KTH ROYAL INSTITUTE
OF TECHNOLOGY

Degree Project in Technology

Second cycle, 30 credits

Evaluating retrieval and summarisation performance of AI-Assistants built with Large Language Models and RAG-techniques (Retrieval Augmented Generation) in the domain of a LMS (Learning Management System)

A subtitle in the language of the thesis

LUDWIG KRISTOFFERSSON

Stockholm, Sweden, 2024

Evaluating retrieval and summarisation performance of AI-Assistants built with Large Language Models and RAG-techniques (Retrieval Augmented Generation) in the domain of a LMS (Learning Management System)

A subtitle in the language of the thesis

LUDWIG KRISTOFFERSSON

Master's Programme, Computer Science, 120 credits
Date: April 28, 2024

Supervisors: Michael Welle, Fredrik Enoksson
Examiner: Danica Jensfelt
School of Electrical Engineering and Computer Science
Host company: KTH IT
Swedish title: Detta är den svenska översättningen av titeln
Swedish subtitle: Detta är den svenska översättningen av undertiteln

Abstract

Foobar

Keywords

Canvas Learning Management System, Docker containers, Performance tuning

Sammanfattning

Foobar

Nyckelord

Canvas Lärplattform, Dockerbehållare, Prestandajustering

Acknowledgments

I would like to thank FEN for having yyyy. Or in the case of two authors:
We would like to thank xxxx for having yyyy.

Stockholm, April 2024
Ludwig Kristoffersson

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

This degree project will investigate Large Language Models (Large Language Models (LLMs)) and Retrieval Augmented Generation (Retrieval Augmented Generation (RAG)) systems in the form of deploying an AI-Assistant in canvas course rooms. The degree project will investigate how to evaluate these systems in very specialised domains and benchmark various models, approaches and techniques.

The reason this research is important is that LLMs have gained widespread attention and we are likely to see large-scale adoption of these models into various applications. Understanding how to benchmark and evaluate these systems in specialised domains will be crucial to understand how to build these systems, which techniques to use, and which models work well.

Many organisations need to, due to commercial and regulatory compliance, host all AI-models themselves. This aspect is also interesting to evaluate, i.e. how well open source and commercially licensed models compare against the closed source models, such as GPT-4 by OpenAI.

1.1.0.1 Where the research is taking place

The research will be carried out within the e-learning management object at KTH, who are responsible for the digital learning environment at KTH. The object consists of two teams at the KTH IT department and one team at the digital learning unit at the ITM-school. The university hosts thousands of courses with domain specific information, such as assignments, lectures and schedules, that aren't part of the public domain and therefore not part of the training set of LLMs.

All the work done by KTH IT aims to improve the operations at the university. Among this is reducing the administrative burden undertaken by teachers and teaching assistants (TAs). KTH IT wants to investigate if AI-assistants can be deployed into the canvas course rooms to reduce the workload of teachers and TAs which would help them focus on teaching, helping students and improve the quality of the education. KTH IT wants to see if it's feasible to deploy an AI assistant into the canvas course rooms.

1.2 Problem

Large Language Models (LLMs) have gained widespread use since its popularisation by ChatGPT. Their abilities to summarise large bodies of text and follow user instructions have proven very useful in many contexts. However, considering their limited context window (and drawbacks of models with larger context window [**?**]) deploying useful applications with a chat based interface still rely upon integrating a RAG system, introduced by Lewis et al. [**?**]. These can retrieve relevant information needed to answer a user's query from outside data sources and inject them into the conversation.

Some unreleased models, such as the gemini family of models [**?**], have been reported to show great recall performance and reasoning abilities over millions of tokens. This could significantly reduce the importance of RAG systems in applications which utilise LLMs and external datasets to create intelligent systems with domain specific knowledge. However, even though no exact figures are presented by the Gemini time, inference speed (the time taken to produce a response to a prompt) seems to be significantly slower than shorter contexts. This would again highlight the importance of efficient RAG systems. Still, other approaches than traditional GPUs have been shown recently [**?**] by the Groq team to greatly increase inference speed.

