UPPSALA
UNIVERSITET

UPTEC XX XXXXX

Degree project 30 credits

February, 2026

# AI Support in Project Management

## Exploring Automation Potential within Vattenfall's Technical Workflows and Processes

Sebastian Laggeby

Master's Programme in Sociotechnical Systems Engineering

UPPSALA
UNIVERSITET

# AI Support in Project Management

Sebastian Laggeby


## Abstract

Projects consisting of maintaining nuclear power plants often involve tasks that are documentation-intensive, and must satisfy strict requirements in safety, confidentiality, and traceability. These tasks, that are associated with knowledge work, can be highly time consuming. This thesis investigates how large language models (LLMs) can be implemented to assist in project management workflows from both technical and social perspectives. This was done through a qualitative study of a process within Forsmark nuclear power plant.

The results indicated that interview participants saw value in AI-support for tasks such as assistance and drafting in writing reports, evaluating and summarizing documents, coding, translating, and educating. However, there were also fears of implementing AI, as it might cause over-reliance on the technology and can hallucinate. Based on identified requirements an on-premise solution is proposed, centered on RAG with different modes depending on use case. Deployment of such a system is recommended to be incremental to allow for early testing, starting with a semantic search only mode first, as that was identified to potentially be of value. The study concludes that an LLM-integration is possible if designed around the identified requirements, and implemented as socio-technical system, with clear guidelines for use and training for employees.



**Faculty of Science and Technology**

**Uppsala University, Place of publication Uppsala**

Supervisor: Björn Molander   Subject reader: Nicolas Pielawski

Examiner: Elísabet Andrésdóttir

# Sammanfattning

Underhållsprojekt inom kärnkraft innehåller ofta moment som består av hantering av stora mängder dokumentation som måste uppfylla strikta krav på säkerhet, konfidentialitet, och spårbarhet. Dessa moment, som kan kopplas till kunskapsarbete, är ofta mycket tidskrävande. Detta examensarbete utforskar hur stora språkmodeller (LLMs) kan integreras som hjälpmedel inom processer och arbetsflöden från både en teknisk och social ståndpunkt, genom en kvalitativ studie av konstruktionsprocessen på Forsmarks kärnkraftverk.

Resultatet indikerar att respondenterna upplevde ett värde i AI-stöd för uppgifter som att utforma underlag och skriva rapporter, utvärdera och sammanfatta dokument, koda, översätta och utbilda. Samtidigt upplevdes vissa rädslor för AI, då det skulle kunna leda till överberoende av tekniken, och att LLM:er kan hallucinera.

Baserat på krav identifierade från intervjuerna så föreslås en lokal (on-premise) lösning centrerad kring tretrieval-augmented generation (RAG), med flera olika inställningar beroende på användningsområden. Införandet av detta system rekommenderas att göras stegvis för att möjliggöra tidig testning, med första steget bestående endast av semantisk sökning då detta identifierades som en potentiellt värdefull funktion. En kostnadsuppskattning gjordes, som indikerade att en sådän här investering potentiellt kan vara finansiellt gångbar. De tre största kostnadsdrivarna identifierades vara konstruktion-, hårdvaru- och personalkostnader.

Slutligen dras slutsatsen att en LLM-integration är möjlig om den konstrueras utifrån de identiferade kraven, och implementeras som ett socio-tekniskt system snarare än endast teknisk, med tydliga riktlinjer för användande, och kurser som detaljerar användning av LLM:er.

# Contents

# 1 Introduction

As we move towards an ever developing digital world, the demand for increased electricity production is apparent; in Sweden consumed electricity is expected to roughly double in the coming 20 years [16]. Meeting these demands, while maintaining fossil-free generation requires both development of current systems, and increased electricity production. Due to this, nuclear energy has been portrayed as a potential main contributor to a renewable, reliable, and expanding electricity generation by the government. Currently, after long periods of dismantling nuclear in Sweden, the government has started the proceedings of building new nuclear power plants within a ten year time frame [22].

Supplying the increased demand does not only entail building new nuclear power plants, it also entails maintaining production in the current ones. Currently a significant part of operating nuclear power plants consists of upgrading, modernizing and replacing outdated systems, while adhering to strict safety and regulatory policies. These resource intensive projects usually involves multiple different stages that require a highly skilled and educated staff, but the energy sector is currently facing pressing skill shortages which may cause delays in current operations and a delayed expansion of the sector [23]. This raises a question: how can productivity be increased with the current resource constraints, without compromising quality or safety? One avenue that might prove fruitful is targeting time spent on document-heavy knowledge work tasks by improving information handling.

Generative artificial intelligence is a field that has developed rapidly as of late, and has also created new possibilities for supporting knowledge work. Mainly, large language models (LLMs) have been adopted across many companies due to their language and reasoning skills. However, deploying LLMs in a safety-critical environment such as nuclear energy, where consequences of errors can be dire, can come with some challenges. LLMs have a tendency to hallucinate [21]; writing statements that sound factually correct but actually are not. Furthermore, due to the nature of nuclear, a high degree of confidentiality is needed, and traceability in decision making is paramount. Thus, a general, ”run-of-the-mill” chatbot might not be a sufficient solution.

One popular way to combat hallucinations in LLMs is through retrieval-augmented generation (RAG), where instead of generating answers based solely on the internal parameters of the LLM, it is also connected to an external database from where information is retrieved. Architecture like this can improve accuracy, and allows updating data without retraining the LLM [20].

Still, an LLM is only useful if it is accepted by its intended user. Whether a tool is effective is not only a matter of technically constructing it, but also a matter of ensuring that the transition period is handled smoothly. Implementing an LLM-system for the purposes improving productivity may then, apart form being a technical challenge, also be a challenge of managing cultural change.

This thesis aims to examine these topics through a case study at Vattenfall AB and Forsmark nuclear power plant, by studying project workflows in their construction process.

## 1.1 Purpose, aims and motivation

The purpose of this masters thesis is to explore to what extent LLMs can be integrated in project workflows and tools within construction processes, and to propose possible improvements and systems that could satisfy any requirements. It also examines how such a change might affect the intended users of those systems. Furthermore, the thesis conducts a cost estimation to identify cost-drivers and assess feasibility. This by examining current processes, holding interviews both within Vattenfall and Forsmark, and studying relevant literature.

This is done with the following research questions:

- Can LLMs be introduced in construction processes in order to streamline project-management workflows?
- Which LLM-based approach and system components can meet the requirements identified in construction-related project workflows, and what are the main cost drivers and feasibility trade-offs of those features?
- What social effects could such an integration have on organizational structure and company culture?

## 1.2 Delimitations

In order to limit the scope of the thesis several delimitations have been made:

The objective of the thesis is not to design, develop, implement or test any LLM, instead it proposes requirements and a architectural structure based on the case context.

The empirical focus is limited to the construction process at Forsmark. Thus, it should mainly be considered as dependent on the context rather than being necessarily generalizable.

Any benchmarking of LLMs or RAG is outside of the scope due to the high level of confidentiality regarding documents at Forsmark.

The cost estimation is constructed as an order-of-magnitude estimate based heavily on assumption in order to identify the major cost drivers, rather than an exact cost analysis.

