CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

# Leveraging Large Language Models for Cybersecurity Risk Assessment of Autonomous Forestry Machines

Master's Thesis in Computer science and engineering

Fikret Mert Gultekin, Oscar Lilja

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2024

Master's Thesis 2024

# Leveraging Large Language Models for Cybersecurity Risk Assessment of Autonomous Forestry Machines

Fikret Mert Gultekin, Oscar Lilja

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2024

Leveraging Large Language Models for Cybersecurity Risk Assessment of Autonomous Forestry Machines
Fikret Mert Gultekin, Oscar Lilja

© Fikret Mert Gultekin, Oscar Lilja, 2024.

Supervisor: Rebekka Wohlrab, Department of Computer Science and Engineering
Supervisor: Ranim Khojah, Department of Computer Science and Engineering
Industrial Supervisor: Mazen Mohamad, Research Institutes of Sweden (RISE)
Industrial Supervisor: Marvin Damschen, Research Institutes of Sweden (RISE)
Examiner: Farnaz Fotrousi, Department of Computer Science and Engineering

Master's Thesis 2024
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Gothenburg, Sweden 2024

Leveraging Large Language Models for Cybersecurity Risk Assessment of Autonomous Forestry Machines

FIKRET MERT GULTEKIN, OSCAR LILJA
Department of Computer Science And Technology
Chalmers University of Technology

# Abstract

Large language models are a type of deep-learning model trained on massive data to generate responses to user prompts similar to natural language. Large language models can be specialized into several different domains such as medical, and e-commerce. This thesis investigated how cybersecurity experts can benefit from large language models in the risk assessment process in the forestry domain. This research was carried out in collaboration with the Research Institutes of Sweden and Chalmers University of Technology. This thesis is a part of the EU-funded AGRARSENSE project. We conducted a design science study including 15 interviews, 12 demos, and a survey. We used local Llama 2 7B and developed an RAG application by supplying the model with data relevant to cybersecurity and the AGRARSENSE project. We created a generated risk assessment document using the tool. The tool is the main artifact of this study. Also, we analyzed several articles. The study demonstrated that large language models can be used in multiple ways in a risk assessment process such as an evaluation tool, assisting chatbot, or generating risk assessments. The findings showed that trust remains an issue for large language models. Even though cybersecurity is one of a software system's most critical work areas, experts are willing to use such LLMs. Experts are willing to use evaluation and assisting features more than the generation feature.

Keywords: LLMs, risk assessment, cybersecurity, RAG, autonomous machinery, forestry, artificial intelligence

# Acknowledgements

We want to thank our supervisors Rebekka Wohlrab, and Ranim Khojah, and our industrial supervisors Mazen Mohamad and Marvin Damschen, for their precious feedback, support, and guidance. We would like to thank our examiner Farnaz Fotrousi for evaluating our work and giving feedback. Additionally, we would like to thank all the experts for their participation, and valuable feedback. Lastly, we would like to thank our family and friends for their infinite support.

Fikret Mert Gultekin and Oscar Lilja, Gothenburg, 2024

# List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

| | |
|---|---|
| AI | Artificial Intelligence |
| GPU | Graphical Processing Unit |
| IEC | International Electrotechnical Commission |
| ISO | International Organization for Standardization |
| IACS | Industrial automation and control systems |
| IoT | Internet of Things |
| LLM | Large Language Model |
| ML | Machine Learning |
| RAG | Retrieval Augmented Generation |

# Contents

# List of Figures

# List of Tables

# 1
# Introduction

Autonomous machinery is increasingly becoming an important tool in industries like agriculture, mining, and forestry [57],[51],[25], [60] due to safety issues and cost-effectiveness. Integrating such machines in forestry, a sector with great importance for global sustainability [66], introduces new challenges in cybersecurity such as data confidentiality, identification and authentication control [63]. Both autonomous machines and vehicles critically rely on software systems, and understanding the vulnerabilities of the software is vital to preventing cyber attacks [41] which can lead to significant consequences such as financial loss and loss of reputation. Such an understanding can be obtained by creating risk assessments [74]. The main goal of this study is to explore how Large Language Models (LLMs) can be leveraged to support cybersecurity engineers in creating risk assessment processes. Large language models are pre-trained deep learning models [79], [10] that are used for several purposes such as summarization of a text, question-answering, and chatbot.

Traditional risk assessment in areas such as forestry has been manual [64] and labor-intensive, requiring experts to identify and evaluate potential risks and their impact. Human limitations in handling large volumes of data and a tendency to miss details in complex systems are ever-present challenges [6]. This study aims to evaluate and assess the ability of LLMs to support practitioners in conducting cybersecurity risk assessments by, for example, assisting with the manual work, evaluating performed risk assessments, or even creating these assessments entirely. This research goal is to narrow the gap between advanced AI technologies and their practical application in critical sectors.

To achieve these goals, in this thesis we collaborated with RISE (Research Institutes of Sweden) as a part of the AGRARSENSE EU project (Grant Agreement No. 101095835) [16]. AGRARSENSE is a project launched in January 2023 and provides the context for our research. The AGRARSENSE project includes 52 partners from different European countries and aims to develop cutting-edge technology in agriculture and forestry [16]. AGRARSENSE is a massive project with multiple broad goals for the coming decades. This thesis will focus on cybersecurity, aligning with the overarching goal of improving efficiency and reliability in autonomous systems in the agriculture and forestry domains.

## 1.1 Problem Statement

Autonomous machinery systems have cybersecurity vulnerabilities. For instance, machinery can be accessed and controlled by unauthorized people, which can result in several consequences. The consequences can be threats to human safety, and the environment. Cybersecurity threats can compromise the integrity and availability of an organization's

services in several ways such as operations, mission, functions, critical services, image, and reputation [58]. Some specific cybersecurity issues within the forestry environment are reviewed by Mohamad et al. [57]. The same study claimed that cybersecurity risk assessments are specific to the context and environment. This means that cybersecurity risk assessments are time-consuming, and require expertise and need for domain knowledge which is a problem.

For many sectors, there are standards specific to the domain. Some examples are, the earth-moving machinery and mining standard ISO 17757 2019 [1], Industrial truck standards ISO 3691-4 2020 [4], and "Agricultural machinery and tractors" standard ISO 18497 2018 [20]. However, based on our knowledge, there is a lack of industrial cybersecurity standards for autonomous forestry machinery.

Additionally, cybersecurity experts always need to decide the assessments, threats, threat actors, or similar things. However, they suffer from the absence of support for decision-making in the risk assessment process because there was no specialized large language model for risk assessment. Also, one study suggested that due to the absence of automation in conducting risk assessments, performing manual risk assessments considering safety and cybersecurity is required by cybersecurity professionals [76] manually which is time-consuming and costly. We searched for solutions to the problems mentioned in this study.

## 1.2 Purpose

The purpose of the study was to explore how large language models can help cybersecurity experts in the risk assessment process, specifically within the forestry domain. This was done by evaluating the applications of a large language model within the context of AGRARSENSE. Thanks to this study, researchers and cybersecurity experts might be able to lessen cybersecurity hazards in software systems in this field. Additionally, one of the purposes of this study was to increase the speed and accuracy of the manual risk assessment process by collaborating with LLMs.

In addition, the study identified the potential benefits and limitations of implementing large language models in risk assessment.

## 1.3 Research Questions

The research questions were split into three parts: *generation* (**RQ1**), *assistance and collaboration* (**RQ2**), and *evaluation* (**RQ3**).

**RQ1** *To what extent can large language models generate cybersecurity risk assessments in the context of autonomous forestry machinery?*
This question aimed to explore LLMs' capability to create risk assessments, examining both the quality and usefulness of the output and potential challenges and limitations.

**RQ2** *To what extent can LLMs assist and collaborate with practitioners in creating and enhancing manual cybersecurity risk assessments for autonomous forestry machinery?*
This question focused on the collaborative potential between human experts and LLMs, investigating usefulness, use cases, limitations, and challenges.

**RQ3** *To what extent can LLMs be utilized to evaluate and improve existing cybersecurity risk assessments?*
This research question's goal was to investigate the potential for LLMs to improve already existing sections of cybersecurity risk assessments by understanding their capability, usefulness, and shortcomings.

In order to answer these research questions, we used a combination of interviews, surveys, and demos, and we collected data by investigating the experts' perceptions of several key aspects of the LLMs' output. These were *biases, completeness of output, correctness, level of detail in the output, ability of the LLM to reason, relevance of information, reliability of the LLM, adherence to a predefined standard, understandability of output, and usefulness for each aspect of the research questions.* These metrics provided the foundation of the results and were the basis for our conclusions.

## 1.4 Significance of the study

The result of the study will allow practitioners to start considering using Large Language Models in the risk assessment process. Because the workload of cybersecurity experts in the risk assessment process will decrease thanks to the automatic risk assessment [24]. This means that operational costs would be reduced. Considering the few security experts in this field makes everything more complicated and expensive. The significance of this study increases, considering the scarcity of research in this field. Moreover. The study will bring a chance to review existing risk assessments to find overlooked errors.

**For researchers**, our goal in this study was to contribute to filling the gap in the literature in the context of autonomous forestry machinery because according to our literature scanning, this subject has not been studied deeply. The study provided a vision for the potential limitations and benefits of LLMs for risk assessment by providing empirical evidence.

**For practitioners**, this study facilitated an improved decision-making process by conducting more comprehensive evaluations of potential risks. One of the goals was for the findings of this study to serve as a foundation for future guidelines. The study might enhance the effectiveness and productivity of cybersecurity strategies within the forestry sector.

# 2
# Background

This chapter introduces relevant terminology and topics related to large language models and cybersecurity risk assessment. In addition to that, the chapter briefly explains the AGRARSENSE project.

## 2.1 Large Language Models

Large language models are a type of deep-learning model [79], [10]. They are trained on massive data. Their purpose is to generate responses to user prompts. LLMs can be specialized for tasks, such as question answering, text summary, text generation, text classification, and other text-related tasks.

Generally, pre-trained models are easier to use than training a large language model. Because training a large language model is time-consuming and requires lots of data, computational resources, and expertise. For instance, the Llama 3 model is trained on over 15 trillion tokens. Llama 3 8B model training took 1.3 million hours on GPU using the hardware H100-80GB [1]. Pre-trained LLMs are usually general-purpose language models and may struggle with domain-specific tasks [37]. Some techniques, such as retrieval augmented generation [44],[83], fine-tuning, and few-shot learning perform quite well for domain-specific tasks [75].

### 2.1.1 Hallucination Problem

One of the main disadvantages of the large language models is hallucination. LLMs can generate output completely or to some degree unrelated content[35]. The models also make mistakes or ignore instructions.

One of the main reasons for the hallucination is that as the models get bigger, they tend to hallucinate more. However, there is no direct correlation between model size and hallucination. Instead, training data quality, lack of explicit training on facts, and overconfidence in generated responses are directly correlated to hallucinations. [67]. The authors of this study shared that hallucination increased when the model was not trained with the 'reinforcement learning from human feedback' method. Also, one study proposed categorized the RAG applications into three based on their complexity: naive, advanced, and modular RAG. The study suggested that naive RAG applications can lead to hallucinations [26]

[1] `https://huggingface.co/meta-llama/Meta-Llama-3-8B`

### 2.1.2 Benchmark Studies

The rise in LLM usage and research led to several studies comparing LLMs. The comparisons were based on criteria, trustworthiness, toxicity, privacy, and robustness. Some studies included benchmarks such as HELM [47], Alpaca Eval [46], BIG Bench [9], LMSYS chatbot arena [14], and so on. Also, the Hugging Face online platform had thousands of models, datasets, demo applications, and a leaderboard. [8].

Studies such as HuggingFace open llm leaderboard [8] or Helm were composed of multiple benchmarks and metrics. For instance, the HuggingFace platform used TruthfulQA to measure if a model tends to reproduce false information that exists in the online environment[2].

### 2.1.3 Prompt Engineering

Prompt engineering techniques are used for effective interaction with the LLMs [55]. Zero-shot and Few-Shot promoting are just two widely used prompt engineering techniques. [3].

If a prompt is not fed with any labeled example, this is called zero-shot prompting. Instructions are directly passed to a model in this technique without an example.

Few-shot prompting is another prompt engineering technique. This technique provides models with a few labeled example prompts to make accurate predictions. In this way, context learning is enabled. If there is only one labeled example, this is called one shot. N number of labeled examples in the prompt can be called n-shot but this term is called few-shot prompting. [11]

To illustrate few-shot prompting, one can write an example prompt such as:

```
I liked the movie!  // Positive
It is just a bad movie!  // Negative
Amazing, that movie was mind-blowing!  // Positive
What a disaster program is that!  //
```

In this prompt example, there are three labeled examples. An LLM will probably say "What a disaster program is that" which is semantically negative.

However, few-shot prompting has some limitations such as complex reasoning tasks, and based on this paper [11], the technique on the GPT3 model did not always perform well when solving complex reasoning tasks.

### 2.1.4 Retrieval Augmented Generation

Retrieval augmented generation is a technique teaching LLM-specific tasks with additional data. The RAG technique is a less costly alternative than fine-tuning when a

[2] https://huggingface.co/spaces/open-llm-leaderboard-old/open_llm_leaderboard
[3] https://www.promptingguide.ai/

user desires to specialize the LLM applications to reason about closed-sourced data [44]. Because external knowledge updates and data processing requirements are minimal [26]. RAG is an alternative to prompt engineering and fine-tuning for better and more proper responses. The same study stated another benefit of RAGs is reducing the hallucination. The reason is that each output is sourced from retrieved evidence.

As described in the figure 2.1, data is embedded into vector chunks. The retrieval is executed by combining LLM's output with chunks retrieved from the vector database. With this technique, cybersecurity-related documents can be safely integrated into a local LLM.

