Leveraging LLMs for Efficient Religious Text Analysis

Bibiána Lajčinová [1], Jozef Žuffa [2], Milan Urbančok [2]

[1]Slovak National Supercomputing Centre, Bratislava, Slovak Republic
[2]Faculty of Theology, Trnava University, Bratislava, Slovak Republic

Abstract

The study explores the task of information retrieval using embedding models on texts with religious themes in Slovak language, with the aim to make analysis of these texts more efficient for religious scholars. Utilizing open-source *Slovak-BERT*, and *BGE M3* embedding models and closed-source *text-embedding-3-small* from OpenAI, we generated embeddings indices from text chunks with varying sizes and evaluated recall value of this solution across five different topics with sets of queries associated with the topics. Query augmentation and stopwords removal preprocessing techniques were explored too. The results indicate that this methodology is indeed effective in acceleration of research of religious text and can help uncover underlying interpretations and meanings within them. Our findings also emphasize the importance of choosing the proper preprocessing technique for the given model and data.

1 Introduction

The analysis and research of texts with religious themes have historically been the domain of philosophers, theologians, and other social sciences specialists. With the advent of artificial intelligence, such as the large language models (LLMs), this task takes on new dimensions. These technologies can be leveraged to reveal various insights and nuances contained in religious texts — interpreting their symbolism and uncovering their meanings. This acceleration of the analytical process allows researchers to focus on specific aspects of texts relevant to their studies. One possible research task in the study of texts with religious themes involves examining the works of authors affiliated with specific religious communities. By comparing their writings with the official doctrines and teachings of their denominations, researchers can gain deeper insights into the beliefs, convictions, and viewpoints of the communities shaped by the teachings and unique contributions of these influential authors.

This report proposes an approach utilizing embedding indices and LLMs for efficient analysis of texts with religious themes. The primary objective is to develop a tool for information retrieval, specifically designed to efficiently locate relevant sections within documents. The identification of discrepancies between the retrieved sections of texts from specific religious communities and the official teaching of the particular religion the community originates from is not part of this study; this task is entrusted to theological experts.

This work is a joint effort of Slovak National Competence Center for High-Performance Computing and the Faculty of Theology at Trnava University. Our goal is to develop a tool for information retrieval using LLMs to help theologians analyze religious texts more efficiently. To achieve this, we are leveraging resources of HPC system Devana to handle the computations and large datasets involved in this project.

2 Data

The texts used for the research in this study originate from the religious community known as the Nazareth Movement (commonly referred to as "Beňovci"), which began to form in the 1970s. The movement, which some scholars identify as having sect-like characteristics, is still active today, in reduced and changed form. Its founder, Ján Augustín Beňo (1921 - 2006), was a secretly ordained Catholic priest during the totalitarian era. Beňo encouraged members of the movement to actively live their faith through daily reading of biblical texts and applying them in practice through specific resolutions. The movement spread throughout Slovakia, with small communities existing in almost every major city. It also spread to neighboring countries such as Poland, the Czech Republic, Ukraine, and Hungary. In 2000, the movement included approximately three hundred married couples, a thousand children, and 130 priests and students preparing for priesthood. The movement had three main goals: radical prevention in education, fostering priests who could act as parental figures to identify and nurture priestly vocations in children, and the production and distribution of samizdat materials needed for catechesis and evangelization. 27 documents with texts from this community are available for research. These documents, which significantly influenced the formation of the community and its ideological positions, were reproduced and distributed during the communist regime in the form of samizdats — literature banned by the communist regime. After the political upheaval, many of them were printed and distributed to the public outside the movement. Most of the analyzed documents consist of texts intended for "morning reflections" — short meditations on biblical texts. The documents also include the founder's comments on the teachings of the Catholic Church and selected topics related to child rearing, spiritual guidance, and catechesis for children.

Although the documents available to us contained a few duplications, this did not pose a problem for the information retrieval task and will thus remain unaddressed in this report. All of the documents are written exclusively in Slovak language.

One of the documents is annotated for test purposes by experts from the partner faculty, who have long been studying the Nazareth Movement. By annotations, we refer to text parts labeled as belonging to one of the five classes, where these classes represent five topics, namely

1. Directive obedience
2. Hierarchical upbringing
3. Radical adoption of life model
4. Human needs fulfilled only in religious community and family
5. Strange/Unusual/Intense

Additionally, each of this topics is supplemented with a set of queries designed to test the retrieval capabilities of our solution.