## 1.3 Disposition

The thesis is structured as follows:

- Chapter 2 (Background) presents the case specific context of Vattenfall and Forsmark and the organizational structure of the construction process, and introduces the theoretical background of the study. This includes knowledge work, generative artificial intelligence, LLMs and RAG, and lastly the ADKAR model based in change management.
- Chapter 3 (Method) presents the methods used for conducting the study, along with the interview design, choice of sample and thematic analysis.
- Chapter 4 (Results) presents the themes identified from analyzing the interviews. This is firstly done through a general description of working in the construction process, then some use cases and requirements for LLMs is presented, along with the case management study.
- Chapter 5 (Discussion) interprets the findings in relation to the research questions and proposes a technical approach of a high-level design of a system that might satisfy the identified requirements.
- Chapter 6 (Conclusions) summarizes findings, and presents future work.

# 2 Background

This chapter aims to provide the necessary context of the thesis. It starts with detailing the case specific information, regarding Vattenfall, Forsmark, and its processes. Then the general theory regarding Knowledge work, Generative AI, and change management with the ADKAR-model is presented.

## 2.1 Vattenfall

Vattenfall AB is an energy company operating mainly in northern Europe, supplying electricity to 7.7 million customers. The company is the largest energy company in Sweden, and is owned entirely by the Swedish government, with the main offices located in Stockholm [40]. In 2017 Vattenfall committed to a net-zero emission policy by 2040, and to reduce their emissions by 77% by 2030. In order to achieve this, they are phasing out fossil fuels and investing in renewable energy [45], and had as of 2024 reduced carbon emissions from the their own operations by 53% [44]. In 2023 the largest source of power generated at Vattenfall was nuclear, with Vattenfall operating five nuclear reactors in Sweden [43]. In total hydro, wind, and nuclear generated 98% of electricity in Sweden in 2023 [24].

In order to reduce carbon emissions while increasing supply to an expected doubling of consumed electricity by 2040, in 2023 the Swedish government changed energy production targets from being renewable to instead being carbon-free, with an emphasis on increased nuclear power production [24]. Vattenfall owned company Videberg Kraft AB just recently submitted an application for subsidies from the state to build new nuclear plants [41], which marks the launch of of the construction of new nuclear in Sweden.

Three of the currently five active Vattenfall-owned nuclear reactors are in Forsmark nuclear power plant. This thesis uses the construction process at Forsmark as a case for studying the research questions, the process is described below.

### 2.1.1 The construction process

The construction process at Forsmark is the process that all projects regarding construction of various systems and constructions (henceforth referred to collectively as constructions) follows, and aims to detail how to implement these implementations within the confines of all various specifications, requirements, and necessary models of the implementation. This can be how to build facilities, change parts of a measurement tool, install or upgrade new systems, etc.

The construction process is modeled after the aptly named Waterfall model (though the naming is unrelated to the company). Instead, the name refers to how the workflow is linear, only moving forward to the second phase when the first is finished. Typical for projects following the waterfall model is that the desired outcome is fairly determined upfront, leading to a goal oriented way of working. The entire process is often planned from the start, providing structured and predictable results, with not much room for deviations [38]. The construction process is divided into separate phases, visualized in figure 1.

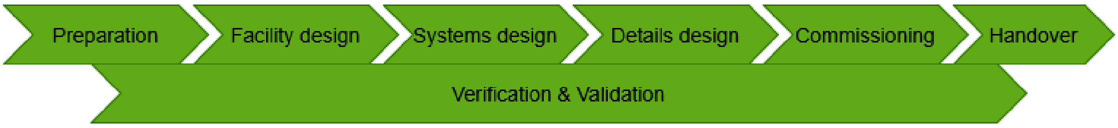


Figure 1: Phases of the construction model

The construction process is part of a larger facility optimization process, and starts when the need for a corrective measure has been identified, and ends when the proper solutions has been designed, documented and handed over to the department that constructs and implements the solutions. Depending on the size of the project, some phases can be excluded if necessary. After each phase a decision is made regarding whether the project is ready to move on to the next phase or not.

The initial phase of the construction process is the technical preparation phase, which largely concerns providing the conditions such that the facility design phase handles the correct instances. A plan for verification & validation is created, detailing how requirements can be tested throughout every phase in the construction process. A radiation safety demonstration plan is also created, detailing how the construction could affect radiation safety assessments. The facility design phase is where the general construction requirements of how a facility operates, or how the construction is intended to operate, are mapped. The systems design phase regards designing individual systems requirements and models. Requirements for every function and its uses is produced, along with requirements regarding maintenance, age analysis, and relevant educational courses. The details construction phase regards how systems can be constructed to fit physically with the current facility. This consists of choosing materials, dimensioning, CAD modeling, creating blueprints, etc. Every previous phase is heavily documented, which is then used in the commissioning phase where the construction is constructed or actualized, by a separate assembly department. When testing of all requirements have been done, documentation is finalized [42].

## 2.2 Knowledge work

The work within the construction process consists largely of knowledge work, that is relatively unstructured and consists of learning, putting to practice, and using knowledge creatively. Knowledge work as a concept is rather ill-defined in literature [34] Instead, focusing on what leads to heightened knowledge work productivity might delimit what is important when managing knowledge workers. Firstly, this means clarifying what their actual task is. In contrast to manual work, where a task might be to assemble parts for a product, it might not be as defined in knowledge work, and part of it is figuring out the task while doing it. This also means figuring out what chores are reducing productivity towards completion of the task. Furthermore, knowledge work often needs a large degree of autonomy to be efficient, and continuous innovation, learning and teaching has to be built into the job [14].

## 2.3 Generative artificial intelligence

The field of Artificial Intelligence (AI) is among the fastest growing, and consists of understanding and building intelligent agents, machines that can act effectively in novel situations with capabilities emulating, or surpassing those of humans [36]. This technology has allowed organizations to change workflows, enabling automation of repetitive tasks, analyzing complex data, deliver personalized assistance, and much more [1]. To define AI is not entirely easy as there are still ambiguities as to what constitutes intelligence. However, for most modern everyday practices the perceived intelligence is created through Machine learning (ML), an approach that has gained popularity in recent decades as it allows the machines to improve and revise their current model by learning from patterns rather than being explicitly programmed, revolutionizing how people interact with AI agents in the process. ML works by allowing the AI to adjust its models parameters based on input in order to minimize the difference between output and a set objective. When the number of these parameters increases this often allows for a more precise output. If these parameters are organized into interconnected layers affecting one another, this is called deep learning. In simple terms, this is the basis for the technological methods that has caused the AI surge that has taken the world by storm in recent years [36].