Evaluating large language models is notoriously difficult. There are objective and automated metrics that can be used for tasks such as evaluating a model's summarisation capabilities, as shown by Basyal and Sanghvi [**?**]. However, for more complicated evaluations it gets trickier. In their seminal instructGPT paper Ouyang et al. at OpenAI try to evaluate *"how well a model can follow instructions"* [**?**] which is a very subjective question. They essentially relied upon human labelers to judge the overall quality of each response generated by the model.

In their Gemini-paper the Gemini team discuss the benchmarks used for their largest model. The team states that benchmarks are often designed to test shorter prompts whereas their longer prompts challenge tests used in

traditional evaluation methods that rely heavily on manual evaluation. This highlights the relevance of good evaluation metrics. Regardless of context size or inference speed, evaluation of models tends to be very general. Which makes sense, when considering their general application.

When releasing their Mixtral model [**?**] the Mistral AI team used a range of benchmark tests, such as MMLU, PIQA, GSM8K etc. MMLU (*Measuring Massive Multitask Language Understanding*) [**?**] benchmarks a LLMs proficiency in understanding and reasoning across various subjects such as humanities, STEM, and professional and everyday knowledge, by evaluating its performance on 57 tasks, to test its ability to generalise and apply knowledge. PIQA (*Physical Interaction: Question Answering*) [**?**], evaluates a language models understanding of physical commonsense by asking them to predict the outcome of physical interactions in various scenarios through multiple-choice questions. GSM8K (*The Grade School Math*) [**?**] tests the ability to solve elementary-level mathematics word problems.

Evaluation of how well LLMs perform is an open research question. As shown above LLM developers often utilise multiple testsuites. These are oftentimes, as shown above, very general tests. When implementing LLMs in practical applications good performance often relies upon very good raw summarisation performance and reasoning abilities. Since the domain specific knowledge is provided to the model, raw built-in knowledge isn't crucial. It is more important for the model to learn the task at hand using very few examples and within the given domain understand the question being asked by a user. Further, as argued by by Siriwardhana et al., the training data of LLMs include the knowledge of datasets such as Wikipedia [**?**] which means that evaluation methods in very specialised domains hold higher value than generalised domains. These brand new domains, that with certainty haven't been seen during training, tests the models zero-shot, and depending on the implementation, few-shot learning abilities.

The research question for this project is *Which language model and which retrieval techniques do students prefer using?* and *Is it possible to deploy an AI-assistant using a completely open source toolchain?*.

I believe the answer to the first question is that the closed source alternatives will be preferred by the students, however, I think the results will show it is possible to deploy an open source based AI assistant too.

1.2.1 Original problem and definition

The core challenge addressed in this thesis is the effective deployment and evaluation of AI-assistants powered by LLMs and RAG techniques in a specialised domain, specifically within the Learning Manegement System (LMS) of Canvas course rooms at KTH. This involves assessing the practicality and efficiency of integrating AI-Assistants built upon LLMs and RAG techniques into the educational settings to aid in reducing administrative burdens on educators and enhancing student interaction with course materials.

The original problem stems from the need to understand whether AI-assistants can effectively handle the domain-specific data intrinsic to educational platforms that are not included in their initial training datasets. Furthermore, the project aims to compare the efficacy and acceptability of open-source versus proprietary AI models in real-world educational applications.

1.3 Purpose

The purpose of this thesis is two-fold: firstly, to innovate within the educational technology space by integrating AI-assistants to potentially reduce workload and improve informational access within Canvas course rooms. Secondly, the thesis aims to contribute to academic knowledge by providing empirical data on the performance of these AI systems in a controlled educational setting. This dual purpose ensures the project not only addresses the immediate needs of KTH's digital learning environment but also enriches the scientific community's understanding of applied AI in education.

This research is intended to benefit educational institutions by potentially offering a tool that improves operational efficiency and students by providing an alternative, possibly more effective way of interacting with course content. In addition the research will bring benefit for researchers within AI and education. Ethically, the study focuses on the sustainable development of AI technologies by emphasising open-source solutions, aiming to democratise advanced technological developments and reduce reliance on proprietary models.

1.4 Goals

The goals of this degree project are organised to comprehensively assess the deployment of AI-assistants within the educational framework of KTH's

Canvas LMS, focusing on technological effectiveness and user receptiveness:

1. **Technological Efficacy:** To evaluate the accuracy, speed, and reliability of responses by AI-assistants utilising both proprietary and open-source models in handling domain-specific content.