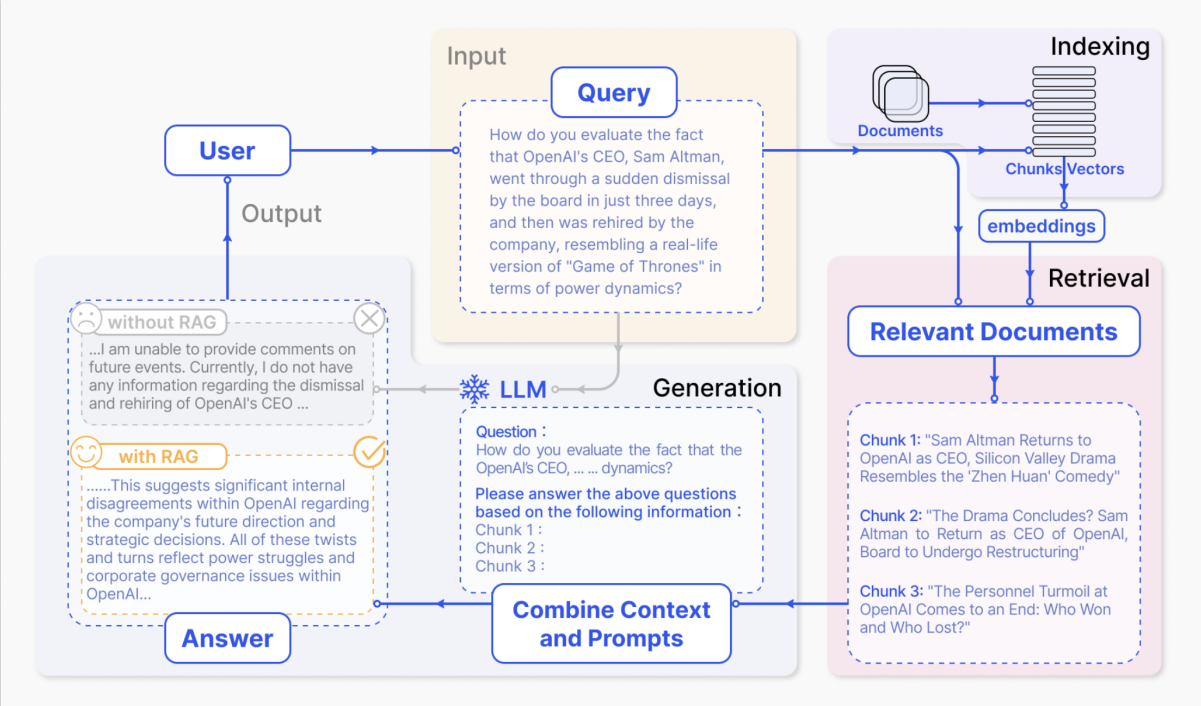



**Figure 2.1:** RAG workflow [26]

### 2.1.5 Quantization

Quantization is a method for reducing the memory consumption of a large language model by representing the data at a lower precision [19], [88], [50], [13], [71]. Normally model weights are represented as 32-bit float numbers. With 4-bit quantization, memory consumption is reduced 8 times to 4-bit weights. With 8-bit quantization, memory consumption is reduced to 8 bits. Without additional data memory, 7B parameter large language model memory consumption is roughly 28GB of VRAM with FP32 representation. Thanks to quantization, this can be reduced to approximately 3.5 GB of VRAM with 4-bit and around 7 GB of VRAM with 8-bit quantization [85], [11]. Dettmers, Tim, et al. showed that memory consumption without performance degradation was possible with multi-billion scale models.

BitsAndBytes, GPTQ, AWQ, and some other quantization libraries are commonly used and are differentiated according to their abilities. For instance, some support GPU, some only support CPU, some support 1-bit and 2-bit quantization but some support 8-bit

quantization, etc [4]. How we used quantization is explained in the methodology section.

### 2.1.6 Architectures in LLMs

A new deep learning architecture was introduced by some Google scientists in 2017 introducing transformer architecture. [79]. The transformer architecture was not designed only for large language models, it is used in various domains including computer vision, machine translation, and speech processing [49], [38], [61] In 2018 GPT, the first pre-trained transformer model, was published.

Generally, 'transformers' architecture LLM models are composed of two blocks; Encoder and decoder. Architectures are differentiated based on these blocks. There are currently three different architectures in transformer architecture. These are encoder-only, decoder-only, and encoder-decoder models. Encoder-only models are good at tasks in which understanding the input is needed. Sentence classification is an example task for encoder-only models. These models are also known as BERT-like or auto-encoding models. Decoder-only models are good at generative tasks such as text generation. These models are also called GPT-like or auto-regressive models. Encoder-decoder models also known as sequence-to-sequence are good at generative tasks with input understanding requirements such as translation [23].

### 2.1.7 Optimizing Inference

Large language models require lots of memory to make inferences [19]. It is possible to use CPU and GPU to make inferences [5] [6]. GPUs support parallel computation [42] which provides more performance gain than CPUs. For GPU inferences, one can benefit from libraries like FlashAttention-2 [17] or BetterTransformer [7] but for CPU it is possible to use TorchScript [8] or BetterTransformer. Such frameworks can optimize inference in terms of performance if they support the LLM.

### 2.1.8 Frameworks and Libraries

Many steps in LLM development require frameworks/libraries to facilitate effectiveness. Some open-source frameworks such as LangChain and LlmamaCpp, libraries such as "Transformers", and platforms like CUDA are just a few examples that were used in this study.

LangChain is an open-source tool that helps build prompt chains and customize existing templates. With the LangChain developers do not need to re-train or fine-tune the model[9].

Llamacpp is another open-source framework that helps inferences in some large language models. Its GitHub description says that the goal of this tool is to minimize the setup

[4] https://huggingface.co/docs/transformers/v4.44.0/quantization/overview
[5] https://huggingface.co/docs/transformers/perf_infer_cpu
[6] https://huggingface.co/docs/transformers/perf_infer_gpu_one
[7] https://huggingface.co/docs/optimum/bettertransformer/overview
[8] https://github.com/pytorch/TensorRT
[9] https://www.langchain.com/

and use the performance of different hardware for inferencing [27].

The 'transformers' is an open-source library that provides API for downloading pre-trained models. This library provides tools for training [84]. Note that it is not the same concept as transformer architecture. Llama 2 model which was used in this thesis based on 'transformers' architecture

CUDA is a parallel computing platform that hosts several libraries, SDKs, and profiling and optimization tools. On the other hand, the CUDA toolkit is a development environment that enables GPU inferences for LLMs. CUDA Toolkit includes GPU-accelerated libraries, a compiler, development tools, and the CUDA runtime [10].

### 2.1.9 Overfitting

Overfitting is a problem in machine learning models. If data do not generalize well from the training data, this is called overfitting [2]. For instance, a rag model is only fed with football data and prompted for general questions, in that case, the output is likely related to football rather than general sports. One of the solutions for preventing overfitting is enriching the training data [87]. Similarly, the data quality affects the model performance noted by the same study. To improve data quality, noise in the data should be eliminated by using data pre-processing methods [52].

## 2.2 Cybersecurity Risk Assessment

Cybersecurity risk assessment refers to the process of identifying and evaluating vulnerabilities and threats in a software system. This involves assessing risks, estimating likelihoods, and determining the impact of risk events. Creating such risk assessments is a labor-intensive process requiring great effort, time, and money [36].

The goal when conducting cybersecurity risk is to combat threats from threat actors like hackers, insiders, or terrorists in a proactive way. If such threats are realized, they can have long-lasting effects on the economy, infrastructure, environment, and safety, potentially putting human lives in danger. [18].

Human limitations in handling large volumes of data and a tendency to miss details in complex systems are ever-present challenges [36]. Many recent studies have investigated the potential of automating this process [36]. Automating the cybersecurity risk assessment would indeed save both time and money. However, it is important to note the potential risks associated with completely automating the process of risk assessment. The main purpose of performing a risk assessment is not only to procure a list of risks but also to review the system, understand complex relationships, and ultimately improve the system as a whole [22]. If the assessment is entirely automated and only produces the final result, certain aspects of this process might be overlooked. Our research tried to narrow the gap between AI technologies and their practical application in critical sectors.

[10] `https://docs.nvidia.com/cuda/doc/index.html`

### 2.2.1 Standards

Even though different cybersecurity risk assessments may look dissimilar, the assessment documents often share common parts, such as identified threat actors, identification of what needs to be protected, risk identification, and many others. One way of creating more cohesive, comprehensive, and stringent risk assessments is by using standards. Standards ensure the procedure is carried out in the same way each time. Standards are often domain or use-case-specific and tailored to the particular needs and challenges of the domain [36]. Standards are developed by large international organizations such as the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC), which create globally recognized standards. Standards ensure consistency throughout a domain. Practitioners often learn to work with established standards relevant to a particular domain to ensure that risk assessments can be understood by everyone working in the sector and that the results and methods applied are universally recognized.

### 2.2.2 IEC 62443

To the best of our knowledge, no standard has been specifically developed for cybersecurity for autonomous forestry machines. Therefore, this study adopted a standard created for a broader scope. The chosen standard was IEC 62443, which was developed by the IEC. The standard focuses on industrial automation and control systems (IACS) cybersecurity. These standards provide a comprehensive framework to address and mitigate security vulnerabilities in industrial environments [15].

This standard is widely recognized and uses best industry practices to assess security on multiple levels for a system. By having access to previous risk assessments conducted according to this standard done in other domains, the study objected to adopting it in a new context based on the training of the LLM model. A cybersecurity risk assessment document can easily be hundreds of pages long thus to maintain a manageable scope focus was put on the initial steps of the risk assessment.

## 2.3 The AGRARSENSE Project

AGRARSENSE is a research project launched in January 2023 and provides the context for our research. The project consisted of 52 partners from 15 European countries and aims to develop new technology for agriculture and forestry [16]. One of these partners is RISE (Research Institutes of Sweden), which is the industry partner for our thesis. RISE is a state-owned research institute with five divisions: Bioeconomy and Health, Built Environment, Digital Systems, Materials and Production, and Safety and Transport [68]. The safety and transportation division has provided invaluable access to experts in both safety and cybersecurity in risk assessment to assist in the research and evaluate the results. The overarching goal of AGRARSENSE is to develop cutting-edge technology related to seven explicit use cases, which are [7]:

- **Greenhouses**
- **Vertical farming**
- **Precision viticulture**
- **Agriculture robotics**
- **Forestry machinery**
- **Optimal soil management and fertilization**
- **Agriculture-related water management**

Our research was entirely centered on the *forestry machinery* use cases and further zooms in to focus on the cybersecurity aspects of the use case. A forestry scenario with autonomous machinery has been developed at RISE as the basis for safety and cybersecurity research [57]. The scenario consisted of a forestry worksite hosting both manually and autonomously operated machinery. Figure 2.2 shows a drone surveying an active work site where a log-transporting shuttle is working. A potentially hazardous situation can be avoided by having the drone detect the obstacle from a secondary point of view.

**Figure 2.2:** Drone aiding autonomous machine in detecting a hazardous situation from an additional point of view. [57]

Every autonomous machine monitors the worksite with sensors and actively communicates with each other for a more comprehensive picture and a safer worksite. Figure **??** is a general illustration of what the active worksite will look like, including autonomous

forwarders, drones, and human-operated harvesting machines. This scenario forms the centerpiece for the cybersecurity risk assessment created by the LLM in the study.

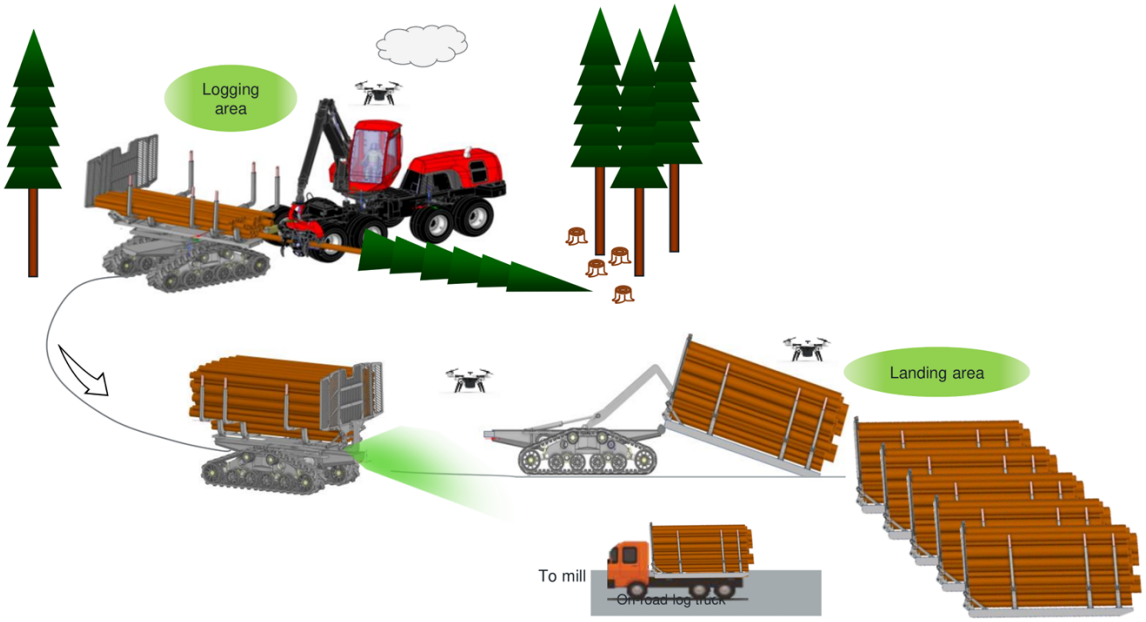



**Figure 2.3:** Overview of active forestry worksite. Image courtesy of Komatsu Forest AB. [57]

# 3
# Related Work

More autonomous machines are deployed in the forestry sector each year. These include driverless vehicles and tree-planting drones, which need to be capable of operating in unpredictable conditions. This new technology aims to increase efficiency and reduce emissions in forestry as a whole. [69] [32].

The 'Forestry 4.0' concept represents a big step towards automating and digitalizing the forestry sector. Like many other sectors, forestry is currently in the process of integrating Internet of Things (IoT) technology, big data analysis, and AI technologies into its operations. This can include, for example, using sensors, drones, and other technology to gather and analyze data to make more informed decisions and improve efficiency. These technologies are not only being used on the worksite but are also quickly becoming a part of the entire forestry supply chain. These new interconnected systems are more vulnerable to cybersecurity threats and have to be protected [65].

As Produção, et al. note: *Although the forest sector might be a priori less exposed compared to other sensitive sectors (e.g., defense, banking), cybersecurity is perhaps one of the most challenging issues* [65, p.17]

In more mature domains where autonomous machinery has been used for a long time, there are established standards for risk assessment regarding the machinery. These standards provide a framework for identifying, analyzing, and mitigating risks connected with the autonomous machinery. Examples include technical design standards and safety-related guidelines developed and refined over time to meet the specific needs and challenges of for example the agriculture sector [73].

To the best of our knowledge, the forestry sector, particularly concerning autonomous machinery, lacks such standardized risk assessment protocols. Forestry's unique environment and operations offer new challenges, such as diverse terrain, remote locations, and complex interactions with the ecosystems. These call for specific risk assessment standards. However, these have yet to be fully developed or standardized in the same way as in agriculture. This gap highlights a critical need for the forestry sector to develop and implement comprehensive risk assessment standards.

The use of LLMs in safety and security engineering is an emerging research area highlighted by many recent studies. One such study explored how LLMs and connected concepts, such as prompt engineering, can be used in the process of specifying safety requirements in the automotive domain [59]. In the paper, the author concluded that LLMs showed promise in automating and supporting the requirements engineering process. However, many limitations are still present, such as the need for prompt engineering

and the challenges of guaranteeing accuracy and relevance in the generated requirements. This study slightly differentiates from our work by looking at safety instead of cybersecurity concerns.

Generative large language models have proven their ability to generate impressive responses; however, they face significant challenges due to their tendency to hallucinate truths and yield misleading information [54]. The reliability of large language models is not high due to many reasons including hallucination, or accuracy. Even though LLMs especially ChatGPT are not trustworthy, this paper demonstrated that the usefulness of such tools is high by surveying software engineers in tasks including learning new concepts and making better decisions [39]. Reliability is an essential metric for cybersecurity risk assessment because of its potential consequences. Another study also shows that reliability varies in different domains. It is an issue in topics such as law and science [72]. The study found that even a single character modification might result in different reliability opinions.

# 4
# Methods

This section outlines the general structure of the research by describing how the design science cycles were developed. Secondly, a short overview of each phase is presented and lastly, some of the more significant activities are described in greater detail.

## 4.1 Design Science Research

The thesis adopted a design science research approach that involved two iteration cycles, broadly based on the general steps and distinct phases outlined by *Vaishnavi and Keuchler(2004)* presented in Figure 4.1 [78]. However, these general steps were customized to fit better into our research as described in the next section.