### 2.3.1 Large language models

Some models can through *natural language processing*, a process where words are probabilistically approximated, understand and generate spoken and written language. These models are called language models, and when they are trained on large sets of data and have a great amount of parameters they are commonly referred to as Large Language models (LLMs) [36]. LLMs have become popular largely due to their general knowledge and capabilities among a large set of different tasks, making them general purpose task solvers [49]. These applications of LLMs can be useful in industry [35]. For example, knowledge workers were found by McKinsey Global Institute to spend about a fifth of their time searching for and gathering information. As LLMs can have a vast knowledge of many different fields, and can be connected to a company's internal information, it can act as a virtual expert – quickly providing answers that could increase efficiency greatly rather than employees searching manually. LLMs can also read large libraries of data in a very short time, which could be in assistance when researching. Apart from this, LLMs can be beneficial in many more fields; such as marketing, customer operations, software engineering, and more [10]. For instance within certain areas of knowledge work, LLMs have been shown to increase speed by 25%, and performance by 40%. Furthermore it can act as an equalizer in the workplace, by assisting those who usually perform worse to produce work of a higher quality [13].

LLMs usually go through three stages of training in order to become accurate [21]:

Firstly, *pre-training* where the LLMs processes large data sets of text, and through autoregressive prediction learns to predict the most probable next coming word in a sentence or sequence. When these predictions are accurate, it gives the LLMs an understanding of language, reasoning, and world knowledge [21]. After pre-training the LLM is optimized to minimize the prediction error, but generated response might not be in line with the users instructions [48].

Secondly, *supervised fine-tuning* aims to fix this mismatch, by training the LLM on pairs of instructions and their desired output that the LLM should generate [48]. This has been proven to render remarkable results in generalization abilities and bridges the gap between the previously generated content and the users instructions [21].

Lastly, *Reinforcement Learning from Human Feedback* aims to further improve the user experience by training the LLMs on preference models, created by showing users a prompt and two answers, and having the users answer which one is preferred [21].

Through these stages an LLM is trained, and by scaling up the size of the training dataset, and the numbers of parameters, along with the amount of training computation have been shown to improve performance [25]. However, for a model to improve efficiency on more challenging tasks such as logic or reasoning, up-scaling alone is not always sufficient. Instead, by allowing the model to create intermediate steps, *chains of thought*, that enables the model to reason incrementally it can increase performance, sometimes considerably [46]. Chain-of-thought prompting is also known as thinking mode in multiple commercially available LLMs.

However, increasing the size of an LLM, and using chain-of-thought prompting, increases the cost of using an LLM, also called the inference cost. For companies (like Vattenfall) that are looking to integrate LLMs in their processes, cost mitigation can be a considerable part of the equation when designing and realizing LLM-systems. One way of reducing costs is saving LLM answers from commonly asked prompts to a database, and instead of generating new answers, accessing the already stored ones when similar prompts are submitted. Another is creating an *LLM cascade*, which consists of chaining multiple LLMs, so that a query is first passed to a smaller, cheaper LLM. If the response is not reliable enough, decided by a regression model, the query is then passed to a second larger model. If the response still is not reliable enough, the query can be passed to an even larger model, and so on, until the answer is of a satisfactory degree [9].

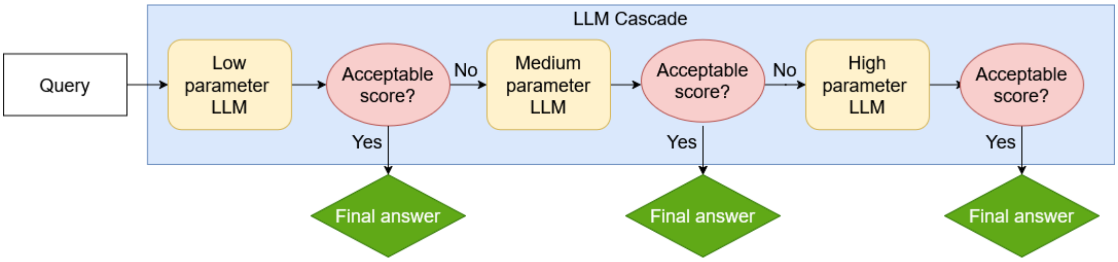


Figure 2: Example of an LLM cascade

### 2.3.2 Technical issues with LLMs

Although LLMs have proven useful, there are still limitations to the technology that needs to be adressed [35]. One of these issues is bias. As LLMs are trained on data from the internet, they can inherit stereotypes, derogatory and exclusionary language towards marginalized groups published there. Social bias can be mitigated throughout multiple stages in the LLM workflow. In the processing of data and prompts, data augmentation techniques can be used to balance data among social groups (for example switching gender pronouns). Reweighting the importance of certain sources, using cost-sensitive learning, and filtering output can also reduce social bias [19].

Bias also extends to non-social factors, for instance LLMs have been shown to sometimes prioritize articles and facts that are more popular than others that might be just as relevant. This is called popularity bias and can lead to filter bubbles [19], and could cause solutions that might be more common to be recommended over those that might be more applicable, which could cause problems when LLMs are used in industry. Users of the LLM might also reduce their usage of the technology if it recommends the most widely known articles. Popularity bias can be mitigated by a surprisingly simple method on the user side, by prompting the LLM to *self-debias* – asking it to not value articles simply based on their popularity [26].

LLMs can be very useful in information seeking due to the incredibly large sets of data that are encoded within them. Though, perhaps the most commonly discussed issue with LLMs are the problems regarding hallucinations, where the LLMs generates content that may seem real but is not actually supported by facts. As LLMs often can be convincing, and generate content that a large amount of the time is factual, these hallucinations can often be hard to detect. This in turn complicates the use case for LLMs [21], especially in safety-critical industries where information needs to be accurate, evidence-based, consistent and verifiable.

Hallucinations are often caused by either flawed data, training, or inference. LLMs are often trained on data from the internet, and this data can contain misinformation, which can be inherited by the LLM. One of the benefits of LLMs are their knowledge of a large amount of fields and subjects, however, when the training data have a limited knowledge regarding a subject, especially in more niche areas, this can cause the LLM to hallucinate more. Models also often have a cut-off date in their data, resulting in sometimes giving answers that are out of date [21].

When pre-training a model, the goal is not necessarily to be factual, but to predict next-coming token in a series of tokens. This may cause a divide between the most likely token, and factuality. When generating from this data, in order to make answers creative and varied, the single most probable token is not chosen, instead stochastic sampling is used – where the token is randomly sampled from a distribution. This randomness, while it has been shown to increase the quality of the text, is also correlated with increased hallucinations [21].

Supervised fine-tuning is used to make the LLM answer in a satisfactory way for the user, but when the model is uncertain it may, instead of generating factual answers or expressing uncertainty, hallucinate answers that seem more satisfactory. These factors may cause generation of non-factual statements, and when the model uses this non-factual context to generate answers it can cause a snowball effect, where answers become more erroneous. Thus, small errors in generation, that could be inconsequential on their own, may lead to larger errors that may mislead users [21].

One way to alleviate hallucinations is through retrieval-augmented generation [21, 20], discussed below.

### 2.3.3 Retrieval-augmented generation

Rather than relying solely on an LLM to have parametric memory from pretraining and fine-tuning, Retrieval-augmented generation (RAG) keeps data in a separate database which the LLM can access on demand. When prompted a retriever accesses the data from the database that is then used as a condition along with the prompt to generate the answer [21]. RAG can substantially improve the accuracy of the answers and reduces the occurrence of hallucinations, which has resulted in RAG becoming popular when a high accuracy of large amounts of data is needed [20]. It also solves the problem of having to retrain or fine-tune LLMs when wanting to update them with new information, by instead simply updating the external database [47]. RAG is not some isolated technology, normally a LLM is created along the steps listed in 2.3.1, and then connected to a RAG system.

