# Research Proposal

Dominik Stipić

Siječanj 2025

## 1 Introduction

The most recent research and commercial events in the field of the natural language processing could be the development of the large language models and retrieval augmented generation systems or RAGs. In general, the LLMs are parametric generative models, they are capable of modelling generative probability distribution of the tokens or words:

$$P(x_1, x_2, ..., x_n)$$

They are sequence to sequence models which are learned through supervision. They do this by embedding the words into the embedding vectors by using embedding matrices. Most applications use open source embedding representations such as word2vec and GloVe which are trained on the massive amounts of data. These are learned in a supervised fashion where each word would be paired with its k-neighborhood. Learned weights from this problem would be considered as the embedding vectors for the target word. The most prominent example of the large language model architecture is the transformer model.

Transformers differ from the typical encoder-decoder deep learning model by adopting the attention mechanism. In summary, this mechanism enables the creation of the contextualized representation which change the word meaning depending on the neighbourhood which surrounds the words.

They try to capture true semantic meaning of the word by assuming the distributional hypothesis. According to this hypothesis, the meaning of the word is also determined by it's neighbourhood.

**Retrieval augmented generation systems** are the systems which are capable of generating the authoritative human like answers from the set of sources or documents. A sources are semi-structured or structured documents which are believed to be authoritative sources about some topic. In practice, the source documents are embedded into the vector space and then stored in vector databases which are optimized for doing fast dot products between high dimensional vectors.

Unlike RAGs, large language models are prone to hallucinations or they produce reasonably sounded arguments which most of the users will evaluate as true. A RAG systems have two parts. The first part is the document retrieval system and second part is the generator system. Retrieval system retrieves set of documents on which generator conditiones his generated content alongside the user query. More specifically, they model the generative probability distribution conditioned on the prompt, query, sources, it's left neighbourhood:

$$P(x_1, x_2, ..., x_n) = \prod_k P(x_1, x_2, ..., x_n | q, R(q), prompt, question, x_j < k)$$

In this example, the prompt, query and sources are used for controlling the form of generated output. There are multiple ways how documents could be encoded. The canonical example encodes them as a bag of words and those are used for calculating document and term frequency statics which are used for calculating the similarity scores between document and query. Document encoder could also be the embedding matrix.

## 2 Conclusion

The rough boundary of this research topic would include the exploration or survey of the latest developments in the large language modelling field, retrieval systems and RAGs. This includes the analysis of the computation methodologies, different architecture approaches to the text generation and evaluation approaches.

When analysing the impact of some newly introduced submodules on the performance of the overall system, ablative studies have emerged as the norm. In the end, the work would also include the implementation of the RAG system which could generate its content conditioned on the database schema and data inside it.