Text	**Annotation**
Veď ak milujeme svojho Boha, ako si to myslíme, alebo aj hovoríme, nemôže nám byť ľahostajný nijaký odklon od jeho svätej vôle.	Directive obedience

Table 1: Example of a text with an annotation Directive Obedience.

3 Strategy/Solution

There are multiple strategies appropriate for solving this task, including text classification, topic modelling, retrieval-augmented generation (RAG), and fine-tuning of LLMs. However, the theologians' requirement is to identify specific parts of the text for detailed analysis, necessitating the retrieval of exact wording. Therefore, a choice was made to leverage information retrieval. This approach differs from RAG, which typically incorporates both information retrieval and text generation components, in focusing solely on retrieving textual data, without the additional step of new content generation.

Information retrieval leverages LLMs to transform complex data such as text, into a numerical representation that captures the semantic meaning and context of the input. This numerical representation, known as embedding, can be used to conduct semantic searches by analysing the positions and proximity of embeddings within a multi-dimensional vector space. By using queries, the system can retrieve relevant parts of the text by measuring the similarity between the query embeddings and the text embeddings.

This approach does not require any fine-tuning of the existing LLMs, therefore the models can be used without any modification and the workflow remains quite simple.

3.1 Model choice

Since the texts available for research are in Slovak language, the choice of a language model was very limited. As of today, there is only one open-source model that exclusively understands Slovak language, along with a few multilingual models that have some degree of proficiency in Slovak. Four pre-trained models were chosen from the sparse number of available options, with *Slovak-BERT* [1] being the first one. *Slovak-BERT* is an open-source Slovak-only transformer-based language model trained using a masked language modeling (MLM) objective. The second model is *text-embedding-3-small* model, which is the powerful third-generation embedding model and can be accessed through OpenAI's API. The third model is a *BGE M3* Embedding model [2], which is a currect state-of-the-art open-source multilingual embedding model supporting more than 100 languages. And the last one is Microsoft's multilingual embedding model *E5*, which is an open-source general-purpose text embeddings model [3].

These four models were leverages to acquire vector representations of the chunked text, and their specific contributions will be discussed in the following parts of the study.

3.2 Data preprocessing

The first step of data preprocessing involved text chunking. The primary reason for this step was to meet the requirement of religious scholars for retrieval of paragraph-sized chunks. Besides, documents needed to be split into smaller chunks anyway due to the limited input lengths of some LLMs. For this purpose, the *Langchain* library [4] was utilized. It offers hierarchical chunking that produces overlapping chunks of a specific length (with a desired overlap) to ensure that the context is preserved. Chunks with lengths of 300, 400, 500 and 700 symbols were generated. Subsequent preprocessing steps included removal of diacritics, case normalization according to the requirements of the models and stopwords removal. The removal of stopwords is a common practice in natural language processing tasks. While some models may benefit from the exclusion of stopwords to improve relevancy of retrieved chunks, others may take advantage of retaining stopwords to preserve contextual information essential for understanding the text.

Index	**Chunk**
8	Podľa tohoročnej sa rozvádza už každé tretie. Tento bolestný spoločenský jav vysvetľujú niektorí skutočnosťou, že dnešní manželia sú náročnejší a od svojho manželstva viac očakávajú než tí, čo žili pred nami. Že by to bola pravda? Od manželstva môže človek čakať, len toľko, koľko doň vloží.
9	Že by to bola pravda? Od manželstva môže človek čakať, len toľko, koľko doň vloží. Ak minulosť týmto bláznovstvom rozvodovosti netrpela, tak zaiste preto, že v manželstve nevidela iba tú sentimentálnu príjemnú lásku, ale aj tú obetavú, živú z viery v Boha a z poslušnosti voči Cirkvi. Zaujímajú nás začiatky, priebeh a dôsledky rozvodov? Nie je ťažko spoznať ich.

Table 2: Example of two text chunks with an overlap.

3.3 Vector Embeddings

Vector embeddings were created from text chunks using selected pre-trained language models.

For the *Slovak-BERT* model, generating embedding involves leveraging the model without any additional layers for inference and then using the first embedding, which contains all the semantic meaning of the chunk, as the context embedding. Other models produce embeddings in required form, so no further postprocessing was needed.