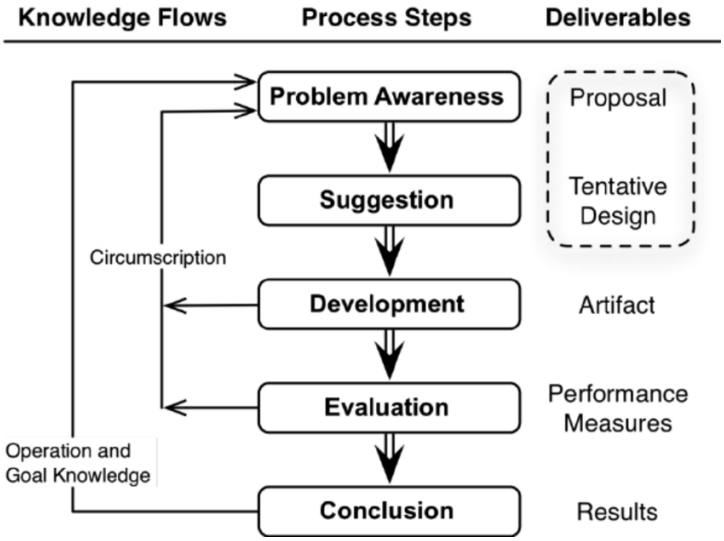



**Figure 4.1:** General steps of design science research [78]

The main goal of the first iteration was to design and develop the main artifact of this study. The artifact was a customized Large Language Model (LLM) specialized by using an RAG implementation for cybersecurity risk assessment in the context of autonomous forestry machinery used to answer the research questions. The second iteration focused on improving the artifact based on interviews and feedback from industry experts.

The study was part of a research team consisting of 15 researchers within AGRARSENSE, specializing in dependable transportation systems with expertise in various domains such as cybersecurity and safety, and provided the majority of the interview participants. Additionally, the team of experienced researchers provided assistance and support throughout each project phase.

### 4.1.1 Customized iteration cycles

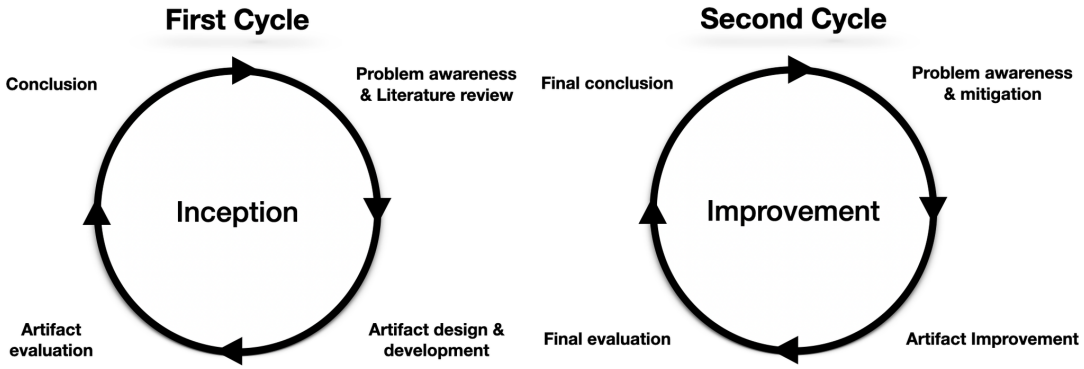
The five general steps of design science research proposed by Vaishnavi and Keuchler [78] and presented in Figure 4.1 provided the foundation for research methodology and planning. However, the phases were adjusted by considering the guidelines outlined by Hevner et al. to be better tailored to our specific context [31]. Additionally, guidelines recommended by Knauss explicitly focused on complementing Hevner's work for master's thesis work in industry were also applied throughout the study [43].

**Hevners' guidelines [78]:**

- *G1:* **Design as an Artifact.** Design-science research must produce a viable artifact in the form of a construct, a model, a method, or an instantiation.
- *G2:* **Problem Relevance.** The objective of design-science research is to develop technology-based solutions to important and relevant business problems.
- *G3:* **Design Evaluation.** The utility, quality, and efficacy of a design artifact must be rigorously demonstrated via well-executed evaluation methods.
- *G4:* **Research Contributions.** Effective design-science research must provide clear and verifiable contributions in the areas of the design artifact, design foundations, and/or design methodologies.
- *G5:* **Research Rigor.** Design-science research relies upon the application of rigorous methods in both the construction and evaluation of the design artifact.
- *G6:* **Design as a Search Process.** The search for an effective artifact requires utilizing available means to reach desired ends while satisfying laws in the problem environment.
- *G7:* **Communication of Research.** Design-science research must be presented effectively both to technology-oriented as well as management-oriented audiences

These guidelines were considered when developing our own customized design science cycles. One example of how these were applied was the design of the evaluation stage, where we made sure the evaluation was rigorous by collecting data from three sources, namely interviews, surveys, and demos, in accordance with G3. On the other hand, one example of divination from the recommendations by Knauss was the decision to use two iterative cycles instead of three. This was done to ensure that enough time was given to research existing literature and develop the initial LLM in the first iteration as we realized this would be a major milestone requiring more time investment than any other phase. Additionally, we decided that having a separate *suggestion* phase was unnecessary and was instead incorporated into the problem awareness stage. This resulted in the following customized iteration cycles.

**Figure 4.2:** Customized steps of design science research




### 4.1.2 Overview first iteration cycle

During the *problem awareness and literature review*, emphasis was put on researching existing studies and literature to make sure we had the information we needed to begin the artifact development. Awareness of the problem came both from the literature and from discussions with the research team at RISE working on the AGRARSENSE project. Major focus areas were understanding what LLM would best suit our purpose and what potential challenges we might face. The *artifact design and development* phase included planning and implementation of a local LLM. This was the most time-consuming phase, as the implementation had to be completely local to avoid leaking confidential information from the AGRARSENSE project. The *artifact evaluation* was done by conducting interviews with industry experts to find improvements and new use cases for the second iteration. The results from the interviews were collected in the *conclusion* phase by identifying key insights before the final iteration.

### 4.1.3 Overview second iteration cycle

For the start of the second cycle, a workshop was held in the *problem awareness and mitigation* phase to discuss how the challenges identified in the first iteration could be met and how to implement some of the improvements suggested by the experts. In the *artifact improvement* stage, new training data was added to the LLM, such as existing risk assessment documents following standard IEC 62443, to fine-tune the model with a focus on following a predefined standard. The output was once again evaluated in the *final evaluation* stage with a focus on the quality of the results and the usefulness of the LLM. Finally, the results were synthesized and analyzed in the *final conclusion* phase.

## 4.2 First Cycle

### 4.2.1 Artifact implementation

Llama 2 7B was selected as the main model to be used as an artifact for this study. Please refer to section 5.1 for the reason behind this selection. After the selection, the basic retrieval augmented generation technique was implemented. The application was encapsulated with docker and some memory issues were handled.

### 4.2.2 Evaluation for first iteration

#### 4.2.2.1 First iteration interviews

For the first iteration, we conducted semi-structured interviews with cybersecurity experts from RISE intending to gather qualitative data as the basis for improvements to the LLM. By interviewing practitioners with real-life experience of risk assessment we aimed to identify weaknesses, potential enhancements, and other recommendations for the next second iteration.

The questions developed for the first iterations were designed to be open-ended in nature to encourage more detailed responses. The interview guide for the first iteration is available in appendix A.1. The questions focused on various aspects of the model and the output to allow a breadth of inputs. Following the same line of thinking the structure of the interviews was semi-structured to allow flexibility in each interview, this allowed each interviewee to focus on what they felt was most crucial or promising in the material. However, each interview was based on the same original questions to ensure consistency.

Each interview participant was picked from RISE and had previous experience with risk assessments as presented in Figure 6.1. The interviews were scheduled to last between 45 to 60 min and were conducted either online or in person according to the preference of the participants. Each interview was recorded and transcribed to facilitate a detailed analysis.

#### 4.2.2.2 First iteration data analysis

The transcripts and recordings were thoroughly analyzed and each statement was assigned a category that corresponded to the context or general theme of the statement. This was similar to a miniaturized content analysis as we tried to find emerging patterns to prioritize the most important changes and improvements for the second iteration. As the focus on the first iteration was to find areas where the LLMs capabilities could be enhanced we chose three categories to encapsulate the findings from the interviews. The categories were, improvements, use cases and challenges as presented in table 6.2.

## 4.3 Second Cycle

### 4.3.1 Problem Awareness

#### 4.3.1.1 Workshop for IEC 62443

Standard IEC 62443 was decided to follow for the risk assessment after the first cycle of interviews. To obtain comprehensive knowledge about the IEC 62443 and its steps, a workshop with a cybersecurity expert from RISE and thesis supervisors was conducted. The expert directed the workshop and shared his knowledge of the standard by giving an example in the railway domain. After that workshop, prompts were updated based on the desired results. These results were similar to the railway domain example in terms of structure.

### 4.3.2 Artifact improvement

More data was provided to the vector database, data processing methods were applied and some hyperparameters were configured to improve the artifact. The risk assessment and evaluation document was created with many prompt trials. Note that the few-shot learning technique was utilized neither in the first nor second cycle.

### 4.3.3 Evaluation for iteration 2

Data was collected in three ways in the second iteration: interviews, demos, and surveying participants. The data collection aimed to get evaluation feedback to answer research questions. Feedback was obtained from the participants by focusing on different aspects of the artifact. The interview focused on the generated risk assessment document and evaluated part by part, the demo focused on the usability of the model and user interaction, and the surveying focused on overall feedback on the Likert scale. Likert Scale includes 7 options namely: Strongly disagree, moderately disagree, slightly disagree, slightly agree, moderately agree, strongly agree, and "I don't know".

#### 4.3.3.1 Interview Design for iteration 2

The purpose of the interview study in the second iteration was to try to answer research questions and get ideas for improvement on the model and its results. For this purpose, we conducted 12 interviews with professionals at RISE and Komatsu. Most of the experts have a cybersecurity background and few of them are in safety however cybersecurity experts mostly interplay between cybersecurity and safety. Experience of the experts changes between 2 years to 25 years. All the information was presented in table 6.3 in the results section.

The interviews were both online. The setting of the interview was semi-structured so that it was possible to ask some pre-written questions as well as follow-up questions based on the conversation. The prepared questions consist of open and some closed questions. During the interview, some probing questions were asked after some open questions to follow the funnel structure [80]. For the interview guide for each iteration, please look at the Appendix A.2.

During the interview, the interface as in the picture 4.3 was used. This interface was a pre-designed interface by the gradio package [5].

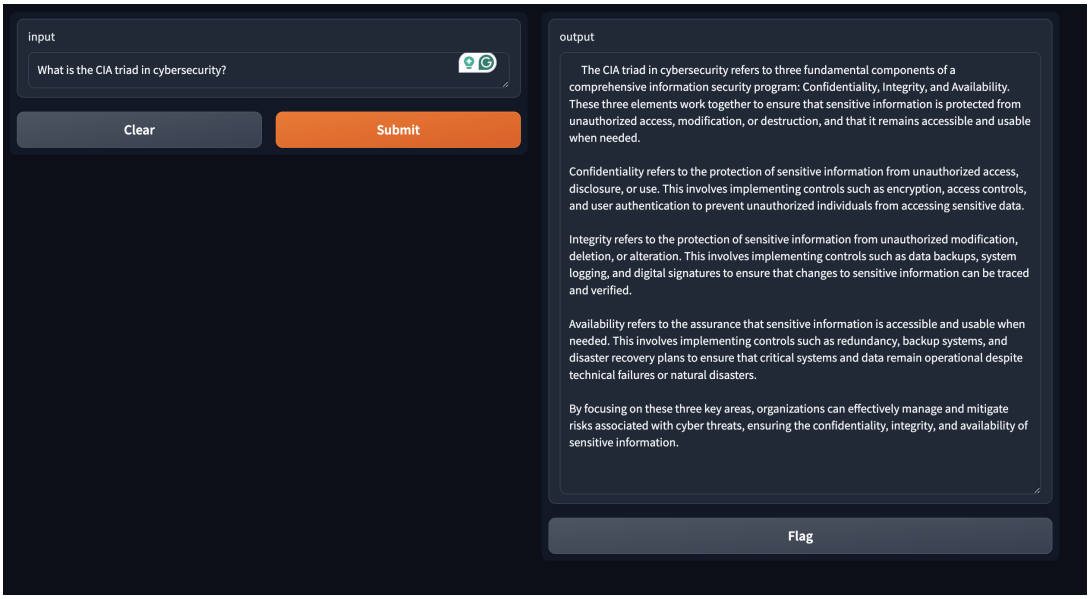



**Figure 4.3:** Example prompt with user interface

#### 4.3.3.2 Survey Design

While interviewing, experts were requested to fill out the survey after the interview. The survey was designed to get a comprehensive insight from the complete document. However, interviewees reviewed each criterion section by section in the interview. Insight into the usefulness, completeness, reliability, level of detail, and relevancy was tried to obtain.

Some short questions were asked during the survey instead of the interview to save time. Some questions were relevant years of expertise or a brief job description. Please refer to the appendix A.4 for the survey questions. Thanks to the survey, qualitative data was converted into qualitative data.

#### 4.3.3.3 LLM Demo Design

The purpose of the LLM demo was to observe the interviewees' interaction with the tool. We were trying to see their prompting style and if they obtained better results, and observe experts if they were ready to use such tools.

After each interview, interviewees were requested to spend 10-15 minutes trying the model. One task was given to the interviewees. At the same time, they were able to construct prompts when they desired. While they were entering prompts and reviewing the outputs, the prompts, outputs, and conversations were recorded to get better analysis material as quantitative data. Please refer to appendix A.3 for the demo guide.

Neither content nor thematic analysis was done to analyze the demo part because the demo parts took between 5-15 minutes and there was not enough transcript to analyze. However, we went through the transcript and highlighted the parts related to usability, and model improvement in this observational study. The reason was that the goal was to observe the participants as mentioned above.

### 4.3.4 Coding in Data Analysis

Content analysis is used to identify the existence of words or concepts within a text so that researchers can analyze and quantify the concepts and their meanings. Thus, they make inferences valuable for their research. There are two types of content analysis. Those are relational content analysis and conceptual content analysis. In relational concept analysis, one of the main goals was to analyze the relationship between codes. In conceptual analysis, researchers look for the existence or frequency of a concept. Those codes and concepts are words, phrases, or sentences [12].

To deeply analyze the interviews, reduce the complexity of the data, and observe the patterns, content analysis was conducted using transcripts and records from the interviews. Conceptual content analysis was selected because codes were independent even though some were close to each other. Another reason was that an attempt was made to see if participants mentioned specific terms that are occurrences of selected terms.

The first step of the conceptual content analysis was to decide the level of analysis. Phrases were better suited for this study because when interviewees explained their evaluations, they provided some examples or described the same concept in more than one sentence. The second step was developing a pre-defined or interactive set of categories or concepts. The decision was to stick with the pre-defined codes but allow some flexibility to insert missing codes. The third step was to decide whether to code for the existence or frequency of a concept. The frequency of a concept would be biased since some of the interviews might only talk about the negative sides however others might give only positive feedback. Because of this, the last decision was to go with the concept's existence. The fourth step was deciding on how to distinguish among the concepts. There were many diversity interviews. When there was a text part that contained semantically the same meaning, the same code was used for that part of the text. The fifth step was to determine coding rules to be consistent and coherent. One of the rules was not to miss anything, text was skimmed at least twice. When anyone came up with an extra coding idea, the information was shared and discussed. The sixth step was to determine what to do with irrelevant data which was ignored. The seventh step was coding the text. Nvivo software, which is used for qualitative analysis, was used to achieve that. The last step was analyzing the results [3].

