**LLM embedding based clustering:**

**Method and models**

BERT embeddings were extracted from the BERT model, which was trained as a transformer-based bidirectional encoder on BookCorpus [23] and English Wikipedia [24]. These embeddings were leveraged for deep contextual understanding, capturing the nuances in semantic meanings across the corpus.

Best embedding model is “textembedding-ada-002” [25] embeddings since they demonstrated the best results among OpenAI’s embeddings on larger datasets for text search, code search, and sentence similarity tasks.

Additionally, other LLM embeddings, computed for Falcon [26] and LLaMA-2-chat [27] models, were included for their respective advancements in performance and efficiency. It optimized auto-regressive transformer—underwent targeted fine-tuning for dialogue and question-and-answer tasks. BERT, Falcon, and LLaMA-2-chat embeddings were obtained from Hugging Face’s transformers library,

The models used in summarization are described ,As an alternative approach to LLM-based models, we used the summarization model BERT-large-uncased.

The combination of the kmeans algorithm and OpenAI’s embeddings yielded the highest values of ARI, F1-score, and HS in most experiments. This may be attributed to OpenAI’s embeddings being trained on a diverse array of Internet text, rendering them highly effective at capturing the nuances of language structures.

In the domain of open-source models, namely Falcon, LLaMA-2, and BERT, the latter emerged as the frontrunner. Given that BERT is designed to understand the context and potentially due to the model’s lower dimensionality, these embeddings demonstrate good effectiveness in text.

In the comparative analysis of open-source LLM embeddings, Falcon-7b outperformed the LLaMA-2-7b across most datasets, demonstrating improved cluster quality and distinctiveness. This superiority may be attributed to Falcon-7b embeddings ability to capture better the salient linguistic features and semantic relationships within the texts since these embeddings were trained on a mixed corpus of text and code, as opposed to the LLaMA-2 embeddings, which are specialized for dialogues and Q&A contexts.

A screenshot of a computer

Description automatically generated

Clustering results from the original texts without generated summaries are higher than those with summarization. This finding suggests that essential details necessary for accurate clustering might have been lost during the summarization process. Alternatively, the inherent complexity and nuances of textual representation might require a more sophisticated approach to text summarization that can maintain essential information while reducing complexity. Additionally, it is important to highlight that we observed low-quality clustering results when using the summarization output from the smaller-sized LLaMA-2-7b and Falcon-7b models, likely due to their limited ability to capture and reproduce the complex nuances inherent in the source texts.

**Conclusions**

In this study, we examined the impact that various embeddings—namely TF-IDF, BERT, OpenAI, LLaMA-2, and Falcon—and clustering algorithms have on grouping textual data. Through detailed exploration, we evaluated the efficacy of dimensionality reduction via summarization and the role of

embedding size on the clustering accuracy of various datasets. We found that OpenAI’s sophisticated embeddings outperformed other embeddings. BERT’s performance excelled amongst open-source alternatives, underscoring the potential of advanced models to positively affect text clustering results.

This research also revealed a trade-off between improved clustering results and the computational weight of larger embeddings. Although results indicate that an increase in model size often correlates with enhanced clustering performance, the benefits must be weighed against the practicality of available computing resources.

A key takeaway from this investigation is that summary-based dimensionality reduction does not consistently improve clustering performance, signalling that a nuanced approach is necessary when preprocessing text to avoid losing vital information. Our exploration also brings to attention the practical challenges imposed by limited computational resources, which are a significant hurdle in the widespread application of large models, especially in expansive text analysis.

These findings point towards continued research focused on developing strategies that leverage the strengths of advanced models while mitigating their computational demands. It is also critical to expand the scope of research to include more diverse text types, which will provide a more comprehensive understanding of clustering dynamics across different contexts and inform the development of more universally applicable NLP tools.

In conclusion, our findings underscore the complex interplay between embedding types, dimensionality reduction, embedding size, and text clustering effectiveness in the context of structured, formal texts. While larger, more advanced embeddings like OpenAI’s present clear advantages, researchers and practitioners must consider trade-offs in terms of cost, computational resources, and the impact of text preprocessing techniques.