In the subsequent results section, the performance of all created embedding models will be analyzed and compared based on their ability to capture and represent the semantic content of the text chunks.

4 Results

Prior to conducting quantitative tests, all embedding indices underwent preliminary evaluation to determine the level of understanding of the Slovak language and the specific religious terminology by the selected LLMs. This preliminary evaluation involved subjective judgement of the relevance of retrieved chunks.

These tests revealed that the *E5* model embeddings exhibit limited effectiveness on our data. When retrieving for a specific query, the retrieved chunks contained most of the key words used in the query, but did not contain the context of the query. One of the explanations could be that this model prioritizes word-level matches over the nuanced context in Slovak language, because it's possible that the training data of this model for Slovak was less extensive or less contextually rich, leading to weaker performance. However, these observations are not definitive conclusions but rather hypotheses based on current, limited results. A decision was made not to further evaluate the performance of the embedding indices leveraging *E5* embeddings, as it seemed irrelevant given the inability to effectively capture the nuances of the religious texts.

On the other hand, the abilities of *Slovak-BERT* model, based on the RoBERTa architecture characterized by its relatively simple architecture, exceeded the expectations. Moreover, the performance of *text-embedding-3-small* and *BGE M3* embeddings met expectations, as the first test, subjectively evaluated, demonstrated a very good grasp of the context, proficiency in Slovak language, and understanding of the nuances within the religious texts.

Therefore, quantitative tests were performed only on embedding indices utilizing *Slovak-BERT*, OpenAI's *text-embedding-3-small* and *BGE M3* embeddings.

Given the problem specification and the nature of test annotations, there arises a potential concern regarding the quality of the annotations. It is possible that some text parts were misclassified as there may be sections of text that belong to multiple classes. This, combined with the possibility of human error, can affect the consistency and accuracy of the annotations.

With this consideration in mind, we have opted to focus solely on recall evaluation. By recall, we mean the proportion of correctly retrieved chunks out of the total number of annotated chunks, regardless of the fraction of false positive chunks. Recall will be evaluated for every topic and for every length-specific embedding index for all selected LLMs.

Moreover, the provided test queries might also reflect the complexity and interpretative nature of religious studies. For example, consider a query ”God's will” for the topic *Directive obedience*. While careful reader understands how this query relates to the given topic, it might not be as clear to a language model. Therefore, apart from evaluating using provided queries, another evaluation was conducted using queries acquired through contextual augmentation.

Contextual/query augmentation is a prompt engineering technique for enhancing text data quality and is well-documented in various research papers [5], [6]. This technique involves prompting a language model to generate a new query based on initial query and other contextual information in order to formulate a better query. Language model used for generation of queries through query augmentation technique was *GPT 3.5* and these queries will be referred to as ”GPT queries” throughout the rest of the report.

4.1 Slovak-BERT embedding indices

Recall evaluation for embedding indices utilizing *Slovak-BERT* embeddings for four different chunk sizes with and without stopwords removal is presented in Figure 1. The evaluation covers each topic specified in the list in Section 2 and includes both original queries and GPT queries.

We observe, that GPT queries generally yield better results compared to the original queries, except for the last two topics, where both sets of queries produce similar results. Also, it is apparent, that *Slovak-BERT*-based embeddings benefit from stopwords removal in most cases. The highest recall values were achieved for the third topic *Radical adoption of life model*, with the chunk size of 700 symbols with removed stopwords, reaching more than 47%. In contrast, the worst results were observed for the topic *Strange/Unusual/Intense*, where neither the original nor GPT queries successfully retrieved relevant parts. In some cases none of the relevant parts were retrieved at all.

