

OPPONENT'S EVALUATION OF THE MASTER'S THESIS

Student: John Tawiah

**Opponent: Assoc. Prof. Martin
Kotyrba, Ph.D.**

Study program: **Information Technologies**
Study course/Specialization: **Software Engineering**
Academic year: **2023/2024**

Master's Thesis topic: **Customized Transformer Model for Efficient Extraction of
Information from Textbooks**

Evaluation of the thesis:

John Tawiah's Master's thesis addresses the area of natural language processing, focusing on developing and evaluating a bespoke transformer model tailored for textbook information extraction. Tawiah's work aims to overcome the limitations of generic language models by proposing solutions to enhance the accuracy and relevance of information extraction. This thesis investigates a solution that enables users to access concise information from large corpora, such as educational textbooks, through the implementation of a Retrieval-Augmented Generation (RAG) application. The thesis demonstrates the feasibility and effectiveness of this approach across various textbooks, regardless of their size, offering valuable insights into the architecture and tools necessary for implementation.

The thesis exhibits a meticulous approach to research methodology, under the expert guidance of Assoc. Prof. Michal Pluháček, PhD. The theoretical framework is robust, incorporating concepts such as recurrent neural networks, PDF text extraction, and sequence modelling techniques. The historical evolution of transformer models and the emergence of large language models are discussed in detail, setting a solid foundation for the study. The thesis is systematically structured into six core sections, but the very labelling of these parts is quite confusing. The thesis has overall 59 pages and 35 literature links. Some pictures are completely unnecessary in the work, such as picture number 8. So, I would expect a much better described chapter with results and I lack a literature review chapter in this thesis

The primary aim of the thesis is to investigate the feasibility of enabling users to query a large language model using custom data, thereby simplifying the extraction of information from textbooks and articles. The thesis also aims to evaluate existing large language models to determine their suitability for the task at hand. The best part of this thesis is proposal of transformer model tailored to efficiently extract and summarize information from educational textbooks.

Through a comprehensive analysis of vector databases and implementation architectures, the thesis presents an interesting approach called Retrieval Augmented Generation (RAG) architecture. This innovative framework combines retrieval-based and generation-based techniques to enhance the efficiency of information extraction from textbooks. The analysis underscores the advantages of using a custom transformer model for educational purposes, including improved comprehension, reduced manual effort, and enhanced learning outcomes.

Questions for defense:

1. How can the RAG architecture be refined and scaled for real-world applications beyond academic research?
2. Why were specific metrics chosen to evaluate the model's performance, and how do they reflect the model's effectiveness?
3. How do external factors, such as varying textbook formats, influence the model's performance?

In conclusion, John Tawiah's Master's thesis represents a significant contribution to the field of natural language processing, offering valuable insights into the development of domain-specific language models for efficient textbook information extraction. Over the many shortcomings mentioned above, I recommend the thesis for defense and I am evaluate

Overall evaluation of the thesis:

The Opponent shall grant a mark according to the ECTS classification scale:

A – Excellent, B – Very Good, C – Good, D – Satisfactory, E – Sufficient, F – Insufficient

An “F” grade also means "I do not recommend the thesis for defence."

I recommend this thesis to be defended and suggest the following evaluation:

B - Very Good

In the case of an evaluation grade of “F – Insufficient”, please supply the main shortages and reasons for this assessment.

Date: 24.05.2024

Thesis Opponent's Signature: