Mittuniversitetet
MID SWEDEN UNIVERSITY

Titel

Subtitle

Alex Olsson

Bachelors Thesis

Main field of study: Computer engineering

Credits: 15

Semester/year: Spring/2025

Supervisor: Alex Olsson

Examiner: Stefan Forsström/Patrik Österberg

Course code: DT099G

At Mid Sweden University, it is possible to publish the thesis in full text in DiVA (see appendix for publishing conditions). The publication is open access, which means that the work will be freely available to read and download online. This increases the dissemination and visibility of the degree project.

Open access is becoming the norm for disseminating scientific information online. Mid Sweden University recommends both researchers and students to publish their work open access.

I/we allow publishing in full text (free available online, open access):

☒ Yes, I/we agree to the terms of publication.

☐ No, I/we do not accept that my independent work is published in the public interface in DiVA (only archiving in DiVA).

....

Location and date

....

Programme/Course

....

Name (all authors names)

....

Year of birth (all authors year of birth)

Abstract

The report shall include two abstracts, one in Swedish and one in English. The abstract acts as a description of the report’s contents. This allows for the possibility to have a quick review of the report and provides an overview of the whole report, i.e. contains everything from the objectives and methods to the results and conclusions. Examples: “The objective of this study has been to answer the question…. The study has been conducted with the aid of…. The study has shown that…” Do not mention anything that is not covered in the report. An abstract is written as one piece and the recommended length is 200-250 words. References to the report's text, sources or appendices are not allowed; the abstract should “stand on its own”. Only use plain text, with no characters in italic or boldface, and no mathematical formulas. The abstract can be completed by the inclusion of keywords; this can ease the search for the report in the library databases. The keywords should be listed in order of importance.

Keywords: Human-computer-interaction, XML, Linux, Java.

Sammanfattning

Also write the abstract in Swedish. If you write your thesis in Swedish, Sammanfattning should be before Abstract.

Nyckelord: Människa-dator interaktion, XML, Linux, Java.

Acknowledgements / Foreword

(Swedish: Förord) Acknowledgements or Foreword (choose only one of the heading alternatives) are not mandatory but can be applied if you as the writer wish to provide general information about your exam work or project work, educational program, institution, business, tutors and personal comments, i.e. thanks to any persons that may have helped you. Acknowledgements are to be placed on a separate page.

Table of Contents

Terminology / Notation

(Swedish: Terminologi/Notation) Choose one of the headline alternatives. A possible list of terms, abbreviations and variable names with brief explanations may be placed after the table of contents but is not required. Note that although a term is explained in the list of terms, it should also be explained in the chapter text where it is used the first time. This list shall be in alphabetical order.

Acronyms/Abbreviations

ACK	Acknowledge.
AWGN	Additive White Gaussian Noise

Mathematical notation

G(x)	RC generator polynomial

1 Introduction

Valmet is a leading company specializing in industrial process technologies, automation, and services, primarily in the pulp and paper industry. To enhance the management of its vast amount of report data, Valmet is exploring the use of AI-powered solutions especially now that Large Language Models (LLMs) have transformed the AI landscape in recent years, offering new opportunities for automating and optimizing data processing tasks.

1.1 Background and motivation

Large Language Models (LLMs) are a type of artificial intelligence designed to understand and generate human language. They are called large language models because of the large amounts of data that they are trained on. These can be texts from books, articles, websites and other sources, all to learn patterns, structures and relationships between words and sentences in languages. An example of this is ChatGPT and Microsoft Copilot.

Local AIs refer to artificial intelligence systems that run directly on local devices such as computers or smartphones rather on relying on cloud-based servers and systems such as ChatGPT that is an online website or application. These local AIs can be specifically designed to fit various purposes, one of which is running language models, without sending all data to external servers and devices.

Because cloud-based services rely on external servers, security can be a problem for companies and corporations that have sensitive information that needs to be handled. Therefor Valmet wants to implement a more local LLM model that can handle large amounts of scientific research reports while maintaining efficiency and security.

1.2 Overall aim and problem statement

Valmet is a large global company that is a leading developer in technology and automation services in pulp, paper and sustainable energy. Because of this they have tons of scientific reports that are currently hard to search through when you need to find a specific one about a specific problem that you have. This is a problem that Valmet believes can be solved by using a locally hosted LLM model to help search through the various amounts of reports to help workers find what they are looking for efficiently.

The aim of this thesis is to find a suitable locally hosted LLM model that can help with report processing at Valmet. The study will compare the pros and cons between two locally hosted LLM models in terms of performance, cost, efficiency, accuracy, scalability and eventual infrastructural requirements.

1.3 Research questions

1. How well do locally hosted LLMs perform in processing scientific reports?
2. What are the infrastructure and resource requirements for deploying locally hosted LLMs?
3. What is the most suitable strategy for implementing AI-based report processing?

1.4 Scope

(Swedish: Avgränsningar) Explain what you have focused your work on and what you have chosen to not focus on. What will be in focus? What will not be in focus?

This thesis will focus on two LLM models that will be evaluated on

Write this under the course

1.5 Outline

1.6 Division of work

2 Theory

(Swedish: Teori) Summarize and introduce the theory chapter. The report's theory chapter should contain additional facts required for the reader's understanding of the remainder of the report. At this point a summary of background material in the area should be provided, i.e. standards, scientific articles, books, magazines, documents on the web, technical reports and user manuals. Explain pedagogically with clear examples and many illustrations. It should be demonstrated that you have an awareness of the context and the background of your work. Also explain the aim of the technology that you describe, and not only how the technology works.

2.1 LLM

Should be changed to whatever the area is called. For example: Machine Learning, Internet-of-Things, Visualization, Network Communication, etc. One good place to find good articles for this section is to search on Google Scholar for your research area with the added keyword 'survey'.

2.2 Open-source AI

Continue with more relevant theory, background knowledge, and technologies so that the reader can understand the rest of this thesis.

Continue the theory chapter with more headings as needed.

2.3 OLLAMA

2.4 Related work

(Swedish: Relaterat arbete) End the theory chapter with a section called related work. Find a number of scientific related works/"competitors". Meaning other works that have done something similar to what you have done. One good place to begin this is work is by searching for similar works to yours on Google Scholar.

2.4.1 Example of a related work domain

Explain how your work will be similar and how you work will be different from their work. Put your work into relation to theirs.

Continue the list with one heading for each concrete piece of related work.

2.4.2 Another example of related work domain

Explain another related work.

2.4.3 Etc. until all related works has been covered

Explain another related work. Etc.

3 Methodology

(Swedish: Metod) Summarize and introduce the methodology chapter. The method section is the point at which your chosen method and intended procedure are presented. This section should contain information for the reader to interpret your results and repeat your work, i.e. in order to check the results. Here, the tools, assumptions, mathematical models, performance measures and assessment criteria are presented.

3.1 Scientific method description

The project will be more of a quantitative study since it will be measuring accuracy, response time and scalability along with assessing what hardware requirements are needed to run the LLMs. It also has some qualitative aspects such as evaluation of feasibility and integration potential along with formulating a strategic recommendation based on the observations made from tests.

The first research question that needs to be answered is how well a locally hosted LLM performs in processing reports at Valmet. This will be done by first picking out two open-source LLMs and implementing them on a computer where tests will be conducted to see how accurate the models are to find data, the response time it has and the scalability. This will determine how well the models handle Valmet's report data in terms of efficiency and reliability.

The second research question to be answered is to see what research and infrastructural needs Valmet has in deploying locally hosted LLMs. This will be evaluated based on what LLM models are picked since that will tell us what system requirements will be needed to run them. During the tests, more and more data will be used to see the scalability of the models. This in turn will tell us if the model itself will be enough to handle the data or if the system and computers that Valmet can offer is good enough to deploy them efficiently. Information will also be gathered to estimate exactly how much data Valmet has that needs to be used and what specs the local devices at Valmet has.

The last question is what the most suitable strategy will be for Valmet in implementing an AI-based report processer. This will be discussed, evaluated and answered by comparing the results of the answers of the two previous research questions to see what the best alternative for Valmet will be. To either implement one of the two tested LLM models, to maybe implement a hybrid of both, or to look elsewhere like a cloud-based solution that can still be safely implemented into the company while keeping high efficiency and scalability.

3.2 Project/Work method description

This thesis will follow the waterfall method, beginning with a theory part where research about how to solve the problem of report processing. An investigation into different open-source LLM models and what requirements they need to run on computers will be conducted to see which are possible to run on Valmet's computers.

Thereafter the pre-study will be done to research more in depth on the models and how to run them. A Pugh matrix will be used to show pros and cons with each model to help pick out the most suitable ones. OLLAMA, a framework for building and running local language models, seems like a good framework to use because of its user friendliness.

After the pre-study is done the implementation will be done where the two picked models along with OLLAMA will be downloaded and deployed on different devices. A laptop, a personal gaming computer with a GPU, a CAD computer and a computational computer, both from Valmet. This is to see if the results differ between devices since they each have different specs. The models will be set up to be able to search through pdf files containing text in the form of reports.

Measurements will then be conducted on the implemented models by sending in data in the form of pdf files of reports or other long texts and, by feeding it a search word like a name or date, it will start scanning the provided data for the keyword. During the search, measurements will be taken to see how fast and efficient it finds or don't find the word or words, how reliable it is depending on the search word and how much data it can search through before losing efficiency. This will also determine the infrastructural needs that need to be met in order to efficiently run the model/models.

Finally the measurements and results will be summarized in a table or Pugh Matrix to evaluate and show where the two models differ in performance and integrity (implementational possibility). These differences will be discussed in chapter 7 as a part of the third research question in finding the most suitable model for implementation. This will also discuss the possible problems and weaknesses encountered during the implementation of the models and the tests that were made. These could be implementation problems or problems during tests where the programs go very slow or crash because of lack of resources depending on the computer since LLM models use a lot of GPU power which a laptop and a computation computer doesn't have.

4 Choice of approach/System design/Pre-study

(Swedish: Val av angrepssätt/Systemdesign/Inledande studie) Choose only one of the headlines. This chapter will be substantially different depending on your thesis topic, direction, and scientific level. But in this chapter, you will present the different options you have faced and made your choice between, or you will explain the work that has led up to your implementation, and your analysis of different requirement.

4.1 Approach alternatives

(Swedish: Kravfångst/Angreppsalternativ) Explain the requirement capturing and the identified requirements, or the different approach alternatives.

4.1.1 LLAMA 3.3

Present and summarize the potential approaches or requirements that you have considered. Make them even in length and style, to be of equal in importance.

Continue the list with all alternatives/requirement, one heading for each.

4.1.2 Falcon Mamba 7B

Another requirement or potential approach.

4.1.3 Mistral

Another, etc.

4.1.4 LLAMA 3

https://ai.meta.com/tools/faiss/

Ha typ en hjärna, kan kolla igenom hårddisk med rapporter osv för att ge sökmotormöjligheter.

4.2 Comparison of approaches // Hårdvaru, telefon, dator, beräkning

(Swedish: Kravanalys/Jämförelse av angreppssätt) Analyze the identified requirements, compare different approaches to another, benefits, drawbacks? Pros vs cons? A table for clear comparison can be used, for example a Pugh matrix.

4.3 Chosen approach // Ramverk, gui, ollama osv

(Swedish: Föreslagen/Utvalt angreppssätt) Present your proposed or chosen approach. It is very important to motivate why this was chosen over the others. Connect back to your overall aim and problem statement, to ensure that the chosen approach actually can answer your scientific goals/research questions.

4.4 Comparison of approach

Pughsmatris

4.5 Chosen approach

Kap 2, mer generella termer. Generative AI, llm, tekniska termer kan komma i kap 4. Kolla igenom, se om andra har gjort saker, se vad man kan implementera av deras.

5 Implementation

(Swedish: Implementation) Choose only one of the headlines. In this chapter you will explain all the technical details of your work. Start this chapter with an overview of your work and an overview block figure with the protocols, algorithms, and technologies written out. Basically, as form of UML component diagram. You can also use storyboards, mock-up designs, state diagrams to show the overall structure. See Figure 1 for an example. **All figures and tables must be referenced from the text**.