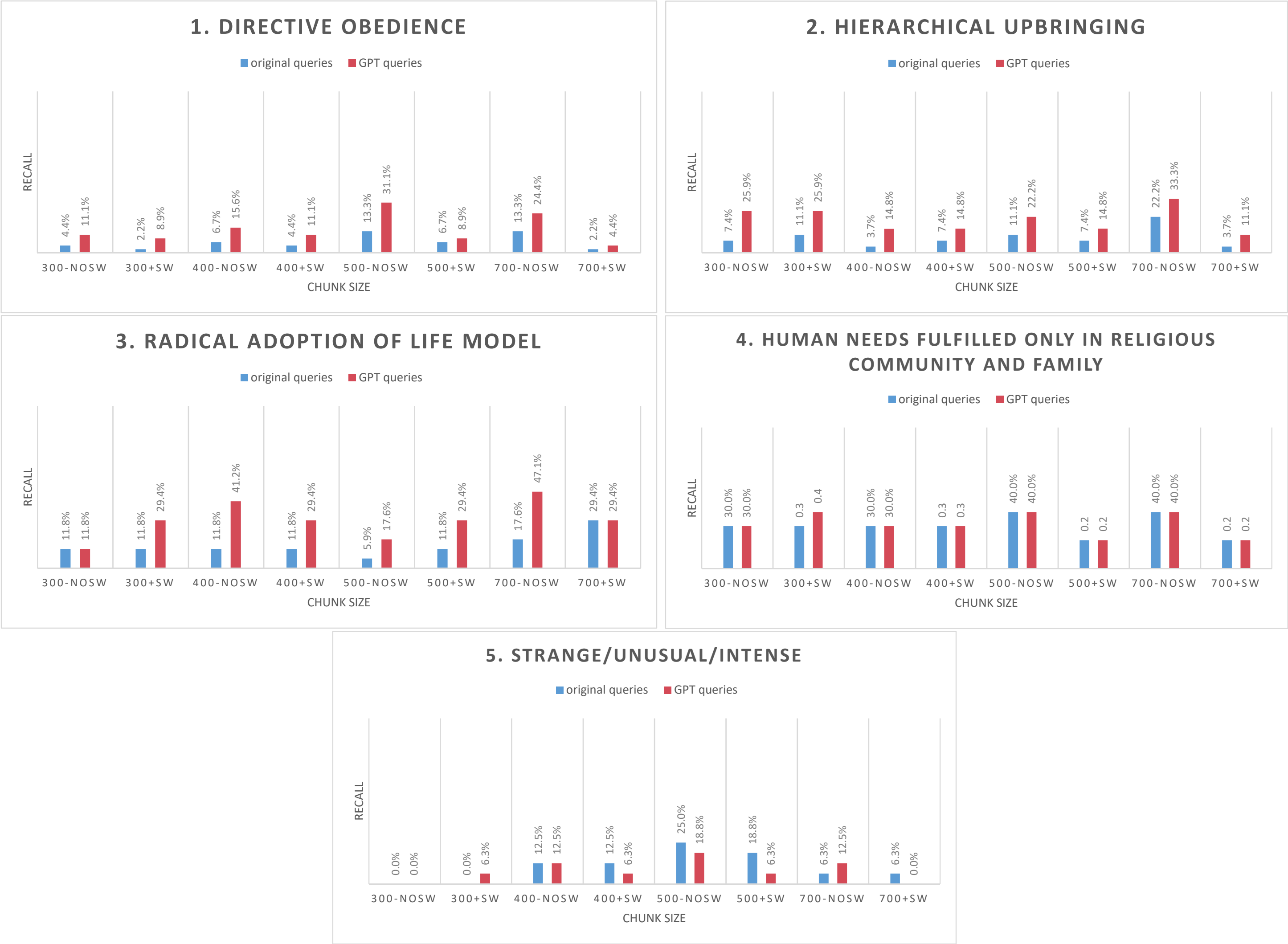

Figure 1: Recall values obtained for all topics using both original and GPT queries, across various chunk sizes of embeddings generated using the *Slovak-BERT* model. Embedding indices marked as +SW include stopwords, while -NoSW indicates stopwords were removed.

4.2 OpenAI's text-embedding-3-small embedding indices

Similar to the evaluation for *Slovak-BERT* embedding indices, evaluation charts for embedding indices utilizing OpenAI's *text-embedding-3-small* embeddings are presented in Figure 2. The recall values are generally much higher than those observed with *Slovak-BERT* embeddings. As with the previous results, GPT queries produce better outcomes. We can observe a subtle trend in recall value and chunk size dependency – longer chunk sizes generally yield higher recall values.

An interesting observation can be made for the topic *Radical adoption of life model*. When using the original queries, hardly any relevant results were retrieved. However, when using GPT queries, recall values were much higher, reaching almost 90% for chunk sizes of 700 symbols.

Regarding the removal of stopwords, its impact on embeddings varies. For topics 4 and 5, stopwords removal proves beneficial. However, for the other topics, this preprocessing step does not offer advantages.

