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

1 Introduction 1

1.1 Background . . . 1

1.2 Problem . . . 1

1.2.1 Original problem and definition . . . 1

1.3 Purpose . . . 1

1.4 Goals . . . 1

1.5 Research Methodology . . . 2

1.6 Structure of the thesis . . . 2

2 Background 3

2.1 Major background area 1 . . . 3

2.1.1 Subarea 1.1 . . . 4

2.1.2 Subarea 1.1.2 . . . 4

2.1.3 Subarea 1.1.2 . . . 4

2.1.4 Link layer Encapsulation . . . 4

2.1.5 IP packet headers . . . 4

2.1.6 Test for accessibility of formulas . . . 4

2.2 Major background area 2 . . . 4

2.2.1 Network layer security . . . 4

2.3 Related work area . . . 5

2.3.1 Major related work 1 . . . 5

2.3.2 Major related work n . . . 5

2.3.3 Minor related work 1 . . . 5

2.3.4 Minor related work n . . . 5

2.4 Summary . . . 5

3 Method or Methods 7

3.1 Research Process . . . 8

3.2 Research Paradigm . . . 8

3.3 Data Collection 8
3.3.1 Sampling 8
3.3.2 Sample Size 8
3.3.3 Target Population 8
3.4 Experimental design/Planned Measurements 8
3.4.1 Test environment/test bed/model 8
3.4.2 Hardware/Software to be used 8
3.5 Assessing reliability and validity of the data collected 8
3.5.1 Validity of method 8
3.5.2 Reliability of method 8
3.5.3 Data validity 8
3.5.4 Reliability of data 8
3.6 Planned Data Analysis 8
3.6.1 Data Analysis Technique 8
3.6.2 Software Tools 8
3.7 Evaluation framework 8
3.8 System documentation 8

4 What you did 9
4.1 Hardware/Software design . . . /Model/Simulation model & parameters/. 9
4.2 Implementation . . . /Modeling/Simulation/. 9
4.2.1 Some examples of coding 9
4.2.2 Some examples of figures in tikz 9
4.2.2.1 Azure's Form Recognizer 9

5 Results and Analysis 11
5.1 Major results 11
5.2 Reliability Analysis 11
5.3 Validity Analysis 11

6 Discussion 13

7 Conclusions and Future work 15
7.1 Conclusions 15
7.2 Limitations 15
7.3 Future work 15
7.3.1 What has been left undone? 15
7.3.1.1 Cost analysis 15
7.3.1.2 Security 15

7.3.2 Next obvious things to be done 16
7.4 Reflections 16

References **17**

A Supporting materials **19**

B Something Extra **21**

List of Figures

2.1 Lots of stars (Inspired by Figure x.y on page z of [xxx]) 3

List of Tables

2.1 xxx characteristics 4

List of acronyms and abbreviations

OS	operating system

Chapter 1

Introduction

1.1 Background

As one can find in RFC 1235 [1] multicast is useful for xxxx. A number of different operating systems (OSes) have been used in this work, such as the following OSes: UNIX, Linux, Windows, etc. The main focus will be on one OS, namely Linux.

1.2 Problem

Longer problem statement
If possible, end this section with a question as a problem statement.

1.2.1 Original problem and definition

1.3 Purpose

1.4 Goals

The goal of this project is XXX. This has been divided into the following three sub-goals:

1. Subgoal 1
2. Subgoal 2
3. Subgoal 3

1.5 Research Methodology

1.6 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.

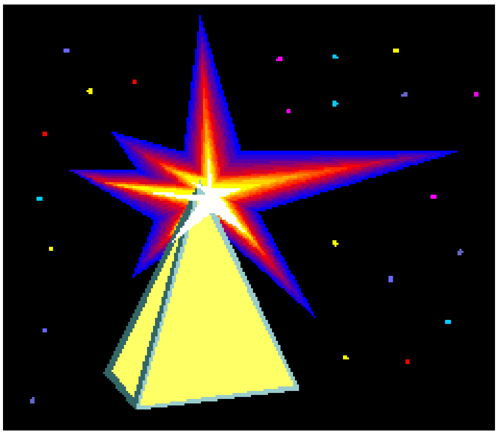

Figure 2.1: Lots of stars (Inspired by Figure x.y on page z of [xxx])

Table 2.1: xxx characteristics

Characteristics α	**Description** 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].

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 [3].

2.1.3 Subarea 1.1.2

Using the nearest data center may improve performance, see [4]

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 [5].

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

[1] J. Ioannidis and G. Maguire, "Coherent File Distribution Protocol," *Internet Request for Comments*, vol. RFC 1235 (Experimental), Jun. 1991. doi: 10.17487/RFC1235. [Online]. Available: http://www.rfc-editor.org/rfc/rfc1235.txt [Page 1.]

[2] Y. S. Kim, G. Q. Maguire, and M. E. Noz, "Do Small-Mass Neutrinos Participate in Gauge Transformations?" *Advances in High Energy Physics*, vol. 2016, pp. 1–7, 2016. doi: 10.1155/2016/1847620. [Online]. Available: http://www.hindawi.com/journals/ahep/2016/1847620/ [Page 4.]

[3] G. Q. Maguire Jr., M. E. Noz, H. Olivecrona, M. P. Zeleznik, and L. Weidenhielm, "A New Automated Way to Measure Polyethylene Wear in THA Using a High Resolution CT Scanner: Method and Analysis," *The Scientific World Journal*, vol. 2014, pp. 1–9, 2014. doi: 10.1155/2014/528407. [Online]. Available: http://www.hindawi.com/journals/tswj/2014/528407/ [Page 4.]

[4] K. Bogdanov, M. Peón-Quirós, G. Q. Maguire, and D. Kostć, "The nearest replica can be farther than you think," in *Proceedings of the Sixth ACM Symposium on Cloud Computing - SoCC '15*. Kohala Coast, Hawaii: ACM Press, 2015. doi: 10.1145/2806777.2806939. ISBN 978-1-4503-3651-2 pp. 16–29. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2806777.2806939 [Page 4.]

[5] A. Roozbeh, A. Sefidcon, and G. Q. Maguire, "Resource Monitoring in a Network Embedded Cloud: An Extension to OSPF-TE," in *2013 IEEE/ACM 6th International Conference on Utility and Cloud Computing*. Dresden, Germany: IEEE, Dec. 2013. doi: 10.1109/UCC.2013.36. ISBN 978-0-7695-5152-4 pp. 139–146. [Online]. Available: http://ieeexplore.ieee.org/document/6809350/ [Page 5.]

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": "1,23"},
"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}
```