RAG works through storing data in a vectorized database (embeddings of text chunks, metadata, and a similarity search index). Connecting an LLM with the database is a *retriever*, consisting of an *embedding model*, and a search index. The database is created through dividing documents into smaller chunks, where one chunk ideally has one semantic meaning, and converting these to vectors using the embedding model. When a query is created, the retriever runs the query through the same embedding model to embed it, and then compares the query vectors to those in the database through a similarity search, to find the most similar ones that are then supplied to the LLM. Usually, after being retrieved, chunks are re-ranked by a more precise ranking model to deliver the most relevant ones to the LLM [47].


Figure 3: Relation between LLM, retriever, and vectorized database.

## 2.4 Change management

In organizations change is ever present. Hence, how to manage such change is a highly required skill [1, 39], especially in industries that rely heavily on technology, where the rate of innovation is high. Change management can be defined as ”the process of continually renewing an organization’s direction, structure, and capabilities to serve the ever-changing needs of external and internal customers”[39]. The emergence of AI technology has had a transformative impact in many organizations as it can be used to automate and streamline processes, workflows, and tasks. Failure to realize upon this

AI potential could lead to organizational stagnation and reduced competitiveness, but failure to manage a change like the implementation of AI could also lead to a loss of competitiveness as employees might not have the knowledge or skills to utilize the new technology or have the proper communication channels to express concerns [1].

While traditional change management principles have prioritized structured planning, clear communication, and stakeholder engagement, AI technology has called for an evolution of change management focusing instead on innovation, agility and continuous learning. This to be adaptable to the ever present technological disruptions. The rapid pace of technological advancement has caused many organizations to move towards more agile methods, prioritizing iteration, quick adaptability and learning. Building an organizational culture where leadership actively communicates this, along with the visions and goal alignment of the changes is paramount, if changes are to be understood and accepted by the workforce [1]. In order to increase the chance for a successful transition between systems, a framework like the ADKAR model can be used, described below.

### 2.4.1 The ADKAR model

The ADKAR model is common in the change management field, and stands for Awareness, Desire, Knowledge, Ability, and Reinforcement, which outline an individuals journey through change. The model provides a framework for both the planning and execution of change in the workplace. The model is useful in diagnosing an employees resistance to change, helping employees transition through the change process, creating a successful action plan for advancement during a change initiative, and developing a change management plan for employees [33]. The elements of ADKAR are explained below.

- First element of the ADKAR model is awareness – if an individual is not **aware** that change is necessary or required, they may be hesitant to accept it. If awareness of change is needed management should discuss and convey the reasons and benefits of the change, both for the organization and the individual [33].
- The second element is desire – even if an employee is aware of a change that is needed, they may not feel a desire to adopt the change. To adress this the motivations for overcoming this persons threshold must be great enough [33].
- The third element is knowledge – while an employee have both awareness and desire for a change, they may not have the proper knowledge of how to implement or utilize it. Here is where education comes into play. Only after making sure

employees are aware of the need and understand the benefits of change should management start educating employees, as they might not be preconditioned otherwise [33].

- The fourth element is ability – the employee may now hold the knowledge of how a change can be made, the may not have the ability to put that knowledge into practice. This can be remedied by having the employee practicing the new system, with a hands-on coach. Both the right amount of practice time and ongoing coaching and support is required for new abilities to develop [33].
- The fifth and final element of the ADKAR model is reinforcement – In order for employees to fully utilize the new system or aspect of their work, there need be proper reinforcement. Positive recognition is a good way of reinforcing change, management should also make sure that change is incentivized, as this encourages adoptation of change [33].

# 3 Method

This chapter details the choice of method used as to fulfill the purpose of the thesis, and the motivations behind those choices.

## 3.1 Research design

### 3.1.1 Qualitative study

The study uses a qualitative research approach, as this can be preferred when detailing different subjective experiences, and how individuals interact with their surroundings. In qualitative studies the context of the research is of high importance, and the measurability of the data is dependent on the methods used. Hence, data usually consists of words rather than numbers [7]. As the purpose of the thesis is to study how LLMs can be introduced in construction processes and the social effects of such an integration, which are both heavily dependent on the views of those working within that process, and thus highly subjective and circumstantial – using qualitative interviews where these aspects can be explored were deemed most appropriate.

### 3.1.2 Interviews

Two sets of interviews were held. The first set were mostly unstructured, and the second semi-structured. These are detailed in sections 3.1.3 and 3.1.4 respectively. All interviews were conducted digitally as employees at Vattenfall usually works remotely, and gaining access to the companies facilities is not generally granted unless necessary. Interviews were recorded and transcribed at the interview participants permission, and will be deleted after the study is finished.

### 3.1.3 First set of interviews

To gain insight of the first research question, a series of interviews with engineers were firstly held. These interviews were aimed mainly at finding out how LLMs could be integrated into the construction process, which they were either currently working with or had worked with in the past. These interviews mostly likened unstructured interviews, where the interviewer has a few major areas of discussion prepared that the interview circles [7], but the interview is open to go in any direction that the discussion might take it. This was in order to not narrow down the scope of the second set of interviews before understanding the process and employees within it. The first set of interviews were thus kept fairly simple, circling three major areas. Those areas were:

- Their current pain points in the construction process.
- Tasks that they believed could be automated.
- Whether LLMs could be used to solve those problems.

Five of these first interviews were held.

### 3.1.4 Second set of interviews

The second set of interviews were conducted in a semi-structured manner, meaning that there existed a set list of questions that the interview participants were asked, but of which they are free to stray away from if the discussion demands it. Follow up questions can also be asked if necessary [7]. The questions were formed based on the discussions from the first set of interviews, along with an initial literature study. Seven of these interviews were held. For further explanation of the interview questions, see section 3.2.2.

## 3.2 Data collection

### 3.2.1 Choice of sample

In qualitative research the sample is usually selective, meaning that instead of randomizing samples in order to be representative of a larger population, they are actively chosen in order to be relevant to the research question. This means that the purpose of qualitative studies is not necessarily to be generalizable, but instead to gain a deeper understanding of behaviors, values and beliefs within studied context [7]. As the research questions of this study concerns construction processes, those that work within and adjacent to the process at Forsmark were chosen. All respondents are listed in Table 1.

For the first set of interviews, where the purpose was to identify whether and how LLMs could be integrated with the construction process, specialists and engineers that had worked with previous digitalizations and likely had some knowledge of AI were interviewed. This as they were deemed to have a higher likelihood of knowing how AI could be used to improve the rate of task completion within the relevant workflows. These interview participants are listed as R1-4 in Table 1. An extra instructured interview with R11 was held at the end of the project, discussing technical implementations and current platforms and systems at Vattenfall.

