Posudek vedoucího bakalářské práce

Studijní program: **Aplikovaná informatika**
Akademický rok: **2024/2025**
Název práce: **Inovace velkých jazykových modelů (LLM) pomocí doladěné metody Retrieval-Augmented Generation (RAG): Návrh a vyhodnocení**
Řešitel: **Šimon Pivoda**
Vedoucí práce: **Ing. Richard Antonín Novák, Ph.D.**
Oponent: **Ing. Jiří Korčák**

	Hlediska	Stupeň hodnocení
1.	Jasnost a srozumitelnost formulace tématu a cíle práce	**1**
2.	Rozsah a relevance popisu současného poznání	**1**
3.	Náročnost řešeného tématu práce	**1**
4.	Adekvátnost metod k řešení stanoveného problému, správnost jejich výběru a použití	**2**
5.	Rozsah, hloubka a preciznost popisu výsledku	**2**
6.	Relevance a správnost diskuse výsledku	**2**
7.	Věcný přínos výsledku dosaženého v práci	**1**
8.	Relevance informačních zdrojů a korektnost jejich citování	**1**
9.	Logická stavba práce a vzájemná konzistence jednotlivých částí	**1**
10.	Gramatika, jazykový styl, terminologie a celková úprava práce	**1**
11.	Iniciativnost studenta a spolupráce s vedoucím práce	**1**
12.	Využití analytických metod a metod zpracování dat	**2**
13.	Naplnění zásad etiky a udržitelnosti	**2**
14.	Schopnost kritického a tvůrčího myšlení	**2**

Stanovisko k originalitě práce:

The originality check report from the VŠE Validator application showed a 22% match. The supervisor reviewed the individual matches in detail and, given that there are a large number (276) of minor matches (<1%), mostly consisting of 2-4 word overlaps, most often English terminology, Authors and article names (70 lit.sources included), or parts of code, the supervisor’s conclusion is that the work can be considered acceptable in terms of originality.

Konkrétní připomínky a dotazy k práci:

The student demonstrated a very active approach throughout the thesis, consulting intensively, sometimes too intensively without prior self-reflection from the supervisor’s point of view.

The thesis has a very good formal level in terms of its logical structure, typography, and use of sources, including appropriate artifacts such as images listed in the index. This makes the thesis clear and readable.

The objective of the thesis was set as follows:

• How can Retrieval-Augmented Generation (RAG) be used to enhance a general purpose Large Language Model’s performance as well as factual correctness on domain-specific queries?

In the theoretical part, the student did an above-standard job of processing the topic, applying the PRISM method within SLR, where he identified 164 sources based on keywords, which he then read and included 70 citations in his work, which shows the intensity and labor intensity of this part of the work.

In the practical part, the student evaluated the implemented system’s ability to answer factual, domain-specific queries by comparing the performance of three different configurations:

1, Basic Pipeline, which uses a large language model (LLM) in isolation. 2, RAG Pipeline, which augments the model with relevant retrieved data. 3, ChatGPT assistant with direct access to a CSV file containing

To compare these three LLM methods and their extensions, a QnA use case was selected from a specialized dataset from Kaggle on the technical parameters of specific cars, where the following information was provided for each vehicle type in the dataset: Torque(lb-ft), Electrical motor torque, CO2 Emissions, Turning circle, Cylinders or Displacement among others.

To compare the relevance of the QnA process over LLM alternatives, the student considered several methods such as: BLEU, ROUGE or embedding similarity. In the end, considering the scope of the work, he only implemented: Binary scoring reflects practical utility in an expert use case (e.g., “Is the answer right or not?”).

This binary evaluation was performed on a set of 15 questions, where the student, as a human expert, evaluated the answer obtained from the LLM in comparison with the correct answer known in advance and named as: ”ground truth”.

Summary of Results:

1, Basic Pipeline: 5/15 correct (33 %) 2, RAG Pipeline: 15/15 correct (100 %) 3, ChatGPT Assistant: 15/15 correct (100 %)

In his bachelor thesis, the student set himself an unnecessarily ambitious goal in the extensive theoretical section, resulting in a final thesis of 102 pages and a total of 70 citations. He also set himself very ambitious goals in the practical section.To address this challenge, a complete RAG pipeline was designed and implemented. The architecture leveraged OpenAI’s GPT-4o-mini as the core generative model, used text-embedding and employed a PostgreSQL database with pgvector to facilitate efficient similarity-based retrieval over a transformed dataset. He ultimately succeeded in achieving the goal of comparing multiple LLM extensions by comparing them according to a binary evaluation of QnA correctness.

However, the student himself acknowledges in the discussion that there is still a lot of work to be done within the scope of the topic and a correct comparison of LLM and RAG such as: contextual filtering, ranking refinement, interpretability layers, ROUGE and BLEU scoring evaluation, different embbeding then cosin F...etc.

Overall, I rate the work as excellent despite all the limitations in the treatment of this extremely broad topic, as its scope and execution far exceed the requirements of a bachelor’s degree.

Questions: 1, In your work, your own RAG solution and ChatGPT Assistant proved equally relevant. For which use cases do you think it would make sense to replace ChatGPT Assistant with your own RAG solution?

2, What are the main criteria to select specific embbeding functions?

Závěr: Bakalářskou práci doporučuji k obhajobě.

Navrhovaná výsledná klasifikace práce: **1**

Datum: 27. 5. 2025

Ing. Richard Antonín Novák, Ph.D.
vedoucí práce