Topics 4 and 5 exhibited the weakest performance among all topics. This may be due to the nature of the queries provided for these topics, which are quotes or full sentences, compared to queries for other topics, that are phrases, keywords or expressions. It appears that this model performs better with the latter type of queries. On the other hand, since the queries for topics 4 and 5 are full sentences, the embeddings benefit from stopwords removal, as it probably helps in handling the context of sentence-like queries. Topic 4 is very specific and abstract, while topic 5 is very general, making it understandable that capturing this topic in queries is challenging. The specificity of topic 4 might require more nuanced test queries, as the provided test queries probably did not contain all nuances of a given topic. Conversely, the general nature of topic 5 might benefit from a different analytical approach. Methods like Sentiment Analysis could potentially grasp the strange, unusual, or intense mood in relation to the religious themes analysed.

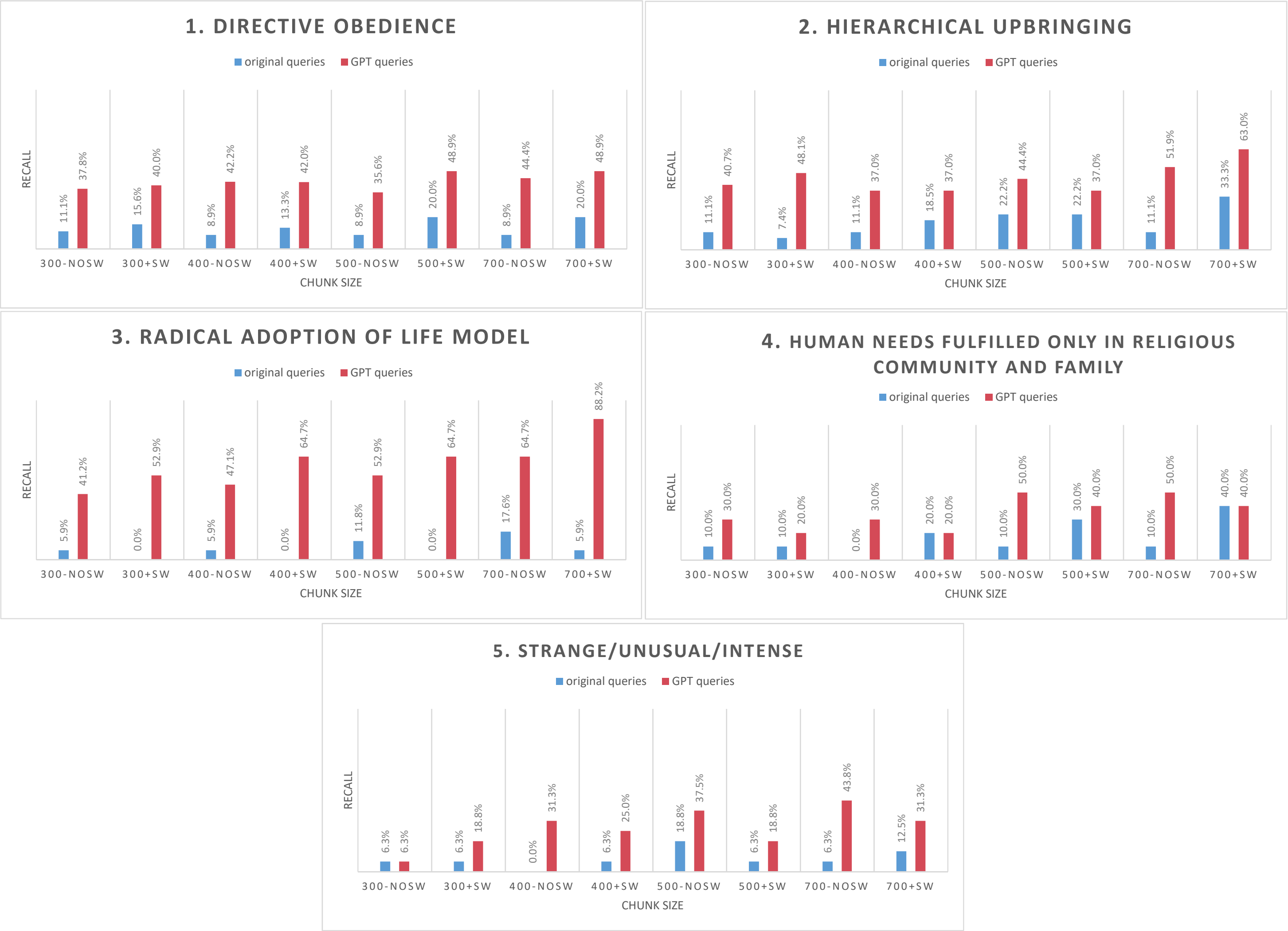

Figure 2: Recall values assessed for all topics using both original and GPT queries, utilizing various chunk sizes of embeddings generated with the *text-embedding-3-small* model. Embedding indices labeled +SW include stopwords, and those labeled -NoSW have stopwords removed.