For the second set of interviews, where the purpose was mainly to propose the LLM and gain and understanding of how the LLM integrations could be received, engineers working directly within the construction process were targeted for interviews since they were working directly in the studied process. The goal of these interviews were to gain an understanding of the differing attitudes, experience and willingness to integrate AI elements. A few managers were also interviewed to provide an overview of the workflow process. Importantly, interviewees were chosen irregardless of their previous experiences and attitudes of AI, as to gain a wide range of perspectives in order to present how to move forward with LLM integrations. These interview participants are listed as R4-10 in Table 1. Note that R4 was interviewed two times, first through a general unstructured discussion, and later through a semi-structured interview. A second shorter discussion also occurred with R5, with the purpose of estimating values for the cost estimation in section 5.5.

All interviews were conducted over Microsoft Teams, and typically lasted from 40-60 minutes. In table 1 respondents are listed, along with their role, and a short description of their involvement. The order of interview participants is listed by the order of when the interviews were conducted.

Table 1: Overview of interview participants (anonymized).

| ID | Role | Primary involvement |
|---|---|---|
| R1 | Unit Manager | Operational Support at Ringhals |
| R2 | Digitalization Manager | Innovation within nuclear facilities |
| R3 | Head of the Safety Governance Unit | Radiation and data safety |
| R4 | Head of Systems Automation | Managing groups in the construction process |
| R5 | Technical Advisor | Previously worked in the construction process |
| R6 | Project Manager | Coordination, meetings, reporting |
| R7 | Engineer/Group Manager | Electric construction at Forsmark |
| R8 | Engineer | Electric construction at Forsmark |
| R9 | Engineer | Mechanical Engineer at Forsmark |
| R10 | Engineer | Mechanical Engineer at Forsmark |
| R11 | Project manager (Student Worker) | Developing an LLM for general use at Vattenfall |

### 3.2.2 Interview guide

Before the second set of interviews an interview guide was created (attached in appendix A), containing the questions that participants were asked. At the beginning of interviews a short presentation of the reason for the study was held. Questions were divided into two parts, with the first one (questions 1-10) inquiring into how the participant experiences work in the construction process and how an LLM could implemented. The second part (questions 11-17) focused on the change management aspects of such an implementation, based on the ADKAR model.

Interviews were conducted in Swedish, with questions both phrased and answered in Swedish. The interview guide, any quotes cited in the result or any ideas presented from interview participants have been translated to English.

## 3.3 Analysis

The interview material was analyzed using thematic analysis. Thematic analysis is a qualitative method for analyzing, organizing, and presenting themes from qualitative data. The method consists of identifying patterns within a dataset, *themes*, and is suitable when the aim is to connect different perspectives from interviews [5]. In this thesis thematic analysis was used to describe work in the construction process and use cases for LLM, derive requirements based on discussions with the interview participants, and formulate the change management study.

Braun and Clarke [5] describes thematic analysis to be made up of different phases. These were followed in this thesis. Firstly interviews were transcribed using a digital transcribing tool and read again to note down any ideas with the goal of writing down initial thoughts and semantics. Secondly, the transcripts were systematically coded based on their semantic meaning, by making comments in the transcribed text. Codes were applied to any bit of text that were in reference to work tasks, pain points in the construction process, opinions towards LLMs, and demands/requirements towards LLMs or likewise. In forming the second set of interviews, codes from the first set was used. When all interviews had been conducted, codes were then separated and compiled into themes. Phases four and five, where themes are reviewed and refined, were done concurrently. Finally, based on these themes the results section was written.

# 4 Results

This chapter details the main themes and responses from the interviews that has become apparent after analyzing them. How employees work within the construction process is explained in section 4.1, with 4.2 detailing identified use cases for LLMs, and requirements for an LLM. Then, the change management study is detailed in 4.3, with section 4.4 discussing skepticism towards AI. Section 4.5 discusses a few other improvements that might be of interest to Vattenfall.

## 4.1 Working in the construction process

Most respondents presented the same picture regarding projects within the construction process, that it is very common for projects to take a long time from start to finish. When asked why this is the case, a number of reasons were given. For some projects, changes to the systems such as the reactors can only be made a few times a year, for others, there might be disagreements or differences in how to approach a problem that could be time consuming. Overall though, the most common answer as to what was the most time-consuming and repetitive tasks were writing reports and reading various documents.

When writing reports there often needed to be multiple reports written, targeted for different stakeholders. For example, one report might be written for decision makers that are further away from the technology, requiring reports to be more general and explanatory of the contexts, while possibly keeping technical specifications vague or simplified as to keep them understandable for managers. A second report would then be written for engineers, which would be highly technical, and could omit some contexts as they were expected to be understood by the reader. While these two reports were different in tone and purpose, a large part of the content was highly similar, making some employees feel that some report-writing can be redundant. Respondent R7 estimated that his group spent about 50% of his time writing at the average project. A separate respondent (R9), that were not tasked with writing reports, estimated instead spending about half of their time handling documentation.

Writing reports, designing solutions, and making decisions in nuclear energy often come with a large amount of reading previous reports, documentation, and articles, which also can be time consuming. A common theme among all respondents were that Forsmark has access to an incredibly large amount of data gathered both locally and globally, to such an extent that it is not possible to read or process a majority of it manually. Finding the right documents even when directly looking for them could sometimes be hard due to the current search engines. Respondent R9 detailed how they sometimes, instead of searching for documents directly, would instead search for a previous employee that had been retired since ten years back, as that employee often would work similar projects. This would not be sustainable in the future as new personnel will not be aware of these retired employees and their projects, per the respondent.

R8, R2 and R5 all discussed how an LLM could be used as an educator for employees that are either new or in searching for knowledge. Often educational moments would happen through informal talks with senior experts. This was also a common occurrence among most interviews when discussing how respondents would go about finding information when faced with a problem. More often than not the first step was to ask a colleague that had expertize within an area, some of which had worked there since shortly after the construction of the plant. When these experts would later retire, all of their tacit knowledge would disappear with them, resulting in a loss of valuable knowledge. Even if all of their projects and written reports were digitized they are still hard to find through current systems, even if you knew what to search for.

## 4.2 Needs and opportunities for AI support

### 4.2.1 Use cases

In this section a few possible use cases of an LLM that became apparent during interviews are presented.

**Writing.** Firstly, a large amount of time within construction processes is being spent documenting and writing reports. Some respondents expressed that there were multiple reports regarding the same decision or specification, but to different departments. Furthermore, different writers have different writing styles, causing reports and specifications to vary depending on who wrote it. An opportunity to reduce time spent writing could be present here, by using an LLM to assist in writing. Depending on the type of document, the LLM could be more or less involved in the writing process. Multiple respondents expressed concern regarding the writer of a document not knowing or actually learning the contents of a text if written in assistance with AI.

**Summarization.** A commonality between interviews were that there exist to much data to be navigable. Often a respondent would want to access a small part of a large document, and would have to read or skim many pages before finding what they were looking for. Often, documents refer to other documents, meaning that engineers have to follow trails of documents before finding the actually wanted data. Sometimes engineers might want to compile data from different sources, meaning they might have to go through this process many times.

**Evaluation.** Most documents that are written within the construction process were also evaluated by a second person, either a colleague or someone from a different department. R5 mentioned during the interview that a report they had written recently had just been denied, due to some minor errors, such as spelling errors, or two different fonts being used. The respondent also mentioned that a majority of the time spent evaluating could be reduced by having an LLM evaluate first.