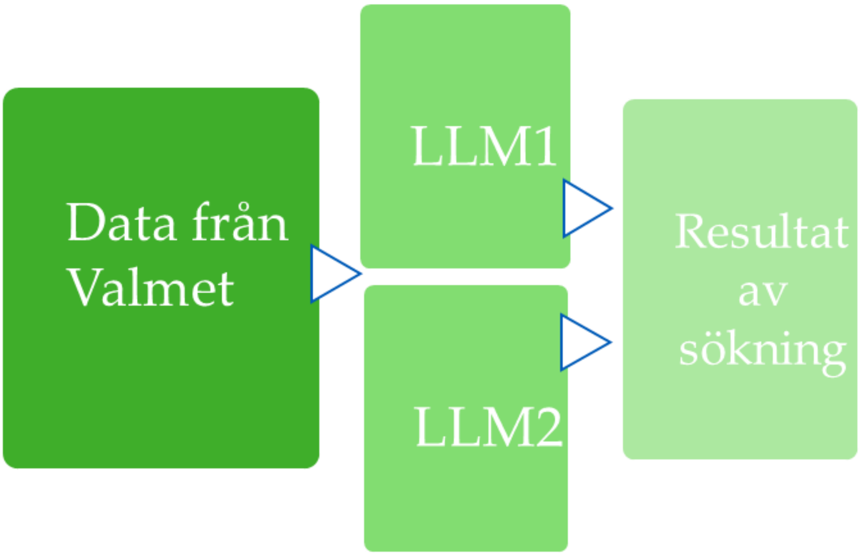

Figure 1: System overview

5.1 First part of the overview figure

Use a top-down approach, divide and conquer, split the figure into relevant pieces. Explain each piece in detail, use titles that can be found in the overview figure. Use flow charts, UML diagram, pseudocode etc. but no real code. Refer to the appendix for the actual source code. Motivate your implementation choices.

5.2 Another part of the overview figure

Another piece of the solution.

5.3 Etc. until all parts are covered from the overview figure

Another piece of the solution. Etc.

5.4 Measurement/Evaluation setup

(Swedish: Mätuppställning) Remember to explain how you will measure/evaluate your implementation and how your test bench has been built/setup/implemented.

6 Results

(Swedish: Resultat) Summarize and introduce your results

6.1 Resulting application/system

(Swedish: Färdiga applikationen/systemet) Present your resulting work. Screenshots, features, etc.

6.2 Measurement results

(Swedish: Mätresultat) Present your measurements. How well does it perform? Present the results objectively, with as little bias as possible. Use tables, graphs, etc.

You can find an example table below. See Table 1.

Table 1. Measured times for the model.

	Shortest (ms)	Longest (ms)	Average (ms)	Stdev (ms)
Model one	10	30	20	5
Model two	20	40	35	3

Here is an example of a line graph. See Figure 2.

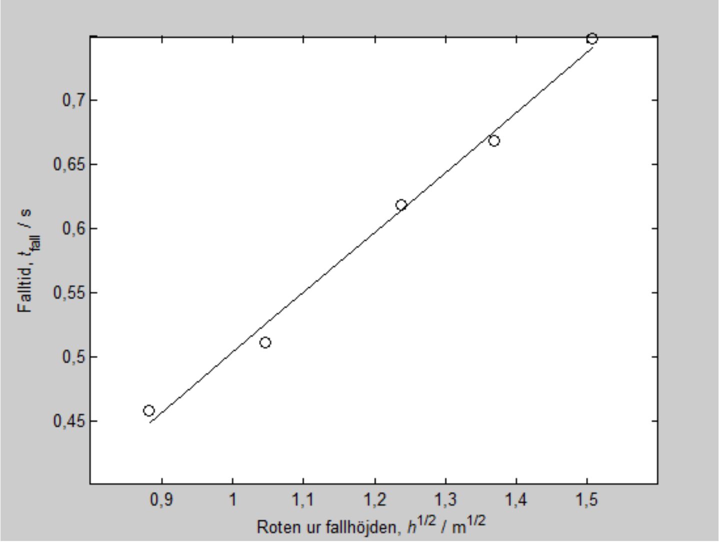

Figure 2: Line graph example.

Here is an example of a bar chart with standard deviation whiskers. See Figure 3.

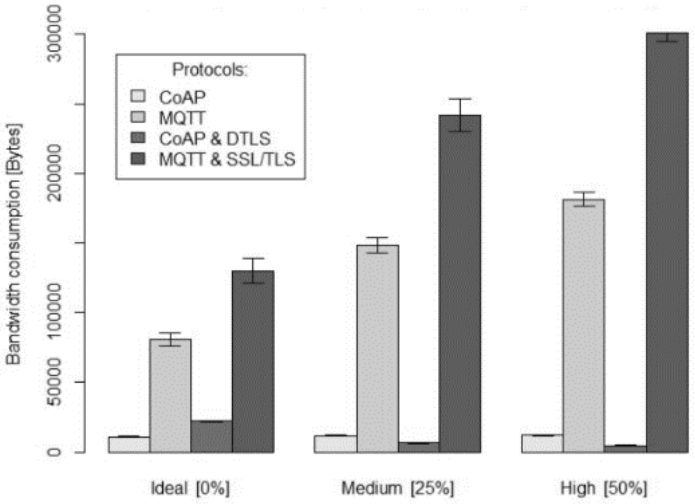

Figure 3: Bar chart example.