2. **User Preference:** To ascertain the preferences of different user groups (students, faculty) regarding the usability, information quality, and overall experience of interacting with various AI-assistant models and retrieval techniques.

3. **Operational Feasibility:** To assess the feasibility of integrating a fully open-source AI-assistant within an academic setting, considering logistical, technical, and regulatory constraints.

4. **Educational Impact:** To explore the potential of AI-assistants to reduce administrative burdens on educators and improve information accessibility for students.

5. **Comparative Analysis:** To perform a comparative study between open-source and proprietary models concerning their deployment costs, maintenance needs, and infrastructure requirements.

Each goal is designed to address a specific aspect of AI technology integration within educational practices, ensuring that the project outcomes are relevant and useful for both academic research and educational administration.

1.5 Research Methodology

This project employs a hybrid research methodology combining empirical data collection with qualitative insights to evaluate the implementation of AI-assistants in an educational setting effectively:

1.5.1 System Design and Implementation

Model Selection Different models, including proprietary and open-source, with different training data/methods, will be evaluated to determine their performance in educational environments.

RAG Integration Various configurations of Retrieval Augmented Generation systems will be tested to identify the most effective method for

enhancing the AI's responses with relevant information from KTH's course-specific data.

1.5.2 Evaluation Design

Study Participants The study will involve students using the AI-assistant and providing feedback on their experiences. There are various types of students participating in the study.

Experimental Setup Controlled experiments will be conducted where participants use different configurations of the AI-assistant for typical student questions.

Data Collection Methods Data will be collected through integrated survey tools within the chat interface, capturing real-time feedback on the AI-assistant's performance and student satisfaction.

1.5.3 Analysis Techniques

Quantitative Analysis Statistical methods will analyse usage data and response accuracy to quantitatively assess the AI-assistant's performance.

Qualitative Analysis Feedback and open-ended responses will be analysed textually to understand user perceptions and contextual effectiveness of the AI-assistant.

This methodology was chosen for its ability to provide a comprehensive evaluation of both the technical capabilities and the practical usability of AI-assistants, offering insights into their potential benefits and limitations in the specific context for this study.

1.6 Delimitations

This project has several delimitations that define the scope and boundaries of the research to ensure a focused and manageable study. The key delimitations are;

- **Model Scope:** The project will not involve the development of new models or the fine-tuning of existing models. This includes LLMs and embedding functions. The study will utilise pre-trained models offered by bigger vendors or the open source community.

- **Data Limitations:** Only existing courses within KTH's Canvas LMS will be utilised for the study. No new course content will be created, and no modifications will be made to existing course materials beyond what is necessary for the integration and testing of the AI-assistants.
- **Course Data Access:** The project will not use Canvas APIs for data integration. All interactions with the Canvas platform will be through existing interfaces, or data will be scraped and used from the Canvas web interface.
- **Geographic and Cultural Constraints:** The study is limited to the KTH environment, which may not represent other educational settings in different cultural or geographic contexts. The findings might not be directly transferable to other institutions or countries without additional localization and adaptation.

These delimitations are set to clarify the focus of the research and define what is outside the scope of this thesis project. They help in managing expectations and provide a clear framework within which the study operates.

1.7 Structure of the thesis

Chapter 2 presents relevant background information about xxx. Chapter 3 presents the methodology and method used to solve the problem. ...

Chapter 2

Background

This chapter provides basic background information about xxx. Additionally, this chapter describes xxx. The chapter also describes related work xxxx.

2.1 Major background area 1

There are xxx characteristics that distinguish yyy from other information and communication technology (ICT) system, as shown in Figure 2.1. Table 2.1 summarizes these characteristics.

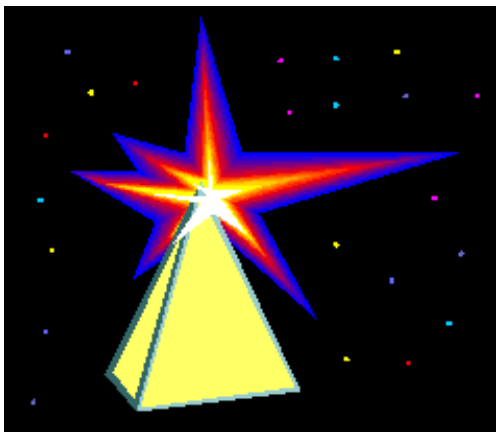

Figure 2.1: Lots of stars (Inspired by Figure x.y on page z of [xxx])

Table 2.1: xxx characteristics

Characteristics	**Description**
α	β
1	1 110.1
2	10.1
3	23.113 231

2.1.1 Subarea 1.1

Entangled states are an important part of quantum cryptography, but also relevant in other domains. This concept might be relevant for neutrinos, see for example [**?**].