Pre-defined codes were completeness, correctness, improvement, standard, level of detail, reliability, relevancy, usefulness for generation, usefulness for assisting, and usefulness for evaluation. Usefulness for generation, assisting, and evaluation mean the same. However, we separated them since we were trying to understand the usefulness of different features of LLM. During the coding process, three extra codes were included; bias, reasoning, and understandability. Also, to understand the intent and emotions of the participants, 4 sub-codes are created as positive, negative, neutral, and mixed for each code. The figure 4.4 below with a few examples demonstrates how the transcript was coded.

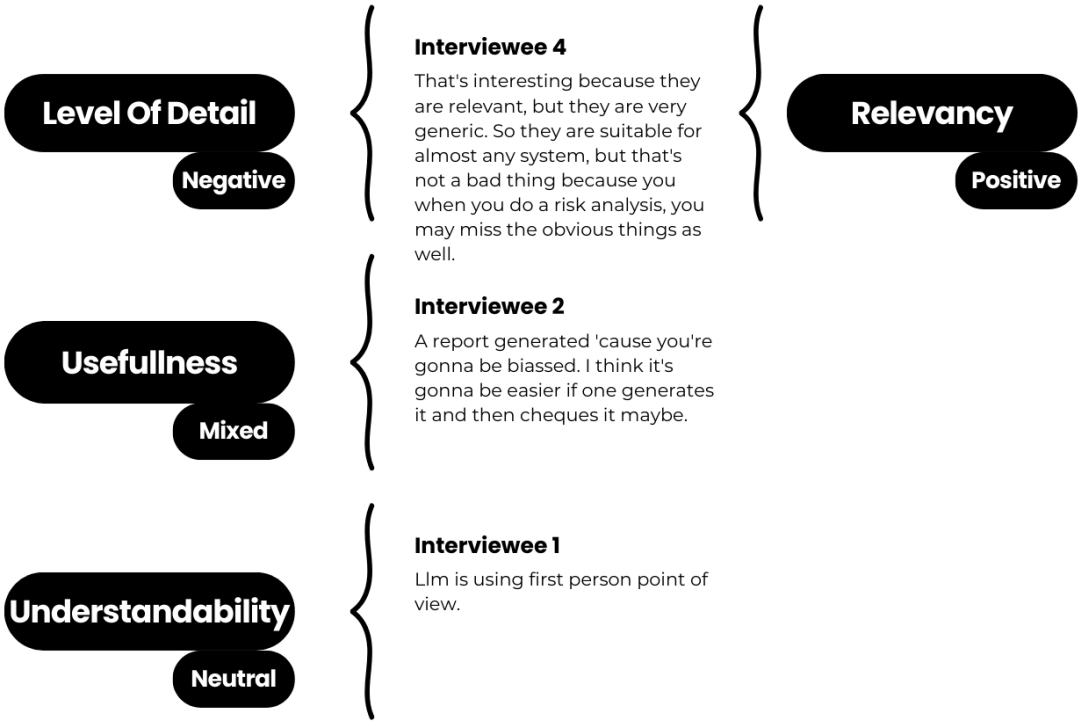



**Figure 4.4:** Coding Example

The example figure 4.4 above had 4 different codes for three different quotes from the interview cycle 2. For instance, the quote from interview 4 was coded as negative level of detail and positive relevancy. The part below is proof of the negative level of detail.

> "They are very generic so they are suitable for almost any system " - Interview 4

At the same time, the part below is evidence of positive relevancy.

> "That is interesting because they are relevant" - Interview 4

The next chapter will cover results from interviews, surveys, and demos.

# 5

# Artifact

This chapter explains the technical details of artifact implementation and improvement in both iterations of design science. First, we reasoned the model selection for the artifact, and then basic implementation and solutions for the first problems. After that, we explained the architecture of the artifact followed by a solution for the memory issue. Next, we explained how we improved the artifact from the first iteration to the second iteration and how we used the artifact to create a risk assessment document. The artifact was an LLM enhanced for the cybersecurity risk assessment process, utilizing the RAG technique.

## 5.1 Selection of Large Language Model

In the first cycle, we decided to focus on a reasonably small model that can run locally and on consumer GPUs to keep the data provided confidential. Considering the time constraints for this study, we determined to evaluate one LLM in depth using experts' opinions. The model had to answer questions since cybersecurity experts may want to interact with the chatbot during risk assessment. They might seek clarification during the process. Additionally, one of the main functions of the selected large language model was text generation. That is, generating completely new risk assessments for specific use cases, finding vulnerabilities or overlooked points, listing possible hazards or rare but possible threats, etc. This functionality was tightly connected to RQ1. The main goal of this functionality was to generate comprehensive risk assessments in the context of autonomous forestry machinery. Lastly, the selected large language model had to be conversational. Conversational functionality would also improve the efficiency of the risk assessment process in addition to instructional mechanisms.

Many criteria and metrics needed to be considered. Those criteria were **accuracy, hallucination, and truthfulness**. Parameter size was also significant because the performance of the large language models increases as parameter size grows despite exceptions [86]. However, to fulfill the goals of this study, one of the requirements was the most suitable LLM rather than the best chat model since some of the best models do not provide local access to the model. The possibility of hallucinations should be acknowledged, as they can affect the correctness of the output. For a risk assessment process, made-up risks were not desirable. Truthfulness was another criterion that was tightly connected to accurate results. Guiding cybersecurity experts with false claims might be costly.

Benchmark studies presented here were reviewed to select and evaluate the appropriate large language models.HELM had main versions for the large language models: one

was the classic [1] where they compared 142 models with 87 scenarios and 50 metrics, and the other one was the lite version [2] which was a simplified version of Helm Classic with 30 models and 10 core scenarios. Authors keep the benchmark updated continuously. Helm Classic showed the Llama 2 (70B) had the best accuracy. However, Llama 2 (70B) was the third-best open model after Yi (34B) and Mixtral (8x7B 32K seqlen) in Helm Lite.

For the hallucination criterion, There was a benchmark study compared the models based on hallucination in order; InstructGPT-davinci-002, InstructGPT-davinci-003, ChatGPT, Claude-2, Claude, LLama-7B-Chat, ChatGLM, Falcon, Vicuna, Alpaca [45]. Another benchmark study in the hugging face platform suggested that Llama 2 (70B) was the best open model followed by Llama 2 (7B) and Llama 2 (13B) [8].

The truthfulness criterion measured how large language models make false statements in specific contexts [48]. Truthfulness was under question answering task and discussed under the TruthfullQA scenario in Helm Classic and Hugging Face platform. Llama 2 (70B) was the best open model for this criterion in Helm Classic [47].

There was no large language model that perfectly met all criteria. For instance, GPT3 had a 175 billion parameter size however hallucination was a big problem for this model, unlike smaller models like T5 with an 11 billion parameter size or Dolly with a 12 billion parameter size [67]. Also to the best of our knowledge, no benchmark study consisted of all the criteria and metrics selected for this study.

To make a final selection, closed-source large language models must be eliminated first. The reason behind that model queries could be executed through APIs. The data to be trained was sensitive, it was not worthwhile incurring risks by sending data to third-party firms. [81]. Unfortunately, we could not use large models such as 70B parameter size models due to a lack of computing resources. Considering factors such as accuracy, truthfulness, and hallucination, Llama 2 (7B) is one of the best open models according to some benchmark studies [47], [46]. Considering these, Llama 2 (7B) was the most suitable model for this study.

## 5.2 Basic Implementation and Solution

As mentioned above, Llama 2 with 7B was used for the first implementation. The main focus was prompting and getting one simple risk assessment to move forward and get initial feedback from the experts. At the same time, we learned how to program large language models. Even though extra time was spent on learning, valuable lessons were figured out such as quantization, few-shots, and getting the model run without any memory issues.

We identified a requirement to encapsulate the implementation of the large language model in a docker which is a common practice for deploying open-source LLM effectively. Control of the dependency management and program maintenance were easier thanks to the docker and local model.

[1] HELM Classic: https://crfm.stanford.edu/helm/v0.4.0/
[2] HELM Lite: https://crfm.stanford.edu/helm/lite/v1.0.0/

To provide ease of use, a docker file was created with Ubuntu as the main OS for the container application. Required packages and libraries were installed such as CUDA drivers for GPU computation, and llamacpp for inference.

At first, the Llama 70B model was tried, however memory issues occurred. Thus, the Llama 7B model was decided to be used instead. Another way to handle memory issues was quantization. 4-bit quantization would be very good for the first implementation since one-fourth of the model weights were utilized. However, the performance loss was apparent. With 4-bit quantization, approximately 4GB of GPU memory was used which was affordable based on the hardware requirements. It was possible to use Llama 70B with 2*24 GB of GPU memory [3] but the hardware that was provided did not meet the requirements specified. Another reason was that the token limit for prompting and output would be limited and it was not desired for this case since experts might want to provide more context in the prompt.

## 5.3 RAG architecture

Before the implementation of the LLM, a technique called "Retrieval Augmented Generation" was determined to be used. The main advantage of this technique was that a large language model could fetch information from the data provided. In the first phase, just 2 documents were included in the vector database while waiting for the rest of the data. Vector databases keep the text, image, video, or audio as a mathematical representation. Embedding functions transform vectors to raw data. Unlike other databases, retrieval happens through a similarity search instead of an exact match. Similarity search looks for semantically similar data. [62], [56], [29]. There were many vector database providers and ChromaDb [4] was selected because of its ease of use and comprehensive community support. Additionally, HuggingFaceEmbeddings [5] was used to embed the semantic meaning of a text into the vector database. Moreover, the RAG implementation can be categorized naive RAG application based on this paper [26]. The naive RAG applications have three parts: indexing which is vector database, retrieval, and generation.

AutoTokenizer [6], AutoModelForCausalLM [7], and transformers pipeline [8] were used which are suitable with Llama 2 architecture. Also for the RAG implementation, RetrievalQA was employed to do question-answering.

[3] https://www.answer.ai/posts/2024-03-06-fsdp-qlora.html

[4] https://github.com/chroma-core/chroma

[5] https://api.python.langchain.com/en/latest/embeddings/langchain_huggingface.embeddings.huggingface.HuggingFaceEmbeddings.html

[6] https://huggingface.co/docs/transformers/model_doc/auto#transformers.AutoTokenizerLink

[7] https://huggingface.co/docs/transformers/model_doc/auto#transformers.AutoModelForCausalLMLink

[8] https://huggingface.co/docs/transformers/v4.44.0/en/main_classes/pipelines#transformers.pipelineLink

## 5.4 Memory Handling

4-bit quantization was implemented for the first implementation but 8-bit quantization was tried in the second cycle because of the need for the performance. However, memory issues were still encountered after a couple of prompts. The reason was each time a prompt ran, memory consumption increased cumulatively. As a solution, the cache was cleaned every time a prompt was run. After that, the average memory consumption was around 6 GB of VRAM with the 8-bit quantization and the 7B model. Depending on the prompt length, memory consumption increased proportionally.

## 5.5 Improvement on RAG Implementation

Based on the feedback from the interviews in the first iteration, we improved the RAG implementation. After getting the data to feed the model, data cleaning methods such as removing unrelated pages and non-English content were implemented. A program was written for data cleaning to remove HTML tags, stop words, emojis, emoticons, special characters, and punctuation. Another thing in data cleaning was normalizing the text using the same text character encoding standard. Also, this normalization included lowering all the letters for data in the vector database and each prompt entered because users can enter different capitalizations of the same word such as AGRARSENSE, AgrarSense, or agrarsense. The data included some other standards related to forestry, some definitions in AGRARSENSE, IEC 62443 documents, risk assessment documents related to the shipping and railway domain, and some EU regulations.

Also, MITRE attack techniques from the official website[9] were included. The data was obtained from this website. The database is relatively small since AGRARSENSE is an ongoing project and a total of 45 pdfs were inserted in the vector database.

The vector database was already constructed in the first cycle. The code for the data orchestration was updated for more data. The responses from the mixed vector database were not as we desired when the first prompts were attempted. When something related to MITRE attack techniques was prompted, the LLM did not recognize what was the MITRE attack technique correctly. Consequently, the database was separated into two: one with only MITRE attack techniques data and one with other data. After that, according to our trials, the model was performing better.

One of the reasons for selecting the Llama 7B model was to provide enough context length for input and output so that the model would have enough computing resources. To do that, the MAX_NEW_TOKENS, which configures input and output lengths, the parameter was increased to 4096. Although as little creativity as possible was needed, a little randomness was required to generate new ideas. the temperature which determines the randomness of the output, was decreased to 0.2 to achieve that.
If someone needs to use the tool for another domain, including data from their domain would be enough. However, temperature and MAX_NEW_TOKENS parameters can be arranged based on the requirements.

[9] `https://attack.mitre.org/techniques/enterprise/`

## 5.6 Creation of Risk Assessment & Evaluation Document

The risk assessment and evaluation document was the output of the artifact for this study which is available in this link. This document was created to present during the interviews and get evaluation feedback from the experts. Based on the feedback obtained in the first cycle of interviews IEC 62443 was selected. To get valuable feedback for the second cycle interviews, standard IEC 62443 steps were followed. This document included some of the steps in IEC 62443. All the content for the generation risk assessment part was related to the AGRARSENSE project. The document included primary and secondary assets, threat actors, threat landscape, initial risk assessment table, threat initiation likelihood, and two attack tree examples. Apart from the generation of risk assessment, there was one example of existing risk assessment evaluation in the railway and shipping domains.