4.3 BGE M3 embedding indices

Evaluation charts for embedding indices utilizing *BGE M3* embeddings are presented in Figure 3. The recall values demonstrate a performance falling between *Slovak-BERT* and OpenAI's *text-embedding-3-small* embeddings. While, in some cases, not reaching the recall values of OpenAI's embeddings, *BGE M3* embeddings show competitive performance, particularly considering their open-source availability compared to OpenAI's embeddings, that are accessible through API, which might pose a problem with data confidentiality.

With these embeddings, we also observe the same phenomenon as with OpenAI's *text-embedding-3-small* embeddings: shorter, phrase-like queries are preferred over quote-like queries. Therefore, recall values are higher for first three topics.

Stopwords removal seems to be mostly beneficial, mainly for the last two topics.

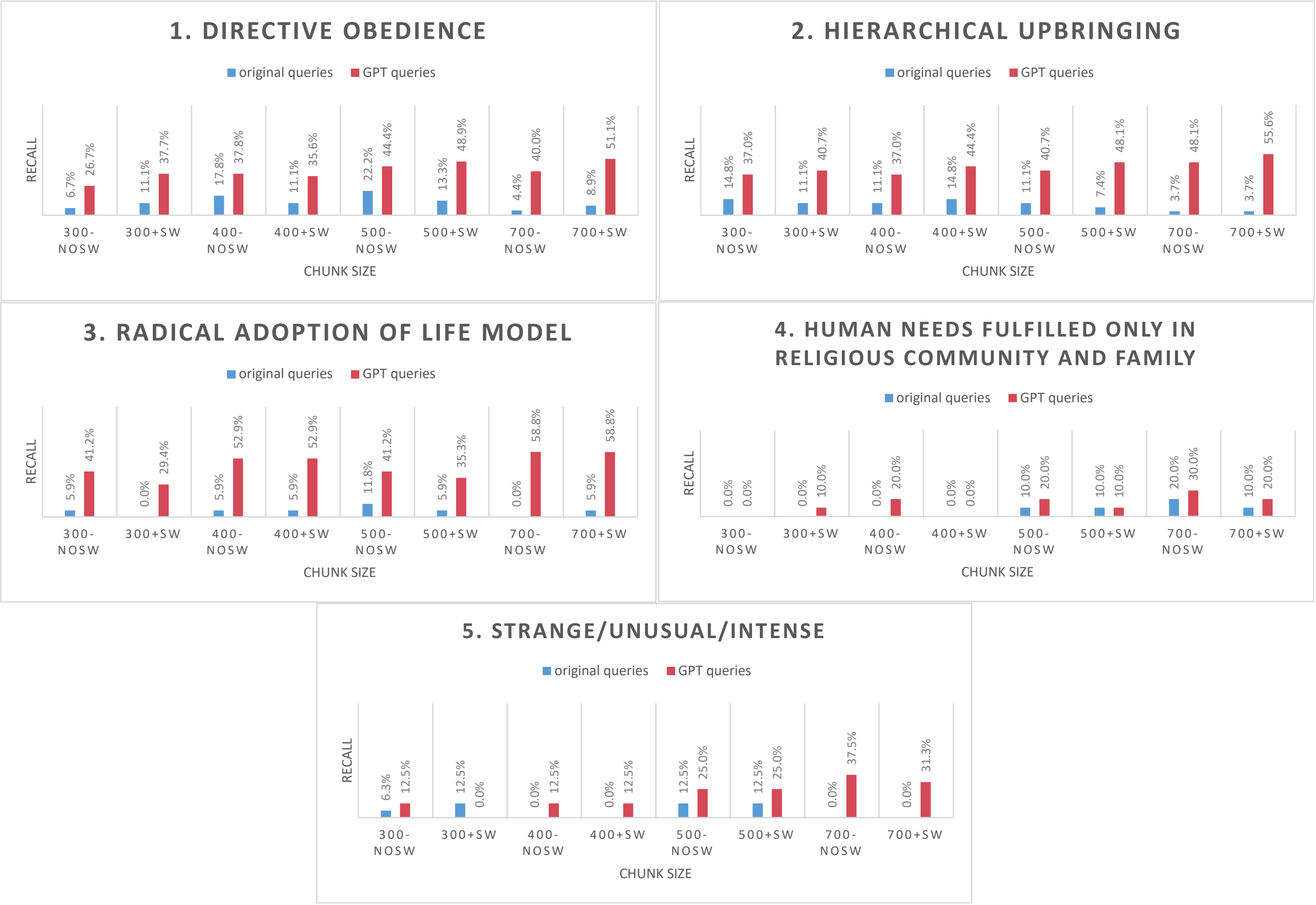

Figure 3: Recall values for all topics using original and GPT queries, with embeddings of different chunk sizes produced by the *BGE M3* model. Indices labeled as +SW contain stopwords, while -NoSW indicates their removal.

5 Conclusion

This paper presents an approach for analysis of text with religious themes with the use of text numerical representations known as embeddings, generated by three selected pre-trained large language models: *Slovak-BERT*, OpenAI's *text-embedding-3-small* and *BGE M3* embedding model. These models were selected after it was evaluated, that their proficiency in Slovak language and religious terminology is sufficient to handle the task of information retrieval for a given set of documents.

Challenges related to quality of test queries were addressed using query augmentation technique. This approach helped in formulating appropriate queries, resulting in more relevant retrieval of text chunks, capturing all the nuances of topics that interest theologians.

Evaluation results proved the effectiveness of the embeddings produced by these models, particularly the *text-embedding-3-small* from OpenAI, which exhibited a strong contextual understanding and linguistic proficiency. The recall value for this model's retrieval abilities varied depending of the topic and queries used, with the highest values reaching almost 90% for topic *Radical adoption of life model* when using GPT queries and chunk length of 700 symbols. Generally, *text-embedding-3-small* performed best with the longest chunk lengths studied, showing a trend of increasing recall with the increase in chunk length. The topic *Strange/Unusual/Intense* had the lowest recall, possibly due to the uncertainty in topic specification.