**Education.** Most of the respondents would discuss with and learn from colleagues early in the process of gaining knowledge of a form them previously unknown topic. This, coupled with the fact that finding information through the current databases could mean that an LLM could be useful for junior engineers in educating themselves, as argued by R8.

**Coding.** Some projects included segments were coding was needed. Respondents R7 and R9 had both used ChatGPT in the past to assist with coding, though kept it general as to not break confidentiality. The sentiment that they would be able to use an LLM that could write code for their precise needs were welcomed.

**Translation.** R7 had also used ChatGPT to translate when communicating with foreign stakeholders. They felt that tools like Google Translate could not capture the meaning of their messages, so AI was a better choice for them. The same problem as in the coding use case regarding having to be general was present here.

**Search Engine.** As presented earlier, all respondents that handled documentation felt that they spent excessive time both trying to find documents that they knew were relevant to their current project, and discovering additional documents that could be relevant to their project. Vattenfall has attempted to improve the search engine by implementing a ”Super-search” function, but the same problems still exists to an extent, per R5 and R11. An improved search engine could therefore improve information gathering, possibly substantially.

### 4.2.2 Requirements

In order to design an LLM-system, some technical requirements for the LLM could be observed from the interviews.

- Confidentiality – The LLM can never have access to documents above the users security clearance, as to not show data to unauthorized personnel. Furthermore some data can not leave Forsmarks premises, meaning that the LLM needs to be hosted locally on premise.

- Traceability – If some part of a construction proves to be wrong, there needs to be a trail to find where the error occurred. The LLM therefore needs to have all of its queries saved for future reference in case of it recommending something that causes problems.
- Accuracy – The LLM needs to have a large degree of accuracy in its answers, partly from the technical standpoint to be usable at all, and partly from a more human standpoint, for users to be able to trust it.
- Consistency – The LLM needs to be consistent in its answers among different queries, as not to give different solutions to different users regarding the same problem.
- Referentiality – The LLM needs to be clear regarding where presented information is sourced from, and preferably link to sourced documents.
- Integrable – The LLM needs to be able to be able to integrate with already existing databases, which may have different structures.
- Extractive summarization – When summarizing text from documents the same words and sentence formatting need to be used, not synonyms or newly generated sentences.
- Fairly affordable – In order to motivate the investment, the LLM should be affordable enough that it either saves costs on salaries spent on employees spending unnecessary time searching, reading and writing, or increases the quality, employee satisfaction, or other metric such that the costs are motivated.

## 4.3 Change management study

The purpose of the change management study was to inquire how susceptible the users would be to involve an LLM in their workflow. A large variety of perspectives and attitudes were observed towards AI. To present the varying degrees of change resistance and AI acceptance The ADKAR model was used.

Regarding *Awareness* of the benefits that LLMs could contribute with, most were aware with positively inclined attitudes. Though, one of the respondents (R4) that were not optimistically inclined, and were not convinced of the benefits. This person felt that there were more within their same department that were at the same level of awareness. In addition to this, most respondents knew a colleague that they believed were not aware.

Overall though, the majority were aware of the benefits that LLMs could bring, and deemed that most of their colleagues were the same. The colleagues that the respondents deemed had a lower level of awareness were often described to be of older ages, and could be perceived as ”set in their ways”. Respondent R7 described how some of these older employees had previously declared their worries when the entire documentation were to be digitalized, and that they might have similar worries towards AI.

Those respondents that *desired* change often coincided with those that were aware of it. Multiple respondents working in the construction process already used ChatGPT as assistance in their work, though due to confidentiality they could only use general prompts, not directly related to Forsmark's facilities. These employees described themselves as having a desire to implement LLM assistance in their work. There were also one respondent (R4) that were aware of the potential usefulness of an LLM, but were not convinced and did not necessarily desire to use it. This was due to something that was encountered multiple times throughout the interviews, even from those that were positively inclined towards the AI; namely skepticism towards it. Skepticism is further discussed in section 4.4. When asked all of these skeptics agreed that they would feel more comfortable to use LLMs if they had a greater understanding of how they work.

A majority of the respondents felt that they could benefit from being educated in LLMs. While using an LLM was deemed to be fairly straight forward by most, a common theme was that a course in prompting could improve the effectiveness of using an LLM. The respondents that felt that they were critical to LLMs to some extents also felt that greater insight into the technology could improve the trust and understanding of AI, which could heighten the support of AI-technology. A two-part course, firstly detailing the structure, architecture and technology of the LLM, and secondly prompting techniques could be beneficial for a wide LLM integration.

Those who felt that they had the necessary knowledge to use an LLM also felt that they had the *ability*. LLMs were generally deemed to be easy to use, and some had used similar technologies in the past. R9 thought it to be *”fairly straight forward, and learn by doing”*. Having the knowledge to use an LLM translated into having the ability to do so, due to LLMs having a low barrier of entry.

When asked how a change like LLMs could be best implemented, both R5 and R9 thought that starting small would be best, with a group of people already interested. This in part to work out any flaws, but also in order to let those initial users be ”culture bearers” (R9), spreading the word of the new technology. R5 stated that there is generally a large degree of freedom in how work is done. Tools may exists, but it is up to each engineer to use or not use them as long as deadlines are met. *Reinforcement* of new technologies should then try to show the value of the tool, and how it might improve an employees workflow.

## 4.4 Skepticism of AI

There were a few different skepticisms towards the use of AI that would be beneficial to tackle if an integration were to be made. These could be divided into: Degree of LLM integration, skepticism towards accuracy, and general skepticism towards new technological functionalities. These are presented below.

Firstly, there were various degrees of skepticism against AI in the writing process. R8 shared a sentiment that was unanimous among most participants:

*”That’s what I’m afraid of with AI... That it becomes cut and paste.”*

While close to all respondents agreed that LLMs could be beneficial, they were divided on the degree of LLM involvement. Some like R5 were more open to LLMs writing a majority of reports, while those like R8 were wary of it leading to reduced understanding of the report, which could cause engineers to be less qualified in the long term, and miss any errors that the AI might write. This over reliance on AI could lead to them being integral to the workflows, which could have consequences if it were to cease functioning. If LLMs were to write something faulty, it could even cause a snowball effect if later reports are based upon these earlier faulty ones, leading to greater errors. The consensus among respondents were then that there needed to be clear guidelines as to how an LLM would be used if it were ever introduced.

This connects to the second point of skepticism, that regarded hallucinations and that the LLM might generate answers that can be incorrect. This could also be described more as an untrustworthiness towards that the AI actually delivers accurate answers often enough that it could be trusted. R5 stated that it *”It needs to be correct almost always, if it’s wrong half the time there’s no use.”* Meanwhile, if an AI were to be correct to the degree that engineers feel they can trust it, when it does hallucinate the error might go unnoticed due to it being trusted. An AI that is less accurate and therefore less trusted, might be safer to use, due to its responses being more examined by humans.