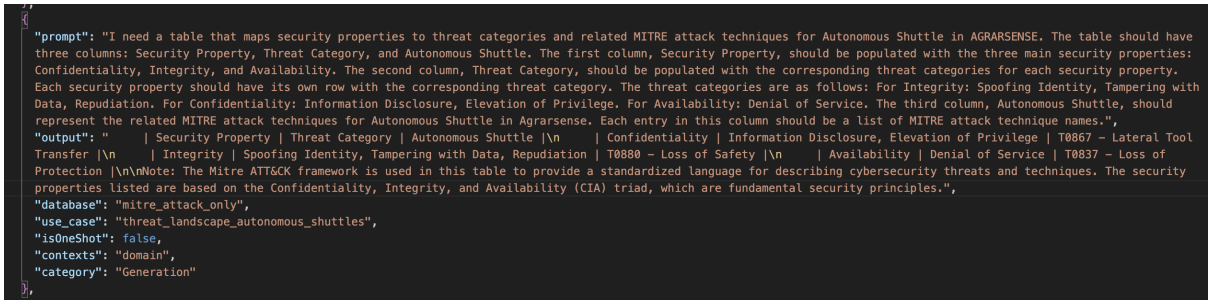
Prompts were created based on the desired output and discussions with the supervisors. We updated the prompts iteratively to get the best result possible. Also, the written prompts were reviewed by supervisors and updated in one of the weekly meetings. We tried to be as clear as possible for each instruction. Each prompt targeted a specific part of the created document. A prompt template was used based on the Llama 2 documentation to define the role of the tool and set the tone for better behavioral traits. The code snippet below was used as a prompt template in our model.

```
template = """<s>[INST] «SYS» You are an assistant for
question-answering tasks.  You are helpful and friendly.  Use the
following pieces of retrieved-context to answer the query.  If you
don't know the answer, you just say I don't know.  Keep the answer
concise.

«SYS»

context
question [/INST] """
```

The model generated the content but we wrote the prompts. The same prompts were run multiple times to get the best result among different outputs. All the distinct outputs were saved into a JSON file and the best outputs were selected based on our knowledge. The selected outputs are merged into a document. However, the document was organized manually by forming tables and giving section headings. The figure 5.1 below is an example prompt-output pair that was selected output for the document.



```
},
{
  "prompt": "I need a table that maps security properties to threat categories and related MITRE attack techniques for Autonomous Shuttle in AGRARSENSE. The table should have
  three columns: Security Property, Threat Category, and Autonomous Shuttle. The first column, Security Property, should be populated with the three main security properties:
  Confidentiality, Integrity, and Availability. The second column, Threat Category, should be populated with the corresponding threat categories for each security property.
  Each security property should have its own row with the corresponding threat category. The threat categories are as follows: For Integrity: Spoofing Identity, Tampering with
  Data, Repudiation. For Confidentiality: Information Disclosure, Elevation of Privilege. For Availability: Denial of Service. The third column, Autonomous Shuttle, should
  represent the related MITRE attack techniques for Autonomous Shuttle in Agrarsense. Each entry in this column should be a list of MITRE attack technique names.",
  "output": "     | Security Property | Threat Category | Autonomous Shuttle |\n     | Confidentiality | Information Disclosure, Elevation of Privilege | T0867 - Lateral Tool
  Transfer |\n     | Integrity | Spoofing Identity, Tampering with Data, Repudiation | T0880 - Loss of Safety |\n     | Availability | Denial of Service | T0837 - Loss of
  Protection |\n\nNote: The Mitre ATT&CK framework is used in this table to provide a standardized language for describing cybersecurity threats and techniques. The security
  properties listed are based on the Confidentiality, Integrity, and Availability (CIA) triad, which are fundamental security principles.",
  "database": "mitre_attack_only",
  "use_case": "threat_landscape_autonomous_shuttles",
  "isOneShot": false,
  "contexts": "domain",
  "category": "Generation"
},
```


**Figure 5.1:** Example Prompt Output Pair

As presented in the figure 5.1 above, the same prompt was run several times. Based on our knowledge, we selected the output for this part. To distinct the prompts, databases, techniques, and domains we used different JSON keys. It is worth noting that none of the selected outputs were implemented using few-shot learning techniques.

# 6
# Results

## 6.1 Results from First Iteration

### 6.1.1 Interview Results

The primary goal of the interviews conducted in the first iteration was to find areas where the LLM could be improved. We grouped our interview results under three categories as presented in table 6.2. The three categories were *improvements, use cases, and challenges* as these made up the most significant feedback provided by the experts. Demographic information about the participants from the first iteration can be found in table 6.1 which shows the interview participant's role and years of experience.

**Table 6.1:** Interview participants

| *Participant* | *Role* | *Experience* | *Interview cycle* |
|---|---|---|---|
| Interview 1 | Cybersecurity Researcher | 2 Years | 1 |
| Interview 2 | Safety expert | 18 Years | 1 |
| Interview 3 | Cybersecurity expert | 25 Years | 1 |

**Table 6.2:** Categories of feedback from the first iteration of interviews

| *Categories* | *Subcategories* |
|---|---|
| *Improvements* | **Standards and Structure:** Follow functional safety and cybersecurity standards like ISO 12100 [34], IEC 62443, and ISO 27001 [33].<br>**Comprehensiveness:** Detailed explanations of reasoning processes should be provided.<br>**Information:** The information provided should be more specific to the domain. |
| *Use cases* | **Risk Identification:** Assist in identifying hazards through, for example, brainstorming; LLMs can speed up the risk assessment process and act as a redundancy check.<br>**Dynamic Interaction:** Use conversational LLMs to dynamically interact and improve the risk assessment process and check for completeness. |
| *Challenges* | **Completeness and Reliability:** Concerns about LLM's ability to fully capture all possible risks without human oversight; Verification by humans is crucial.<br>**Specificity and Training:** The model needs more specific training data and detailed knowledge about the domain. |

#### 6.1.1.1 Improvements

By interviewing experts working with risk assessments we identified several key aspects to improve the LLM artifact in the second iteration of the study. Firstly, all experts stressed the importance of following predefined standards when conducting risk assessment. The goal of following a standard is to ensure consistency, reliability, and accuracy regardless of who conducts the risk assessment. This was especially important as our goal was for the LLM-generated outcome to be as close to human as possible.

> "Following a standard would make the output more structured and closer to human-made risk assessments" - Interview 1

Secondly, all participants agreed that the output was too general in nature. A fine-tuned model with specific domain knowledge and more in-depth information would likely perform all tasks examined better. Lastly, two experts expressed that more details in the reasoning and motivation of the LLM would make the results more comprehensive.

> "The risks are accurate but general. A better output would include risks associated with the specific machine." - Interview 2

#### 6.1.1.2 Use cases

Another goal of the first iteration was to explore how practitioners want to interact and incorporate LLMs into their established workflow of creating risk assessments. All of the interviewees agreed that risk identification can be a time-consuming and laborious process. Therefore, using our LLM model can be beneficial as a brainstorming partner. This can both speed up the process of creating a risk assessment and act as a redundancy check to make sure the assessment covers easily missed risks or scenarios.

> "I would use the LLM to brainstorm risks early in the risk assessment process to then build upon it myself" - Interview 2

Furthermore, while some experts expressed their preference of using the LLM early in the risk assessment creation process, they saw promising potential in interacting with an LLM at several other points of the process. For example, evaluating and improving a human-made risk assessment or identifying potentially dangerous scenarios.

### 6.1.2 Challenges

We also asked the experts about potential challenges that may hinder the use of LLMs when working with cybersecurity risk assessments. The most common feedback was that reliability and trust limited the practical applications of LLMs, especially in safety-critical areas. LLMs nowadays do not provide complete reliability, therefore when it comes to liability, humans remain accountable when using the outcome of LLMs [81] [53]. The interviewees also reiterated that AI in general and our LLM model in particular would be more likely to be used with more specific knowledge and information for each domain, highlighting the importance of domain-specific training data.

> "To me trust is the biggest issue right now, I cant rely on the information given by the AI to be correct" - Interview 3

The results from the first iteration of interviews formed the basis for the further development and improvement efforts in the second iteration.

## 6.2 Results from Second Iteration

### 6.2.1 Interview Results

The primary objective of the second round of interviews was to gather qualitative and quantitative data to answer each research question. Demographic information about each participant such as years of experience and area of expertise can be found in the below in table 6.3. The interview participants evaluated the quality of the output and usefulness of the customized LLM. The interviews focused on the experts' perceptions of key aspects of the LLMs output such as correctness, completeness, and relevance of information. A total number of 12 interviews were conducted, the interviews transcribed, and every statement was coded as presented in Figure 6.1 using content analysis.

**Table 6.3:** Interviewees Data

| Interviewee | Company | Expertise | Years of experience | Length | Date | Cycle |
|---|---|---|---|---|---|---|
| Interviewee 1 | RISE | Researcher | 4 | 66 min | 2024/06/07 | 2 |
| Interviewee 2 | RISE | Cybersecurity | 11 | 71 min | 2024/06/17 | 2 |
| Interviewee 3 | RISE | Researcher, safety & cybersecurity | 2 | 79 min | 2024/06/19 | 2 |
| Interviewee 4 | RISE | Safety & Cybersecurity | 30 | 59 min | 2024/06/20 | 2 |
| Interviewee 5 | Komatsu | Project Manager | 23 | 60 min | 2024/06/20 | 2 |
| Interviewee 6 | RISE | Researcher | 18 | 39 min | 2024/06/24 | 2 |
| Interviewee 7 | RISE | Researcher | 13 | 51 min | 2024/06/24 | 2 |
| Interviewee 8 | RISE | Safety | 30 | 67 min | 2024/06/27 | 2 |
| Interviewee 9 | RISE | Safety | 20 | 49 min | 2024/06/27 | 2 |
| Interviewee 10 | RISE | Researcher | 20 | 48 min | 2024/06/28 | 2 |
| Interviewee 11 | RISE | Researcher | 15 | 51 min | 2024/07/01 | 2 |
| Interviewee 12 | RISE | Research Leader | 9 | 55 min | 2024/08/09 | 2 |

The results from each code are presented in this section. The numbers in this section signify in how many interviews a certain opinion was expressed not how many times. In some cases, the sum of coded statements is greater than the total number of interviews, meaning that the same interviewee expressed, for example, both positive and negative opinions. It is important to note that when interviewees mention something is complete or correct, they refer to specific parts.

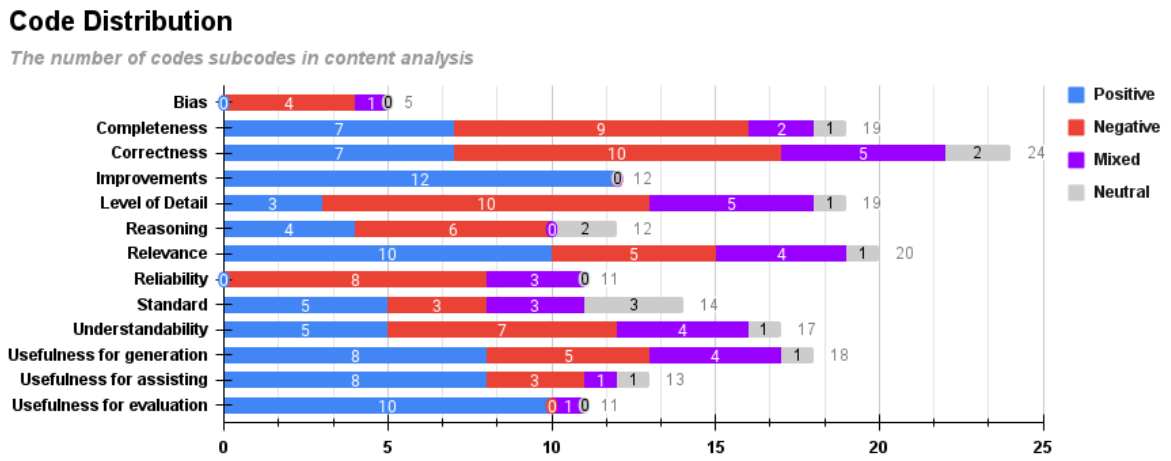



**Figure 6.1:** Code Distribution

#### 6.2.1.1 Completeness

The completeness code referred to the extent to which output contained all necessary information. In the specific case of cybersecurity risk assessment, it was about covering all areas that experts would expect to find in human-created risk assessments. The interviews showed a mixed perception regarding the completeness of the LLM's output.

In some sections, experts noted that information was missing that they would have expected to be present in a risk assessment created by practitioners.

**Negative**

> "This part is not complete because there would be a base station and there would also be communication systems that would include 5G network and communication" - interview 1

> "The definition of risk as we are using it in our work, says that risk is a function of impact and probability or likelihood. And that those aspects are not covered here." - Interview 4

Participants also expressed positive opinions about the general completeness of the LLM output.

**Positive**

> "Almost everything is covered here. This is very well documented. These detectors as well and taught in many courses, I have seen these in couple of courses at Chalmers so nothing out of ordinary, so it was complete, I would say." -Interview 7

#### 6.2.1.2 Correctness

Correctness assessed the accuracy, factual correctness and logic of the information provided by the LLM. Correctness was another code where the perceptions of the experts were mixed. There were also some disparities between the different tasks performed where the LLM generally provided correct information when evaluating preexisting risk assessments. Correctness is crucial code as inaccurate could potentially lead to dangerous

decisions.

The negative points related to incorrect information which can be critically important in risk assessment.
**Negative**

> ”Amplification attack can lead to network exhaustion; that is not a problem. But network exhaustion leading to unforeseen access is not logical.” - Interview 8

Other sections were deemed as having very accurate information.
**Positive**

> ”Examples of possible attacks on the network and the system are correct. I could see that it was very interesting for me, the left hand side as well because this is my area of expertise that networks can be can be compromised. So that was quite good here.” - Interview 7

#### 6.2.1.3 Level of detail

The level of detail evaluated how specific and precise the LLMs outputs were, particularly in information specific to the domain and use cases. The feedback related to the level of detail of the information in the outputs was primarily negative, often referring to information not being specific to the domain and specific use case.
**Negative**

> ”So the structure is somewhat accurate to the standard, but the actual detail of the content is not.” - interview 10

> ”This is a generic attack vector. Doesn’t go into details, it’s not adapted to any specific drone type, for example, so this applies for all drones.” - Interview 1

#### 6.2.1.4 Relevance

The relevance code aimed to examine how well the output and information were shaped to the specific case of cybersecurity risk assessment for autonomous forestry machines. The impression of the interview participants was that the information provided by the LLM was always relevant to the domain and the tasks it was asked to perform.
**Positive**

> ”Yes, the identified threat actors are relevant.” - Interview 2

#### 6.2.1.5 Reliability

Reliability captured how trustworthy and dependable the experts perceived the output. It became apparent from the interviews that trust and perceived reliability of the LLM were weak among most experts.

One area where trust was especially low was in the LLMs’ ability to compute mathematical calculations.
**Negative**

> ”The one thing that would worry me a bit is the also. I wouldn’t trust LLMs to do math for me at this point.” - Interview 3

Another area that affected reliability negatively was a lack of motivation for facts or missing information.

**Negative**

> ”Which again, without having proper motivation for those. We don’t know whether we can trust the data.” - Interview 2

#### 6.2.1.6 Standards

Standards evaluated how well the LLM follows predefined risk assessment standards, in this case IEC 62443, and whether the outputs are structured according to the standard. The experts judged the ability of the LLM to follow a predefined standard very differently. Some focused on the overall structure and were impressed, while others focused on the details of specific sections and deemed it to not follow the standard closely enough.

**Negative**

> ”Secondary assets are the components that enable primary asset functionality. So this primary asset classification doesn’t follow 62443.” - interview 1

Others felt like the structure followed the main points.

**Positive**

> ”I felt that the structure was quite good. It has the main ingredients and that actually intrigued me” - Interview 8

**Mixed**

> ”Is similar to a structure of a risk assessments I’ve been doing in the past. But it’s not exactly the same” - Interview 5

#### 6.2.1.7 Bias

Bias assessed the extent to which the LLMs output exhibited any biases. None of the experts noted any of the usual suspects like gender or racial bias commonly found in AI-generated material [28]. However, some training data biases were present, closely resembling overfitting. As the LLM used training data from existing cybersecurity risk assessments from the railway and train domain, some residual information also showed up in the output. One interview suggested:

**Negative**

> "I see railway at some part where I expect it should be AGRARSENSE.” - Interview 11

#### 6.2.1.8 Improvements

The improvements code represented statements focused on further enhancements to improve the generated output and new use cases of the LLM. Each expert was asked what

could be improved and if they could come up with additional use cases. The experts were and saw the potential for future improvements in several areas and additional uses such as in regulations, requirements, education, and new areas like machinery and safety. One participant wanted to try the LLM in new areas and said:
**Positive**

> ” I work with machinery, machinery, safety mostly. So I think this is very interesting and it would be interesting to apply it to say, machinery directive, ISO 12100." - Interview 8

We grouped the improvements into three parts: The first part includes improvement ideas for the content of the generated risk assessment document. Those can be achieved with prompting. Explanation of each property, reasoning for each part, and combination of risk in different perspectives are some of the examples of this category

> “I think the reasoning to me should be for all of the parts.” - Interview 2

Some of the interviewees requested new features on the LLM's capability such as having chat history for better prompting, data diagram or data flow as an output, and forming the outputs as a template are some of the examples.