For *Slovak-BERT* embedding indices, the recall values were slightly lower, but still impressive given the simplicity of this language model. Better results were achieved using GPT queries, with the best recall value of 47.1% for the topic *Radical adoption of life model* at a chunk length of 700 symbols, with embeddings created from chunks with removed stropwords. Generally, this embedding model benefited most from the stopwords removal preprocessing step.

As for *BGE M3* embeddings, the result were impressive, achieving high recall, though not as high as OpenAI's embeddings. However, considering that BGE M3 is an open-source model, these results are remarkable.

These findings highlight the potential of leveraging LLMs for specialized domains like analysis of texts with religious themes. Future work could explore the connections between text chunks using clustering techniques with embeddings to discover hidden associations and inspirations of the text authors. For theologians, future work lies in examining the retrieved text parts to identify deviations from official teaching of Catholic Church, shedding light on movement's interpretations and insights.

6 Acknowledgements

Research results were obtained with the support of the Slovak National competence centre for HPC, the EuroCC 2 project and Slovak National Supercomputing Centre under grant agreement 101101903-EuroCC 2-DIGITAL-EUROHPC-JU-2022-NCC-01. Computational resources were procured in the national project National competence centre for high performance computing (project code: 311070AKF2) funded by European Regional Development Fund, EU Structural Funds Informatization of society, Operational Program Integrated Infrastructure.

References

[1] Matúš Pikuliak, Štefan Grivalský, Martin Konôpka, Miroslav Blšták, Martin Tamajka, Viktor Bachratý, Marián Šimko, Pavol Balážik, Michal Trnka, and Filip Uhlárik. Slovakbert: Slovak masked language model, 2021.

[2] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024.

[3] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual e5 text embeddings: A technical report, 2024.

[4] Harrison Chase. Langchain. https://github.com/langchain-ai/langchain, 2022. Accessed: May 2024.

[5] Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. Query rewriting for retrieval-augmented large language models, 2023.

[6] Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang, and Michael Bendersky. Query expansion by prompting large language models, 2023.