The last point of observed skepticism was not skepticism towards AI or LLMs specifically, but general skepticism towards new technological functionalities. None of the respondents expressed this skepticism themselves, but most spoke of colleagues that were skeptical of technological change. These were often described to be of older ages and had worked in the construction process for a long time, and had numerous previous experiences of where they felt technological changes had not been integrated well.

## 4.5 Other possible improvements and implementations of AI

In this section a few other observations made throughout the interviews are listed. These were either mentioned in passing by respondents, or had no clear relation to the research questions, but may still be of interest to Vattenfall as they are recommended directly from employees.

Some parts of Vattenfall operations us PLM-systems. PLM generally stands for Project Lifecycle Management, though in Vattenfalls case the P instead stands for Plant. R1 felt that the PLM-systems that were in use were very outdated, and that there are many available better options, some of which have smart functions to improve workflows between different departments.

A project manager (R6) that was interviewed explained that a large part of their day-to-day task consisted of organizing meetings. This was comprised primarily of sending emails, and finding meetings slots that were available to all attendants.

R10 mentioned that some older documents are uploaded as scanned pictures rather than text-readable pdf, meaning that they are not readable by an embedding model or LLM.

# 5 Discussion

This study suggests that an integration of LLM support could potentially increase efficiency in some of the most prominent tasks in the construction process such as reading, searching, and writing documents. It also finds that such an integration is likely possible, but should be viewed as more of a socio-technical change than a mere technical one. While the technicalities of constructing an LLM is a large part of such an integration, the right social culture might need to be in place for it to be successful. This section aims to interpret the result with regard to the research questions.

## 5.1 RQ1: Can LLMs be introduced in construction processes in order to streamline project management workflows?

The result supports that an integration of an LLM could be beneficial in reducing time spent on supportive activities, to increase the time spent on producing solutions. The interviews indicated that a substantial part of the engineers' time were spent on more menial tasks, such as locating documentation, cross-referencing, or changing formulations based on the intended reader. While task like these support the current engineering process and the knowledge work, they often constitute documentation overhead rather than the core value creating activities that provide solutions to the tasks at hand, and could impede productivity, as described by Drucker [14]. Hence, using an LLM to do said overhead tasks could free up time for engineers in the construction process to focus on solving problems. As some respondents already were using LLMs in their work for general purposes it would seem like the perceived usefulness exists.

Though, the results indicate that some functionally easier actions could be taken to streamline the construction process. One of the often discussed pain points were the current search engines, which were deemed to be, by most respondents, tedious and hard to navigate. LLMs with RAG usually use a vectorized database, that store data with their semantics, i.e. their meaning. This means that a search query is matched by the meaning of the content, instead of keywords, which could ease the process of finding the right data without knowing exact titles or terminologies. Although it might make it more intuitive, semantic search does not actually need an LLM, meaning a technically easier first step could be to create a vector database with semantic search from all current databases and introduce it to engineers, and then expand it with an LLM system. This would also be simpler to validate and govern, as none of the output would be generated, instead it would be retrieved from real sources, which also aligns with the requirement of traceability.

## 5.2 RQ2: Which LLM-based approach and system components can meet the requirements identified in construction-related project workflows, and what are the main cost drivers and feasibility trade-offs of those features?

This section mainly discusses the requirements part of the research question, with trade-offs and cost estimation being discussed under sections 5.4 and 5.5.

Based on the requirements identified in section 4.2.2, an interpretation of what an LLM system needs to support becomes apparent. The requirements are strict enough that a general purpose LLM is not suitable at Forsmark unless it adheres to the identified conditions. The requirements can be interpreted as guidelines for designing the architecture of the LLM system, with the following elements and components being implied. Based on this, how the architecture of such a system might look is discussed in section 5.4, and the cost estimation is calculated in section 5.5.

**On-premise hosting**. Based on the requirement of confidentiality along with the requirement of integration the LLM-system should not use publicly hosted API:s as this would mean that confidential information could leave the Forsmark facilities, which would not be in terms of their security policies. Hosting the LLM locally is then a more applicable choice. This also brings the benefit of having increased control of the system configurations and monitoring, which could reduce risks from outside interference.

**Access control.** Confidentiality requires more than local hosting, as the system must ensure that any person using it does not access information above their clearance level. Either this is enforced by only allowing those with high enough clearance level to use the system, or by controlling the clearance level per employee on every request, and only providing permissible data. This would in turn require identity verification, And metadata of document clearance level.

**RAG with vectorized database.** The requirements of accuracy, referentiality and integrable with current databases indicate that a system such as RAG, which does not solely rely on the LLMs internal parameters, is in need. As RAG can improve accuracy [48] this can both increase trust, and minimize the risk of hallucination. This in turn implies data collection pipelines, which collects and then chunks documents, that are then embedded into vectors and stored in a vector database.

**Referential output.** In order to satisfy the requirement of referentiality the system should always cite its sources, and preferably link directly to the retrieved document, and the part of the text which has been retrieved.

**Query and output logging.** Traceability requires that all queries and outputs along with relevant data is logged, in case of later audits.

**Extractive summarization.** There were a need for summaries to preserve the exact wording of quoted text. Though, for education purposes, there might be situations were abstractive summarization (where the meaning is retained, but not wording) is more apt. Hence, an extractive mode that is toggleable could be a solution.

**OCR scanning.** As some documents were not text readable, they require OCR before they can be embedded. Without this the data would likely be incomplete.

**Affordable models and systems.** In order to motivate the investment of building an LLM on-site, the cost of operation must not be larger than the value it provides. Larger models might not be the optimal choice, as inference computation (directly correlated with GPU time) increases with model size [37]. Though, the capabilities of a high parameter LLM might still be needed for some more complex queries. Thus, an LLM cascade might be a good option for keeping costs low, while still providing a service of high quality.

## 5.3 RQ3: What social effects could such an integration have on organizational structure and company culture?

The change management study indicate that the reception of an LLM integration will depend on how it is introduced. The general attitude towards LLMs was positive, with only R4 being directly unconvinced by AI abilities. This could suggest that an AI adoption will likely be welcomed by the majority of its intended users, if their concerns regarding accuracy and to what degree it is safe to use AI have been addressed. It then becomes of importance that guidelines regarding the use of AI is developed and distributed to users, partly for the social aspects, but perhaps mostly for the safety aspect. There was however also general skepticism towards change, described to be mainly from older employees. If this group were to be negatively inclined towards an AI adoption while others are not, this could cause divides regarding what is considered proper or ”real” work, as it could be deemed less trustworthy. This lowering of standards were a common concern, that instead of supporting engineers they would be substituted.

The results also indicated that some part of the workforce might not be aware of the potential of LLMs, or might be critical to change in general. To combat this, the benefits and motivations should be communicated to those individuals, perhaps through an initial lecture. Motivations could be reduced time spent on overhead tasks, which could increase the individuals performance and time for other, perhaps more satisfactory tasks. If employees have never interacted with an LLM before, they could be given the opportunity to do so, to see the potential first hand. As these employees are less likely to be enthusiastic towards AI, they might also be unenthusiastic towards lectures about it. Thus, making sure that these employees are reached with this information is important for them to gain awareness, especially in departments that may have a large number of unaware employees where awareness might not spread from person to person as extensively as other departments.