Below is an example boxplot. See Figure 4.

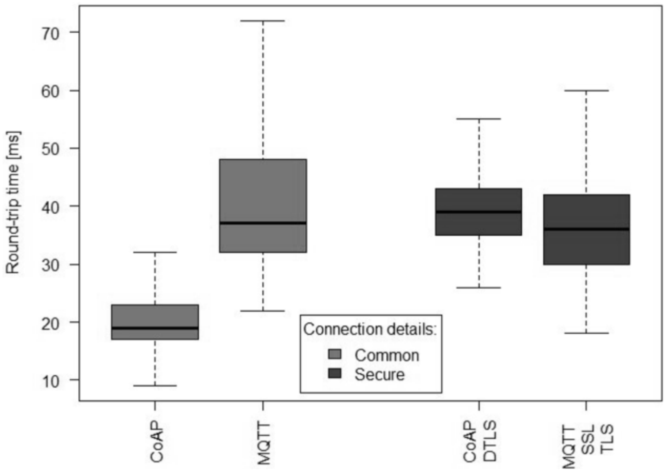

Figure 4: Boxplot example.

And here is an example equation. See Equation 1.

$$P_{i,j} = C\frac{P_i G_{i,j}}{d_{i,j}^{\alpha}}, \qquad (1)$$

7 Discussion

(Swedish: Diskussion) Summarize and introduce what you will discuss and analyze.

7.1 Analysis and discussion of results

(Swedish: Diskussion och analys av resultat) Make a deep analysis and discussion of your resulting application, measurements, evaluation, etc. Here it is ok to be more subjective/biased.

This section can and should be quite long.

7.2 Project/Work method discussion

(Swedish: Metodsdiskussion) Make a deep analysis and discussion of your chosen method, chosen approach, chosen metrics, etc. Connect back to 3.2.

Discuss all of your project milestones/phases. Have they been successful?

7.3 Scientific discussion

(Swedish: Vetenskaplig diskussion) Make a deep discussion of the gained scientific knowledge and your scientific method. Answer and discuss your original research questions/scientific goals/verifiable goals. What can be learnt? Is this knowledge general or specific, etc. Make a discussion by looking back at the related works and put your work into perspective. Remember to discuss and show insights into the scientific possibilities of this work and its limitations.

7.4 Consequence analysis/Recommendation

Discuss the scientific impact and the contribution of your work. What will be the consequences of this new knowledge? If you are working with an industry problem, what do you recommend the company to continue to pursue?

7.5 Ethical and societal discussion

(Swedish: Etiska och samhälleliga aspekter) You will need to include a discussion on ethics, societal impact, and considerations. Use the human perspective, how will we be affected by this work, was people involved in the process, privacy? Remember to discuss and show insights into this project’s role in society and our responsibilities for how it is used.

8 Conclusions

(Swedish: Slutsats) Summarize your outcome and wrap up all loose ends. Make clear conclusions regarding your goals, research questions and problem statement. Include your scientific contribution and impact.

Shortly give an answer to all of your research questions/scientific goals/verifiable goals. Have they been met?

Remember to also give an answer to your problem statement.

8.1 Future work

(Swedish: Framtida arbete) You should also explain potential future work based on your work. What new research problems or potential project have arisen after you have finished? If someone were to continue your work, what should they do?

8.1.1 Example of a future work

Remember to not just point out potential future work, also explain how you would approach this particular future work if you had the opportunity to do so.

8.1.2 Another example of a future work

Another aspect to do future work on.

8.1.3 Etc. until all future works has been covered

Explain another aspect to do future work on. Etc.

References

(Swedish: Referenser) Below is an example of a numbered list of references according to the IEEE reference standard used in technical reports (Vancouver system).

https://vivekupadhyay1.medium.com/falcon-mamba-7b-a-revolutionary-state-space-language-model-for-efficient-long-form-text-processing-7f3c7e8c2934

Appendix A: Source Code

(Swedish: Bilaga) Avoid using actual program code in the report. Instead, provide a link here to a repository or someplace where the code can be downloaded. But remember that you must mention this appendix somewhere in the report e.g. For the complete source code, see Appendix A.