> “Adding easily domain-specific files” - Interview 9

Some interviewees suggested some improvements in the content of the RAG implementation. They requested to add more data to RAG implementation in different ways including training the model on relevant standards, feeding the model with safety standards and regulations, and adapting threat mitigation strategies.

> “Feeding the RAG with machinery directives, safety standards, and safety regulations so that it can work safely. Safety is important since both safety and cybersecurity affect each other.” - Interview 8

> “Mitigation strategies should be included” - Interview 3

#### 6.2.1.9 Reasoning

Reasoning evaluated the LLM's ability to draw conclusions from information and explain its decision-making logically. The opinions on the ability of the LLM to reason were generally negative, with some positive highlights. The negative impressions were often connected to two well-known shortcomings of current-day LLMs: mathematics and complex reasoning [70] [30]. Firstly the capability of the LLM to do calculations and math in general was very limited, even basic calculations that may seem rudimentary proved to be challenging to the model. One expert stated:
**Negative**

> ”So I cannot understand the justification for these numbers. It doesn't explain what it means. Doesn't give an explanation for why.” - Interview 1

Secondly, the experts perceived the LLMs' ability to justify decisions and conclusions as insufficient in the risk assessment's more intricate and complex sections. Explaining logical consequences and inferences in a new context is a well-known challenge for LLMs, and our case was no different [30].

Other experts found the reasoning helpful and were impressed by the ability to reason about for example threat actors. One expert said:
**Positive**

> ”It’s even providing this reasoning, so this is pretty good and pretty helpful” - Interview 12

#### 6.2.1.10 Understandability

Understandability investigated how easily the experts could understand the LLM’s output, considering the clarity of language, structure, and presentation of information. The perception of the understandability of the model was relatively mixed. The negative opinions referred to the language style used in the generated cybersecurity risk assessment. One expert said:
**Negative**

> ”That it wasn’t tuned to sound a scientific assessor, felt to me like the LLM is a layman.” - Interview 1

Others expressed some difficulty understanding the contents of the tables created by the LLM, which negatively impacted the overall understandability. One expert expressed:
**Negative**

> ”I had the most difficulty understanding this table for somebody who performs risk assessment. So I got frustrated not understanding all of these,” - Interview 2

Some experts also expressed positive opinions about the clarity and structure of certain sections and stated that it was easy to understand and close to what a human would have done. One participant stated:
**Positive**

> ”I can understand the security properties and so on availability and integrity and that looks nice to me” - Interview 9

#### 6.2.1.11 Usefulness

Lastly, the usefulness was coded according to each research question. Usefulness measures the practical usefulness of the LLM according to our three research areas, Generation, Assistance, and Evaluation. The response to the usefulness of LLMs in the context of risk assessment in forestry was primarily positive across all three scenarios.

**Usefulness for Generation** Usefulness in relation to generation was related to the first research question (RQ1) where we tried to investigate to what extent the LLM could autonomously generate cybersecuirty risk assessments. This code also entailed specific tasks during the risk assessment process where generation was perceived as negative or positive. One expert noted how the LLM could be useful when autonomously generating the initial structure of the risk assessment:
**Positive**

> "The final or the full sort of risk assessment cycle, I think you can speed that up quite a bit by having a large language model will help you sort of prepare the structure and to." - Interview 5

> "The generated risk assessment could be used as a baseline at the beginning or as a proof-check at the end of the risk assessment process." - Interview 12

**Usefulness for Assistance** Usefulness in relation to assistance encompassed tasks and scenarios where the human practitioner and LLM would collaborate to complete a task. This code was connected to the second research question (RQ2) and the perception by the experts where primarily very positive. One expert saw great potential in assisting when identifying threat actors:
**Positive**

> "For example think about some new threat actors that perhaps an assessor might not have thought about." - Interview 1

> "I see LLM usage when they can help an assessor to save time in doing some time-consuming work." - Interview 1

> "So you can view this as a partner to get tips, suggestions, and assistance" - Interview 4

**Usefulness for Evaluation** The final usefulness code captures experts' opinions about the evaluation of already existing sections or full risk assessment documents. The experts' opinions were overwhelmingly positive with 10 positive and 0 negative. One expert expressed interest in using it as a reviewer to find missing information:
**Positive**

> "This will be very helpful to get something more. If you have a risk assessment, then the LLM can go over it and find flaws or examples, and so on that, you didn't think about. It would be very useful." - Interview 9

> "I think what it gave here was quite good. A good example and it gives good guidance on what to think about" - Interview 4

### 6.2.2 Demo Results

Demo Results are divided into three parts. In the statistics subsection, some statistical information is presented. In the inferences subsection, improvement suggestions and usability feedback are shared.

#### 6.2.2.1 Statistics

The table includes columns with an order of the number of prompts, the number of updated prompts, the number of characters for all prompts for one interviewee, and the average prompt length
12 interviewees queried 65 prompts in total. 23 of those prompts are updated. Updating the prompts means that the meaning of the prompt is the same but has a slight change

**Table 6.4:** Modified Demo Data

| Interviewee | #prompts | # updated prompts | # of char for all prompts | avrg prompt len |
|---|---|---|---|---|
| Interviewee 1 | 3 | 2 | 880 | 293.3 |
| Interviewee 2 | 11 | 4 | 1552 | 141.09 |
| Interviewee 3 | 5 | 0 | 611 | 122.2 |
| Interviewee 4 | 3 | 0 | 519 | 173 |
| Interviewee 5 | 9 | 2 | 525 | 58.3 |
| Interviewee 6 | 5 | 3 | 479 | 95.8 |
| Interviewee 7 | 4 | 1 | 873 | 218.25 |
| Interviewee 8 | 5 | 3 | 487 | 97.4 |
| Interviewee 9 | 2 | 0 | 156 | 78 |
| Interviewee 10 | 3 | 1 | 151 | 50.3 |
| Interviewee 11 | 5 | 3 | 507 | 101.4 |
| Interviewee 12 | 10 | 4 | 934 | 93.4 |
| Total | 65 | 23 | 7674 | 1522.44 |
| Average | 5.42 | 1.92 | 639.5 | 126.87 |
| Max | 11 | 4 | 1552 | 293.3 |
| Min | 2 | 0 | 151 | 50.3 |

in the language. The majority of the prompts were related to cybersecurity. The total number of characters prompted is 7674.

The average number of prompts is 5.42 with a 1.92 update rate. The average prompt length is 639.5 chars for all prompts.

Interviewee 2 entered 11 cybersecurity-related prompts, which were the most during the demos. Also, the same participant and interviewee 12 updated 4 prompts which is the largest number. Average prompt length reached its peak with interviewee 1.

Interviewee 9 just tried 2 prompts. 3 interviewees did not update the prompts. Even though Interviewee 10 entered 3 prompts, the total prompt length was lower than Interviewee 9. This makes interviewee 10 have the lowest average for prompt length with 50.3 characters.

#### 6.2.2.2 Improvement suggestions

Some interviewees proposed extra improvement ideas for the model and the tool itself. One interviewee suggested including chat history to ask for extra prompts related to previous outputs and prompts. The same interviewee also questioned if he could ask questions about mitigation strategies however data available did not include any cyber threats mitigation strategy. We are considering this question as a suggestion for improvement.

Another interviewee suggested writing a program or function to which a user can easily add additional domain-specific documents or relative information when the model does not understand the task or its context.

Additionally, interviewee 9 wanted to use the tool to search for specific requirements in the document.

> "It will be really helpful with an AI just telling me in this document in this row, for example, you get the answer to some requirement." - Interview 9

#### 6.2.2.3 Usability

One of the goals of doing demos on top of interviews was to get insight into the usability of our tool. Interviewees asked many questions regarding how to use the prompt such as "What does the flag button do?", "Does it save the output somewhere?", "Can I write multiple sentences?", and "Should I press clear here?" One of the participants asked "What does input/output do" to understand how to use the interface. To visualize what they mean, please refer to figure 4.3 to look at the tool interface.

Only interviewee 6 complained about slow inference. We observed that 8-bit quantization inferences took roughly 7 to 20 seconds but 4-bit inferences were faster.

> "It's slower than ChatGPT. They must have a lot of hardware." - Interview 6

One of the interviewees entered multiple prompts and observed that

> "So probably if you match the information that is available in the vector database. Then it becomes much better, right?" - Interview 12

It is possible to infer that usability might increase if the amount of documents increases.

Another usability metric inferred the time requirement to write prompts. Interviewee one was the sole person mentioning this need.

> "I think I would need to massage the input more based on the output I get because there seem to be so many things that need to go into one sentence[prompt]." - Interview 1

On the other hand, when more time was spent on prompting, the results were likely to get better quoted by interviewee 3.

> "Just I don't know if you want me to play around more but I think we're slowly getting some better." - Interview 3

Interviewee 6 also suggested some guidance tips about how to use the tool effectively. For instance,

> "It's the same when you play like ChatGPT or Bing, you need to learn how to ask the questions." Interview 6

The interviewee meant prompt engineering methods like learning how to ask questions. The same interviewee also suggested giving tips on best practices for the tool to get good

answers. The same person said,

> "I hope you include a bit about how people use the tool as well. " - Interview 6

Interviewee 12 found that the model might confuse the STRIDE model and the verb stride. This is the conclusion of normalizing all the letters in data and prompt.

In conclusion, demo participants came up with valuable improvement suggestions for further implementation. We observed that they require information more on how to use the tool and to get good results. Participants proved that the results would be better when the prompts and vector database matched.

### 6.2.3 Survey Results

All 12 interviewees filled out the survey and answered all of the 6 questions. Questions were closed-ended and the aim was to get a general overview of the artifact. The difference between the survey and interviews is that the survey gets insight into the general perspective of our model and the generated risk assessment document. However, interviewees answered their questions based on specific parts of the generated content in interviews. After that, all the results from content analysis, survey analysis, and demo analysis were compared and merged to get better insight into answering research questions and future work.

Figure 6.2 below illustrates the questions and responses. Answers collected were on a Likert scale with 7 different options and 6 of them are listed in the legend. The last option was "I don't know" which is not included in the list since it does not provide valuable information on the x-axis. The Y-axis consists of questions in reverse order. Percentages in the x-axis illustrate the degree to which people agree with the questions. For instance, if it is close to 100% on the left side of the x-axis, most respondents disagree with the respective question.

The general pattern shows that participants do not find the LLM reliable, generated risk assessment documents detailed enough and complete. However, they indicate that they would use such tools in their work. They also point out that the content in the generated risk assessment document is relevant. Even though the level of detail seems to be negative, the majority of the people slightly disagree on that question. The $2^{nd}$, $3^{rd}$, $4^{th}$ questions have one "I don't know" option selected which corresponds to 8.33%.

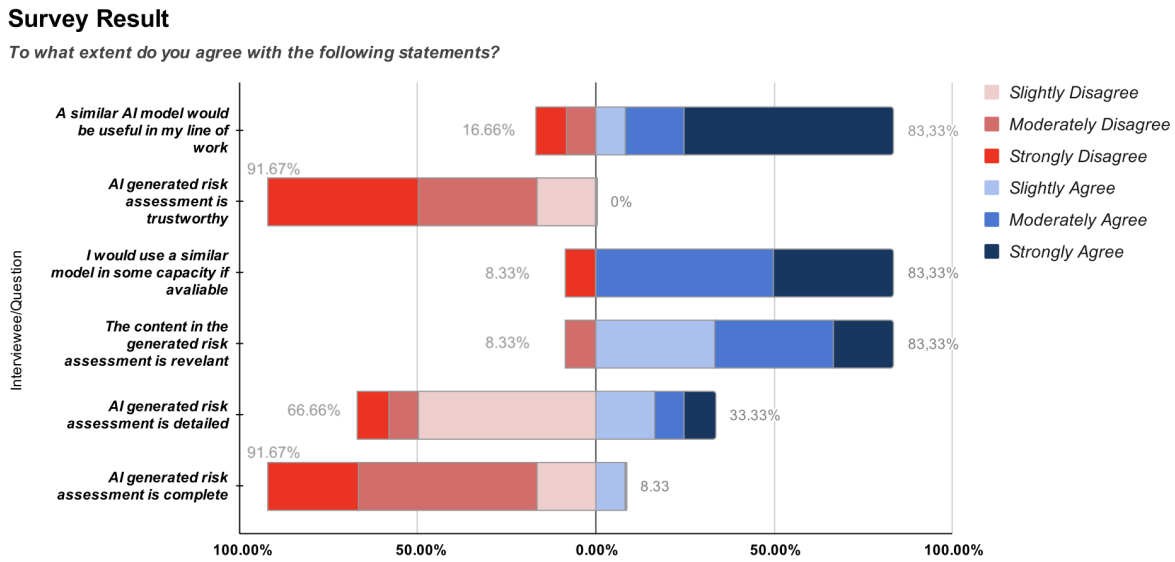



**Figure 6.2:** Survey Results

# 7
# Discussion

We discuss the consolidation of the research space regarding the findings of this study. We are also going to explore if the findings align with existing literature.

## 7.1 Research Question 1: Generation of Cybersecurity Risk Assessment

The first research question investigated the possibility of generating a risk assessment from scratch. We saw that our participants would not directly use the generated risk assessment without going through it and modifying it, primarily due to the poor levels of trust. Instead, the experts showed interest in using the generated risk assessment as a baseline or initial structure to iterate upon when creating their own 6.2.1.11. This aligns with previous work by Khojah et al. [39]. The study found that software engineers use LLM-generated artifacts as a foundation and source of inspiration rather than applying them directly

In both the interviews and the survey the main concerns expressed by the participants about the generated risk assessment were the reliability and level of detail of the output. However, even though the experts did not entirely trust the LLM model, they expressed a willingness to use the tool in the survey, and eight interviewees perceived the usefulness of the LLM generation as positive. Secondly, the experts expressed negative opinions related to the completeness, correctness, level of detail, and reasoning. These opinions coupled with observed hallucinations also contributed to their level of trust. We observed that these aspects were correlated to the quality of the output in the generation of risk assessments. One comment in 6.2.1.5 mentioned the connection between reasoning and reliability.

Poor data quality might lead to undesired outputs, overfitting or underfitting [40]. The same applies to our results. During the demos, some experts were prompting safety-related questions however the vector database does not include any information regarding safety standards or regulations. This way the outputs did not match experts' expectations. The same applies to the generated risk assessment document. We fed the vector database with relatively small numbers of documents. Insufficient detail and completeness issues might be related to a lack of data, RAG architecture, and data quality. For example, the complexity of the RAG implementations may impact the generated content quality. This study suggests that naive RAGs can lead to hallucinations [26]. Additionally, the importance of data quality was highlighted by the demos. The model performed better when the prompt and training data matched, highlighting the importance of high-quality and diverse data. 6.2.2.3.

These findings contribute to practitioners and researchers by showing that results from LLMs should be carefully used and not be relied on completely.

**Key Insights**

- Experts found the generation of risk assessment more useful as a baseline or initial structure to iterate upon rather than using the completely generated documents as it is.
- Though the experts did not entirely trust the LLM model, they were willing to use such tools in cybersecurity risk assessments.
- The quality of output from the LLM depended on the RAG implementation, which in turn relied on high quality and data.