2.1.2 Subarea 1.1.2

Computational methods are increasingly used as a third method of carrying out scientific investigations. For example, computational experiments were used to find the amount of wear in a polyethylene liner of a hip prosthesis in [**?**].

2.1.3 Subarea 1.1.2

Using the nearest data center may improve performance, see [**?**]

2.1.4 Link layer Encapsulation

2.1.5 IP packet headers

2.1.6 Test for accessibility of formulas

2.2 Major background area 2

2.2.1 Network layer security

...

2.3 Related work area

...

2.3.1 Major related work 1

Carrier clouds have been suggested as a way to reduce the delay between the users and the cloud server that is providing them with content. However, there is a question of how to find the available resources in such a carrier cloud. One approach has been to disseminate resource information using an extension to OSPF-TE, see Roozbeh, Sefidcon, and Maguire [**?**].

2.3.2 Major related work n

2.3.3 Minor related work 1

...

2.3.4 Minor related work n

2.4 Summary

Chapter 3

Method or Methods

3.1 Research Process

3.2 Research Paradigm

3.3 Data Collection

3.3.1 Sampling

3.3.2 Sample Size

3.3.3 Target Population

3.4 Experimental design and Planned Measurements

3.4.1 Test environment/test bed/model

3.4.2 Hardware/Software to be used

3.5 Assessing reliability and validity of the data collected

3.5.1 Validity of method

3.5.2 Reliability of method

3.5.3 Data validity

3.5.4 Reliability of data

3.6 Planned Data Analysis

3.6.1 Data Analysis Technique

3.6.2 Software Tools

Chapter 4

What you did

4.1 Hardware/Software design . . . /Model/Simulation model & parameters/. . .

4.2 Implementation . . . /Modeling/Simulation/. . .

4.2.1 Some examples of coding

4.2.2 Some examples of figures in tikz

4.2.2.1 Azure's Form Recognizer

Chapter 5

Results and Analysis

In this chapter, we present the results and discuss them.

5.1 Major results

Some statistics of the delay measurements are shown in table... The delay has been computed from the time the GET request is received until the response is sent.

5.2 Reliability Analysis

5.3 Validity Analysis

Chapter 6

Discussion

diskussion här

Chapter 7

Conclusions and Future work

7.1 Conclusions

7.2 Limitations

7.3 Future work

Due to the breadth of the problem, only some of the initial goals have been met. In these section we will focus on some of the remaining issues that should be addressed in future work. ...

7.3.1 What has been left undone?

The prototype does not address the third requirment, *i.e.,* a yearly unavailability of less than 3 minutes; this remains an open problem. ...

7.3.1.1 Cost analysis

The current prototype works, but the performance from a cost perspective makes this an impractical solution. Future work must reduce the cost of this solution; to do so, a cost analysis needs to first be done. ...

7.3.1.2 Security

A future research effort is needed to address the security holes that results from using a self-signed certificate. Page filling text mass. Page filling text mass. ...

7.3.2 Next obvious things to be done

In particular, the author of this thesis wishes to point out xxxxxx remains as a problem to be solved. Solving this problem is the next thing that should be done. ...

7.4 Reflections

One of the most important results is the reduction in the amount of energy required to process each packet while at the same time reducing the time required to process each packet.

