which outperforms other chunking strategies across all metrics. Among embedding models, “bge-m3” is a better choice due to its complexity and ability to handle nuanced text, though “mini” may suffice for simpler tasks.