## 7.2 Research Question 2: Collaboration and assistance while working with cybersecurity risk assessments

The second research question examined the potential for the LLM to assist a human practitioner with various tasks in the process of creating a cybersecurity risk assessment. Both the interviews and the survey indicate that our experts perceived collaboration with LLMs when working with cybersecurity risk assessment as highly useful. The experts particularly highlighted the ability of the LLM to perform monotonous and time-consuming tasks like data analysis or quickly finding information.

Another aspect of usefulness appreciated by the experts was the potential to have the LLM act as a college or sparring partner when performing highly creative tasks like brainstorming or risk identifications. These findings are supported by previous research by Khojah et al. [39] where they also found LLMs highly useful in many creative scenarios.

Furthermore, eight out of twelve interview participants expressed positive perceptions about the usefulness of assistance from the LLM. This is despite having negative opinions about the level of detail, correctness, and reliability. This shows that even though the output of the LLM needs to be corrected or double-checked by a human, it can still add value and be useful. This was also found by Khojah et al. [39] where they compared asking the LLM for information to asking a college for help. In both cases, you can not be completely certain about the correctness of the information and you are responsible for double-checking it yourself.

Collectively this demonstrates the ability of LLMs to act as a supplementary tool to established practices and processes further supported by 83,33% the survey results saying that they would use LLM in some capacity in their line of work. However, to further increase the efficiency and usability, some key challenges would need to be combated. As with all current-day LLMs, reliability, and correctness of information are still an issue, especially in the more niche domains and use cases[82]. Without improving the ability to produce correct information and provide coherent reasoning, practitioners simply can't rely on the output with enough certainty.

Overall, these research findings align very well with the findings of Nouri et al. [59] who also found LLMs to be a very useful tool in assisting practitioners in certain tasks when working with safety requirements. However, both our study and Nouri et al's study agree that additional research is needed to overcome challenges and limitations such as hallucinations and incorrect information.
These findings can contribute to both researchers and practitioners by highlighting areas where more research and development is needed before it's practical to use a similar tool. Additionally, the findings also show the great variety of tasks where an LLM already can be useful.

**Key Insights**

- Experts saw great potential in collaborating with the LLM during creative tasks like brainstorming but also using it as an assistant when performing monotonous and time-consuming tasks like data analysis.
- A majority of study participants perceived the LLM to be very useful, however, the usability was negatively impacted by the output being too general in nature and information not always correct, and thus not trustworthy.

## 7.3 Research Question 3: Evaluation of Existing Risk Assessments

The third research question investigated the ability of our LLM to evaluate already existing cybersecurity risk assessments. Ten out of twelve experts expressed positive opinions regarding the usefulness of this feature, indicating the perceived value it could bring to the risk assessment process. Many experts interviewed expressed great interest in using LLMs to either find potential improvements or flaws in a risk assessment or to act as a redundancy check, where the LLM can detect missing information or overlooked areas.

Interview participants also saw the potential of combining the evaluation (RQ3) and assistance (RQ2) by having the LLM act as a conversation partner when conducting manual review by, for example, offering guidance or making sure essential steps are carried out correctly. The combination of the two tasks shows the breadth and variety of the LLM's potential.

While the LLM may perform very well when asked general questions, one major challenge still remaining when evaluating domain-specific cybersecurity risk assessment is the level of detail in the output. Both the interview and survey data clearly show the level of detail provided by the LLM is not sufficient to be relied upon as the sole source of evaluation. Additionally, the low trust in LLMs, especially in safety and security-critical areas, also dampens the viability of using the LLM as a stand-alone evaluator, once again manifesting the LLM's primary use as a supplementary or collaborative tool.

Once again, these findings are similar to that of other research in the domain. Nouri et al.[59] also stress that their research highlights the need for human oversight when LLMs are used in safety-critical operations.

These findings contribute to practitioners by showing that LLMs can already be used to improve existing risk assessments but more development and research are needed before they can be relied upon entirely. To researchers, the findings can help guide future research by contributing to uncovering important challenges yet facing the broad implementation of LLM.

**Key Insights**

- The evaluation of existing risk assessments was perceived as the most useful feature by the experts, especially having the LLM as a redundancy check to make sure no critical information is missing.
- However, both the interview and survey data clearly show the level of detail provided by the LLM is not sufficient to be relied upon as the sole source of evaluation.

## 7.4 Reflections on Prompting

Large language models are a relatively new technology. Even though it is a popular topic, many users expressed the need for additional training on how to produce high-quality prompts to get the best results possible. Considering the experts' experiences with AI applications and their statements regarding additional training in constructing prompts, it would be useful to provide a guideline on how to get good results. This could include instruction on using established prompt engineering techniques like few-shot prompting. This further shows yet untapped potential in the use of LLMs.

Another finding is that the quality of the results would increase when more time is spent on prompting. Demo statistics revealed that most of the experts updated the prompts which means that they spent time on that. On the other hand, one expert quoted the need for time for prompting in 6.2.2.3. Also, interviewee 2 quoted that the results were getting better during the demo c as referenced in 6.2.2.3.

## 7.5 Ethical considerations

When conducting interviews and handling, surveys, demos, and interview data, special attention was paid to ethical aspects such as consent and anonymization. The ethical guidelines and checklist suggested by P. E. Strandberg, 2019 were considered in every step of this study by, for example, anonymizing all interview, survey, and demo data. [77]. Additionally, the LLM was chosen and developed with ethical AI principles in mind as presented by Taiwo et al., 2023 [53]. These include, for example, transparency, accountability, safety, security and reliability. By following these guidelines we ensured that AI was used and developed responsibly without causing any harm.

Lastly, the European Union recently passed the *AI Act*, classifying different use cases of AI into risk categories, as shown in Figure 7.1. Each level of increased risk requires the creator or publisher to take measures to ensure that safety and ethical requirements are met. The limited risk category encompasses all AI capable of generating content such as images, texts, video, or audio. The category also includes AI tasked with improving the result of something a human has created, for example, a chatbot capable of improving a novel written by an author7.1. Since the LLM developed in this study is capable of both generating text and improving human-created content, it falls into the limited risk category, requiring all generated outputs to be clearly labeled as such to fulfill the transparency requirement mandated by the *AI Act*. This was done for the entire risk assessment document created by the LLM.

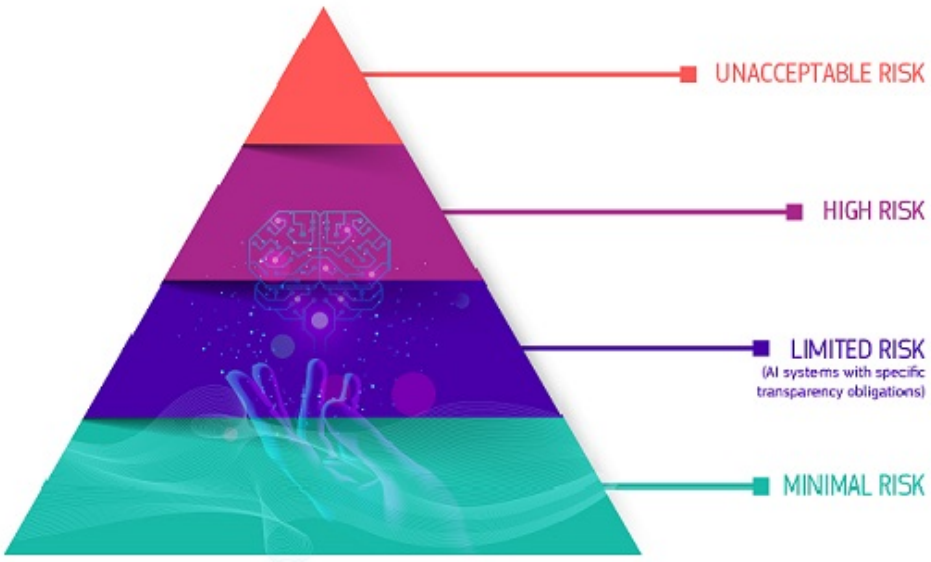



**Figure 7.1:** The regulatory framework defines 4 levels of risk for AI systems [21]

## 7.6 Validity

**Internal validity** In order to mitigate threats to the internal validity, measures were taken to ensure the data from the interviews were consistent and accurate. Each interview was conducted based on the same interview guide to ensure a baseline of data was

collected in each interview. The selected experts had an average of more than 15 years of experience working with risk assessment, this made sure the insights gathered were reliable and relevant to our specific research. Lastly, to further strengthen the internal validity of the research, data triangulation was introduced by asking the experts to both participate in a demo study and fill out a survey in addition to the semi-structured interview. This ensured that potential bias or misunderstandings when interpreting the data was greatly reduced. Another aspect of the internal validity that could be threatened was the consistency of the LLMs output. Variations on the prompt and settings also change the output of the LLM. In order to address this we always ran each prompt and set of settings at least five times and selected the output best representing the general response.

**External validity** This study focuses on one specific research area, namely cybersecurity risk assessment for autonomous forestry machines, which may threaten the study's external validity by being highly specialized. However, the procedure outlined in the study could easily be replicated in any other domain as the training data primarily came from outside the forestry domain that was being researched. Another threat to the validity was the lack of diversity and the relatively small sample size of the interview participants. We recommend that future studies include more diversity from different sectors and domains and potentially a larger sample to enhance the generalization of the results Lastly, the choice of model could also impact the findings. In the study, we were limited by the available memory on the local server, other studies without this limitation might choose a bigger model with 70B parameters or more which according to benchmark studies has slightly better performance when doing more complex tasks.
**Construct validity** Construct validity was ensured by clear definitions of regularly occurring concepts and codes. This ensured a common understanding among researchers, interview participants, and domain experts.

**Reliability validity**
The reliability of the research was ensured by having as much transparency as possible. Interview guides and the survey can be found in the appendix with examples of codes and coding practices in the thesis. Additionally, the answers to the survey and many quotes from the interviews are included in the results section to strengthen the reliability of the evidence.

## 7.7 Future Research

The risk assessment process is a quite long process which contains many steps. In addition to steps, there are lots of domain-specific standards. The risk assessment process includes a risk mitigation strategy section which is also a long process. Including multiple relevant standards would increase the model strength and reduce the bias in the output.

Most of the experts at RISE work in both safety and cybersecurity. Safety and cybersecurity concerns interplay with each other. Because of this reason, some of the cybersecurity concerns might be missed by the model. Safety concerns should be evaluated to minimize cybersecurity concerns in future research including the other steps in cybersecurity risk assessments for instance threat mitigation strategies.

Several new models are arising thanks to rapid developments in large language models. Some models have already surpassed the Llama 2 7B model based on some benchmark studies such as Llama 3 8B. In the risk assessment document, we needed to provide tree representations for the possible attack trees. Also, some interviewees asked about the possibility of adding charts or figures for the generated risk assessment during the interviews. Integrating text-to-image models into more developed LLM may satisfy experts more in this area which is a good future research candidate.

# 8

# Conclusion

This thesis used a design science approach to explore the capabilities of LLMs in cybersecurity risk assessment for autonomous forestry machines by identifying challenges, investigating use cases, and evaluating the quality of the LLM's output. The central artifact developed was a tool aimed to aid practitioners in the cybersecurity risk assessment process. The tool was LLM-based, using risk assessment training data from other domains such as railway and shipping and a RAG implementation. The artifact was capable of generating a cybersecurity risk assessment for the AGRARSENSE project. The risk assessment was evaluated by domain practitioners from RISE and Komatsu Forest. By conducting a mixed methods study using interviews, surveys, and demos, we found that domain experts perceived the LLM as useful as an assistant (RQ2) when performing monotonous or time-consuming tasks and doing more creative activities like brainstorming. Additionally, the study participants found the evaluation (RQ3) of human-made cybersecurity risk assessments another significantly useful area alongside generation of risk assessment (RQ1) which can be used as a baseline or source of inspiration rather than directly applying them. However, challenges and limitations to the applicability of implementing the LLM were identified, such as undesired hallucinations, unreliable information, and inability to perform calculations. Our findings indicate that even though the trust in the LLM's output was low among the experts, they still found it overwhelmingly useful. This highlights the future potential of LLM while underscoring the need for human oversight. The findings align well with similar research findings in the area and show that future work and research are needed to further improve the usability and capability of LLMs when used in safety-critical domains.

# Bibliography


[1] Earth-moving machinery and mining – autonomous and semi-autonomous machine system safety, 2019.

[2] Jul 2022.

[3] Nvivo leading qualitative data analysis software (QDAS) by. `https://lumivero.com/products/nvivo/.`, November 2022. Accessed: 2024-8-11.

[4] ISO 3691-4: 2020. Industrial trucks—safety requirements and verification—part 4: Driverless industrial trucks and their systems, 2020.

[5] Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569*, 2019.

[6] P. Albizu-Urionabarrenetxea, Eduardo Esteban, and Elena Román-Jordán. Safety and health in forest harvesting operations. diagnosis and preventive actions. a review. *Forest Systems*, 22:392–400, 12 2013.

[7] Peter Assarsson. Agrarsense project & komatsu. https://www.komatsuforest.com/media/newsroom/agrarsense-project, 2024.

[8] Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. Open llm leaderboard. `https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard`, 2023.

[9] BIG bench authors. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023.

[10] G Bharathi Mohan, R Prasanna Kumar, P Vishal Krishh, A Keerthinathan, G Lavanya, Meka Kavya Uma Meghana, Sheba Sulthana, and Srinath Doss. An analysis of large language models: their impact and potential applications. *Knowledge and Information Systems*, pages 1–24, 2024.

[11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[12] C Busch, PS De Maret, T Flynn, R Kellum, S Le, B Meyers, M Saunders, R White, and M Palmquist. Content analysis. writing@ csu. *Colorado State University Department of English*, 2005.

[13] Jianfei Chen, Yu Gai, Zhewei Yao, Michael W Mahoney, and Joseph E Gonzalez. A statistical framework for low-bitwidth training of deep neural networks. *Advances in neural information processing systems*, 33:883–894, 2020.

[14] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonza-

lez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference, 2024.

[15] International Electrotechnical Commission. International electrotechnical commission. https://www.iec.ch/blog/understanding-iec-62443, 2024.

[16] Marvin Damschen. Agrarsense. https://www.ri.se/sv/vad-vi-gor/projekt/agrarsense, 2023.

[17] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023.

[18] Prerit Datta, Natalie Lodinger, Akbar Siami Namin, and Keith S. Jones. Predicting consequences of cyber-attacks. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 2073–2078, 2020.

[19] Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Llm.int8(): 8-bit matrix multiplication for transformers at scale, 2022.

[20] ENISO DIN. 18497: 2019-08 agricultural machinery and tractors–safety of highly automated agricultural machines–principles for design (iso 18497: 2018).

[21] Content Directorate-General for Communications Networks and Technolog. regulatory-framework- ai. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework- ai, 2024.

[22] Volkan Evrin. Risk assessment and analysis methods: Qualitative and quantitative. https://www.isaca.org/resources/isaca-journal/issues/2021/volume-2/risk-assessment-and-analysis-methods, 2021.

[23] Hugging Face. The hugging face course, 2022. `https://huggingface.co/course`, 2022. [Online; accessed 2024-8-9].

[24] Michael Felderer, Christian Haisjackl, Ruth Breu, and Johannes Motz. Integrating manual and automatic risk assessment for risk-based testing. In *Software Quality. Process Automation in Software Development: 4th International Conference, SWQD 2012, Vienna, Austria, January 17-19, 2012. Proceedings 4*, pages 159–180. Springer, 2012.