References

Appendix A

Supporting materials

Appendix B

Something Extra

TRITA-EECS-EX- 2023:0000

€€€€ For DIVA €€€€

{
"Author1": { "Last name": "Kristoffersson",
"First name": "Ludwig",
"Local User Id": "u100001",
"E-mail": "ludwigkr@kth.se",
"organisation": {"L1": "School of Electrical Engineering and Computer Science",
}
},
"Cycle": "2",
"Course code": "DA231X",
"Credits": "30.0",
"Degree1": {"Educational program": "Master's Programme, Computer Science, 120 credits"
,"programcode": "TCSCM"
,"Degree": "Masters degree"
,"subjectArea": "Technology"
},
"Title": {
"Main title": "Evaluating retrieval and summarisation performance of AI-Assistants built with Large Language Models and RAG-techniques (Retrieval Augmented Generation) in the domain of a LMS (Learning Management System)",
"Subtitle": "A subtitle in the language of the thesis",
"Language": "eng" },
"Alternative title": {
"Main title": "Detta är den svenska översättningen av titeln",
"Subtitle": "Detta är den svenska översättningen av undertiteln",
"Language": "swe"
},
"Supervisor1": { "Last name": "Welle",
"First name": "Michael",
"Local User Id": "u100003",
"E-mail": "mwelle@kth.se",
"organisation": {"L1": "School of Electrical Engineering and Computer Science",
"L2": "COLLABORATIVE AUTONOMOUS SYSTEMS DIVISION OF ROBOTICS, PERCEPTION AND LEARNING" }
},
"Supervisor2": { "Last name": "Enoksson",
"First name": "Fredrik",
"Local User Id": "u100003",
"E-mail": "fen@kth.se",
"organisation": {"L1": "",
"L2": "UNIT OF DIGITAL LEARNING" }
},
"Examiner1": { "Last name": "Jensfelt",
"First name": "Danica",
"Local User Id": "u1d13i2c",
"E-mail": "danik@kth.se",
"organisation": {"L1": "School of Electrical Engineering and Computer Science",
"L2": "COLLABORATIVE AUTONOMOUS SYSTEMS DIVISION OF ROBOTICS, PERCEPTION AND LEARNING" }
},
"Cooperation": { "Partner_name": "KTH IT"},
"National Subject Categories": "10201, 10206",
"Other information": {"Year": "2024", "Number of pages": "1,29"},
"Copyrightleft": "copyright",
"Series": { "Title of series": "TRITA-EECS-EX" , "No. in series": "2023:0000" },
"Opponents": { "Name": "A. B. Normal & A. X. E. Normalè"},
"Presentation": { "Date": "2022-03-15 13:00"
,"Language":"eng"
,"Room": "via Zoom https://kth-se.zoom.us/j/ddddddddddd"
,"Address": "Isafjordsgatan 22 (Kistagången 16)"
,"City": "Stockholm" },
"Number of lang instances": "2",
"Abstract[eng]": €€€€
€€€€,
"Keywords[eng]": €€€€
Canvas Learning Management System, Docker containers, Performance tuning
€€€€,
"Abstract[swe]": €€€€
€€€€,
"Keywords[swe]": €€€€
Canvas Lärplattform, Dockerbehållare, Prestandajustering €€€€,
}

acronyms.tex

```
%%% Local Variables:
%%% mode: latex
%%% TeX-master: t
%%% End:
% The following command is used with glossaries-extra
\setabbreviationstyle[acronym]{long-short}
% The form of the entries in this file is \newacronym{label}{acronym}{phrase}
%                                      or \newacronym[options]{label}{acronym}{phrase}
% see "User Manual for glossaries.sty" for the  details about the options, one example is shown below
% note the specification of the long form plural in the line below
\newacronym[longplural={Debugging Information Entities}]{DIE}{DIE}{Debugging Information Entity}
%
% The following example also uses options
\newacronym[shortplural={OSes}, firstplural={operating systems (OSes)}]{OS}{OS}{operating system}

% note the use of a non-breaking dash in long text for the following acronym
\newacronym{IQL}{IQL}{Independent Q^^e2^^80^^91Learning}

\newacronym{KTH}{KTH}{KTH Royal Institute of Technology}

\newacronym{LAN}{LAN}{Local Area Network}
\newacronym{VM}{VM}{virtual machine}
% note the use of a non-breaking dash in the following acronym
\newacronym{WiFi}{Wi^^e2^^80^^91Fi}{Wireless Fidelity}

\newacronym{WLAN}{WLAN}{Wireless Local Area Network}
\newacronym{UN}{UN}{United Nations}
\newacronym{SDG}{SDG}{Sustainable Development Goal}

\newacronym{LMS}{LMS}{Learning Manegement System}
\newacronym{RAG}{RAG}{Retrieval Augmented Generation}
\newacronym{LLMs}{LLMs}{Large Language Models}
```