[25] Tarek Gaber, Yassine El Jazouli, Esraa Eldesouky, and Ahmed Ali. Autonomous haulage systems in the mining industry: cybersecurity, communication and safety issues and challenges. *Electronics*, 10(11):1357, 2021.

[26] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

[27] Georgi Gerganov. llama.cpp: LLM inference in C/C++.

[28] Judy Wawira Gichoya, Kaesha Thomas, Leo Anthony Celi, Nabile Safdar, Imon Banerjee, John D Banja, Laleh Seyyed-Kalantari, Hari Trivedi, and Saptarshi Purkayastha. AI pitfalls and what not to do: mitigating bias in AI. *British Journal of Radiology*, 96(1150):20230023, 09 2023.

[29] Yikun Han, Chunjiang Liu, and P Wang. A comprehensive survey on vector database: storage and retrieval technique. *Challenge. arXiv preprint arXiv*, 231011703, 2023.

[30] Chadi Helwe. *Evaluating and Improving the Reasoning Abilities of Language Models*. Theses, Institut Polytechnique de Paris, July 2024.

[31] Alan R. Hevner, Salvatore T. March, Jinsoo Park, and Sudha Ram. Design science in information systems research. *MIS Quarterly*, 28(1):75–105, 2004.

[32] Herman Holgert. Autonomous forest machines: Enable technology shift with new business models, 2021.

[33] International Organization for Standardization/International Electrotechnical Commission. Iso 27001:2005: Information technology – security techniques – information security management systems – requirements, 2005.

[34] EN ISO. 12100: 2010: Safety of machinery—general principles for design—risk assessment and risk reduction. *Reference Source*, 2010.

[35] Adam Tauman Kalai and Santosh S Vempala. Calibrated language models must hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 160–171, 2024.

[36] Ramanpreet Kaur, Dušan Gabrijelčič, and Tomaž Klobučar. Artificial intelligence for cybersecurity: Literature review and future research directions. *Information Fusion*, 97:101804, 2023.

[37] Tobias Kerner. Domain-specific pretraining of language models: A comparative study in the medical field. *arXiv preprint arXiv:2407.14076*, 2024.

[38] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022.

[39] Ranim Khojah, Mazen Mohamad, Philipp Leitner, and Francisco Gomes de Oliveira Neto. Beyond code generation: An observational study of chatgpt usage in software engineering practice. *Proceedings of the ACM on Software Engineering*, 1(FSE):1819–1840, 2024.

[40] Monique F Kilkenny and Kerin M Robinson. Data quality:"garbage in–garbage out", 2018.

[41] Kyounggon Kim, Jun Seok Kim, Seonghoon Jeong, Jo-Hee Park, and Huy Kang Kim. Cybersecurity for autonomous vehicles: Review of attacks and defense. *Computers & Security*, 103:102150, 2021.

[42] David Kirk et al. Nvidia cuda software and gpu parallel computing architecture. In *ISMM*, volume 7, pages 103–104, 2007.

[43] Eric Knauss. Constructive master's thesis work in industry: guidelines for applying design science research. In *2021 ieee/acm 43rd international conference on software engineering: Software engineering education and training (icse-seet)*, pages 110–121. IEEE, 2021.

[44] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[45] Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore, December 2023. Association for Computational Linguistics.

[46] Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`, 5 2023.

[47] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

[48] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

[49] Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. A survey of transformers. *AI open*, 3:111–132, 2022.

[50] Ye Lin, Yanyang Li, Tengbo Liu, Tong Xiao, Tongran Liu, and Jingbo Zhu. Towards fully 8-bit integer inference for the transformer model. *arXiv preprint arXiv:2009.08034*, 2020.

[51] James Lowenberg-DeBoer, Karl Behrendt, Melf-Hinrich Ehlers, Carl Dillon, Andreas Gabriel, Iona Yuelu Huang, Ian Kumwenda, Tyler Mark, Andreas Meyer-Aurich, Gabor Milics, et al. Lessons to be learned in adoption of autonomous equipment for field crops. *Applied Economic Perspectives and Policy*, 44(2):848–864, 2022.

[52] Kiran Maharana, Surajit Mondal, and Bhushankumar Nemade. A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*, 3(1):91–99, 2022.

[53] Kolade Makinde, Esther Taiwo, Edward Tella, Mayowa Akinwande, and Ahmed Akinsola. A review of the ethics of artificial intelligence and its applications in the united states. *International Journal on Cybernetics & Informatics*, 12, 10 2023.

[54] Potsawee Manakul, Adian Liusie, and Mark JF Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*, 2023.

[55] Ggaliwango Marvin, Nakayiza Hellen, Daudi Jjingo, and Joyce Nakatumba-Nabende. Prompt engineering in large language models. In *International conference on data intelligence and cognitive informatics*, pages 387–402. Springer, 2023.

[56] Bhaskar Mitra, Nick Craswell, et al. An introduction to neural information retrieval. *Foundations and Trends® in Information Retrieval*, 13(1):1–126, 2018.

[57] Mazen Mohamad, Ramana Reddy Avula, Peter Folkesson, Pierre Kleberger, Aria Mirzai, Martin Skoglund, and Marvin Damschen. Cybersecurity pathways towards ce-certified autonomous forestry machines. *arXiv preprint arXiv:2404.19643*, 2024.

[58] Michael Mylrea, Sri Nikhil Gupta Gourisetti, and Andrew Nicholls. An introduction to buildings cybersecurity framework. In *2017 IEEE symposium series on computational intelligence (SSCI)*, pages 1–7. IEEE, 2017.

[59] Ali Nouri, Beatriz Cabrero-Daniel, Fredrik Törner, Håkan Sivencrona, and Christian Berger. Engineering safety requirements for autonomous driving with large language models. *arXiv preprint arXiv:2403.16289*, 2024.

[60] Luiz FP Oliveira, António P Moreira, and Manuel F Silva. Advances in forest robotics: A state-of-the-art survey. *Robotics*, 10(2):53, 2021.

[61] Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. Scaling neural machine translation. *arXiv preprint arXiv:1806.00187*, 2018.

[62] James Jie Pan, Jianguo Wang, and Guoliang Li. Survey of vector database management systems. *The VLDB Journal*, pages 1–25, 2024.

[63] Heimo Pentikäinen and Timo Malm. *Cybersecurity in Autonomous Machine Systems Development*. Number VTT-R-01087-19 in VTT Research Report. VTT Technical Research Centre of Finland, Finland, November 2019.

[64] Stephen Phillips, Steve Taylor, Michael Boniface, and Mike Surridge. Automated knowledge-based cybersecurity risk assessment of cyber-physical systems. *Authorea Preprints*, 2023.

[65] Gestão Produção, Yan Feng, and Jean-François Audy. Forestry 4.0: a framework for the forest supply chain toward industry 4.0. *Gestão & Produção*, 27, 12 2020.

[66] Asif Raihan. Sustainable development in europe: A review of the forestry sector's social, environmental, and economic dynamics. *Global Sustainability Research*, 3:72–92, 09 2023.

[67] Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, SM Tonmoy, Aman Chadha, Amit P Sheth, and Amitava Das. The troubling emergence of hallucination in large language models–an extensive definition, quantification, and prescriptive remediations. *arXiv preprint arXiv:2310.04988*, 2023.

[68] RISE. Rise. https://www.ri.se/en/about-rise/operations/organisation, 2023.

[69] Morgan Rossander and Håkan Lideskog. Design and implementation of a control system for an autonomous reforestation machine using finite state machines. *Forests*, 14(7), 2023.

[70] Ankit Satpute, Noah Gießing, André Greiner-Petter, Moritz Schubotz, Olaf Teschke, Akiko Aizawa, and Bela Gipp. Can llms master math? investigating large language models on math stack exchange. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2316–2320, New York, NY, USA, 2024. Association for Computing Machinery.

[71] Sheng Shen, Zhen Dong, Jiayu Ye, Linjian Ma, Zhewei Yao, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Q-bert: Hessian based ultra low precision quantization of bert. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8815–8821, 2020.

[72] Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. In chatgpt we trust? measuring and characterizing the reliability of chatgpt. *arXiv preprint arXiv:2304.08979*, 2023.

[73] John Shutske, Kelly Sandner, and Zachary Jamieson. Risk assessment methods for automated agricultural machines: Current practice and future needs. 01 2022.

[74] Ömer Söner, Gizem Kayisoglu, Pelin Bolat, and Kimberly Tam. Cybersecurity risk assessment of vdr. *The Journal of Navigation*, 76(1):20–37, 2023.

[75] Heydar Soudani, Evangelos Kanoulas, and Faegheh Hasibi. Fine tuning vs. retrieval augmented generation for less popular knowledge. *arXiv preprint arXiv:2403.01432*, 2024.

[76] Michal Sterbak, Pavel Segec, and Jan Jurc. Automation of risk management processes. In *2021 19th International Conference on Emerging eLearning Technologies and Applications (ICETA)*, pages 381–386. IEEE, 2021.

[77] Per Erik Strandberg. Ethical interviews in software engineering. In *2019 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pages 1–11, 2019.

[78] Vijay Vaishnavi and B Kuechler. Design science research in information systems. *Association for Information Systems*, 01 2004.

[79] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[80] Vincent R. Waldron. Interviewing for knowledge. *IEEE Transactions on Professional Communication*, PC-29(2):31–34, 1986.

[81] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. 2023.

[82] Weixuan Wang, Barry Haddow, Alexandra Birch, and Wei Peng. Assessing the reliability of large language model knowledge, 2023.

[83] Xiaohua Wang, Zhenghua Wang, Xuan Gao, Feiran Zhang, Yixin Wu, Zhibo Xu, Tianyuan Shi, Zhengyuan Wang, Shizheng Li, Qi Qian, et al. Searching for best practices in retrieval-augmented generation. *arXiv preprint arXiv:2407.01219*, 2024.

[84] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

[85] Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning*, pages 38087–38099. PMLR, 2023.

[86] Frank F Xu, Uri Alon, Graham Neubig, and Vincent Josua Hellendoorn. A systematic evaluation of large language models of code. In *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*, pages 1–10, 2022.

[87] Xue Ying. An overview of overfitting and its solutions. In *Journal of physics: Conference series*, volume 1168, page 022022. IOP Publishing, 2019.

[88] Ofir Zafrir, Guy Boudoukh, Peter Izsak, and Moshe Wasserblat. Q8bert: Quantized 8bit bert. In *2019 Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing-NeurIPS Edition (EMC2-NIPS)*, pages 36–39. IEEE, 2019.

# A

# Appendix 1

## A.1 Interview Guide - 1

**Date of the interview:**
**Interviewee:**
**Cycle:**
**Start-Time - End Time**
**Actions before the interview:**

- Get verification that we are going to record the interview.
- Send invitation
- Show research questions and some results on the screen but also send them by sending an invitation.

**Introduction:**

- Show gratitude
- Introduce ourselves
- Mention the purpose of the interview
- Introduce our thesis and research questions shortly
- We are going to record the interview with your consent to evaluate the answers after the interview. However, all personal details will remain anonymous, and all responses will be kept confidential.
- This is a semi-structured interview, which means that we have some prepared questions, however the conversation can be evolved based on the responses.
- Please feel free to ask any questions or seek clarifications at any point during the interview. We value your input and look forward to a productive discussion.

**Pre-Questions**

- What is your role in RISE, could you briefly explain?
- How much experience do you have in safety or/and cybersecurity
- Have you used any large language models before?
- Do you use LLM at work?
- Explain that we have 3 RQs

**Mid Part**

- Do you agree with the risk assessment?
  - What parts should be improved?
- What features or capabilities do you think are currently missing in the LLM that would be beneficial for risk assessments? Why?
- How would you compare the risk assessments generated by the LLM to those created by human experts? Why?
- In general, there are 3 ways to interact with an LLM, those are Instruction to perform a task, question answering and conversational. In what way would you like

to interact with the LLM?
- Would you use the LLM in the risk assessment process for:
  - assisting/ getting help
  - generating a new risk assessment - Why and how
  - evaluating a new risk assessment - Why and How
- What are the potential limitations or challenges you foresee in using LLMs for risk assessments?
- What would you do to mitigate them?

**Wrap Up**
- What is your general impression of the result from the model?
- Are there any apparent flaws?
- What would you improve?
- What would you focus on?
- Any additional thoughts you would like to share or feedback for us?

Thank you very much for talking to us

## A.2 Interview Guide - 2

**INTERVIEW GUIDE 2 Date of the interview:**
**Interviewee:**
**Cycle:**
**Start-Time - End Time**
**Actions before the interview:**
- Get verification that we are going to record the interview.
- Show research questions and document

**INTRODUCTION**
- Introduce ourselves
- Mention Thesis
- Confidentiality & Recording

**GENERAL QUESTIONS**
- Could you please describe your overall thoughts on the output of the large language model? (For the document)

**RQ1**
- Are the model's outputs aligning with your expectations? (Expectation is full risk)
  - If not, how do they differ?
- How complete is the risk assessment document generated by LLM?
  - Is it detailed enough?
  - Are there any aspects or details missing?
- How relevant do you believe the identified risks are in the context of AGRARSENSE?
  - Are the identified risks relevant to this context?
- Are you familiar with Standard 62443?
  - If YES, How well do you think the risk assessment adheres to and follows the standard 62443?
  - If NO, just skip
- How do our risk assessments compare to those typically created by human experts?

**RQ2**

- Are there specific tasks or steps in your risk assessment process where the LLM could be particularly beneficial?

**RQ3**

- How well does the LLM find improvements to existing risk assessments?
  - Do you agree with the LLM's evaluations and suggestions?
  - Would you follow the suggestions?

**Future Work**

- What specific improvements would you suggest for the LLM to enhance its capabilities in terms of generating risk assessment, evaluating existing risk assessment, and assisting in those two?
- Are there any biases in the model's output that need to be addressed?

**Discussion Section**

- What would be your concerns or considerations before integrating this technology?
- Would you rely on the LLM's feedback to make changes to your risk assessment?
- Can you say any challenges or difficulties you encountered or might encounter in the future?
- How likely will you use this model for future risk assessments? - Why?

**Wrap Up**

- Any additional thoughts you would like to share or feedback for us?

Thank you very much for talking to us!

## A.3 Demo Guide

We will let you use the model by giving you the link for the interface. We would like you to do some tasks and observe you. Our focus is on your interaction with the model, your prompting style, and your feedback from the output. This part should take no more than 10 minutes. You can either follow our task or prompt any question you want.

**TASK** In task one, you are the cybersecurity expert and we would like you to create a prompt that lists primary assets for the AGRARSENSE project.

## A.4 Survey Guide

To what extent do you agree with the following statements? *

*Markera endast en oval per rad.*

| | Strongly Disagree | Moderately Disagree | Slightly Disagree | Slightly Agree | Moderately Agree | Strongly Agree | Don't know |
|---|---|---|---|---|---|---|---|
| **AI generated risk assessment is complete** | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |
| **AI generated risk assessment is detailed** | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |
| **The content in the generated risk assessment is revelant** | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |
| **I would use a similar in some capacity if avaliable** | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |
| **AI generated risk assessment is trustworthy** | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |
| **A similar AI model would be useful in my line of work** | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

**Figure A.1:** Survey Guide

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**
**CHALMERS UNIVERSITY OF TECHNOLOGY**
Gothenburg, Sweden
www.chalmers.se


CHALMERS
UNIVERSITY OF TECHNOLOGY