The group of employees that desired the implementation could be utilized to spread the potential benefits of LLM support. If they were allowed to take part in a pilot test of the system, and found it to be useful, they might share this sentiment with their colleagues, causing acceptance of the technology to spread organically.

The ADKAR study further suggests that acceptance toward LLMs correlated with perceived understanding of the technology. As there were fears regarding how accurate AI is, and how to manage a tool that might hallucinate – education of the architecture, technology behind LLMs, and how to use it effectively might not only be something that could ease an AI adoption, but something that an successful integration hinges on. Initially testing the technology on a few enthusiastic early adopters, before systematically rolling out the technology along with training an clear guidelines could be a way to reduce resistance and skepticism.

Assuming an LLM adoption is done technically well, the likely cultural risk is not rejection of the tool, but a shift and uncertainty in verification and documentation responsibility. A roll-out that is centered around evidence-first retrieval alongside training and guidelines could possibly reduce the risk for overreliance on the technology while giving a chance for all different groups of AI-readiness to embrace the change.

Based on the results and above discussion, an implementation plan might read as follow:

1. Initial lecture detailing benefits and use cases of LLMs to increase awareness of its potential.

2. Pilot test, for improving usability, measuring contributions of LLMs in knowledge work, and spreading awareness and acceptance organically through ”culture bearers”.

3. More in depth education regarding prompt engineering, how AI functionally works, and guidelines for usage.

4. Incremental roll-out to different departments as to not overload the system and create negative feelings toward AI.

## 5.4 Proposed technical approach

This section proposes a technical approach that satisfies the implications discussed in section 5.2, based on the requirements identified in section 4.2.2. The proposed system is centered around being evidence-first, secure, accurate, and cost-aware. Scalability was also taken into account. The system consists of an on-premise 3-tiered LLM-cascade, utilizing RAG.

This section is divided in two parts, with the first providing a description of a high level description the proposed system, and the second detailing some practical recommendations for choosing models.

### 5.4.1 System overview

At a high level the system consists of technical layers, listed below.

1. **Ingestion & preparation layer**, that collects documents and metadata and prepares them for retrieval.

2. **Retrieval layer**, performs semantic retrieval over a vector database, and re-rank for higher precision.

3. **Answer layer**, routes requests to the best fitting model upon generation, and generates answers.

4. **Governance & logging layer**, conducts identity checks, and logs queries and answers for traceability and reproducibility.

The layers are further described in the sections below.

### 5.4.2 Ingestion & preparation layer

For each database that is connected, a program that connects and transfers it to the LLM system is needed, this transfer of data needs to be done routinely, as documents can be updated and new ones are often added. In order for the LLM system to access all possible data, all documents need to be machine-readable, meaning that documents that are scanned needs to be converted through an OCR program. Document from different databases may also have differently styled metadata. For example, one might have the date of creator in the format of YYYY/MM/DD, while another might have it as DD/MM/YYYY. When collecting these documents, their metadata needs to be normalized so that they are uniform. When documents are readable, and metadata is uniform, the documents are be split into chunks for later embedding.

### 5.4.3 Retrieval layer

In the retrieval layer, when creating a vectorized database, all chunks are embedded into a vector representation by an embedding model, and stored in the database. Due to the requirement of confidentiality, with the information needing to stay at Forsmark, the vectorized database also needs to be stored locally. To improve precision, a re-ranking program should be applied when comparing candidate answers to queries.

If employees of different clearance levels will be able to access the system (which might be wise if it would be implemented copied to other parts of Vattenfall operations down the line), identification should be at least before retrieval of higher clearance documents, but preferably before accessing the program at all, as to not allow unauthorized parties access.

### 5.4.4 Answer layer

In order to provide customizable functionalities, and keep costs low, a 3-tiered system likened to an LLM cascade, with toggleable modes is proposed. It would utilize two different LLMs, one smaller and one larger.

The first mode consists of doing a bare-bones semantic search, without any LLM involved. This would function as an improved search mode, utilizing the vectorized database for semantic search. This would render outputs such as best ranked documents to the search, without any summarizations or generated text.

The second mode would be an extractive summary mode, that would extract paragraphs and sentences in their original wording, by the use of a smaller LLM. The LLM could give some connecting commentaries based on user preference, but the emphasis is on keeping paraphrasing as low as possible.

The third mode would be like that of a more typical LLM, where a query is made and the LLM answers. If the query regards something related to Forsmark operations it should prioritize generating answers using RAG, rather than its internalized parametric knowledge. Importantly, if there does not exist enough data to warrant an answer, the LLM should be clear regarding this, and not try to create one. Factuality should take precedence over user satisfaction.

When designing an LLM-assistance system for practical use, cost is inherently part of the equation. In order to provide a service that is of a high standard an LLM with a high number of parameters should be used for more computationally complex prompts, but if this high parameter model is the only one used, it will be unnecessarily costly as it has higher requirements, that will process more computations. Instead, when simpler computable prompts are generated they should be routed to a model with fewer parameters, that require fewer computations. Depending on the complexity of the query, different LLMs would automatically be chosen by a regression model, and then a satisfactory answer would be generated. For example, if a simple general knowledge question would be asked, a smaller LLM is likely to be enough to provide an answer, while a question regarding writing code might need a larger model.

The three tiers of the LLM-cascade is thus as follows:

- **Tier 1**, Retrieval only, no LLM involvement. Used for more accurate search functions.
- **Tier 2**, Smaller, cheaper LLM, used to answer most queries.
- **Tier 3**, Larger, more costly LLM, used for complex queries.

### 5.4.5 Governance & logging layer

As discussed, traceability was an important requirement, as it is integral to the proposed design. Thus, each interaction and all data such as query text, retrieved documents, final output, timestamps, and LLM model should be logged and saved as this enables the ability to investigate any non-factual texts that the LLMs might generate down the line.

### 5.4.6 Choice of models

This section aims to provide some more concrete recommendations regarding which models might be most suitable to the specified requirements.

There are currently many different models available online. One of the major limiting factors in choosing an LLM to be used at Forsmark is the need for the model to run locally on premise, if it is to access documents of high confidentiality. This means that commonly known cloud based models like ChatGPT, Gemini, or Claud etc is not suited due to their current policys. These are the type of models that usually have the highest performance. Instead open source models that can be run locally are of interest. As they are open source, this also provide the benefit of being transparent as the back-end of the model is more open, and can be more easily modified to satisfy specific requirements if needed.

As there are a large number of models available, picking which one to use is not an exact science. Furthermore, new models are releasing rapidly, so whichever one is chosen to be further developed upon will likely be outperformed by models that are released in the near future. Nevertheless, as a vast majority of Forsmark's documentation is in Swedish (often highly technical, formal, and academical), and most employees are Swedish, using models that are highly proficient in the language was chosen for this system. Using EuroEval, an LLM benchmarking tool for European languages [30], the top open-source with the highest understanding of Swedish can be listed [17]. The top four which were deemed to be of interest are (in order of highest to lower general Swedish understanding):

- cogito-v1-preview-llama-70B
- Mistral-Small-3.1-24B-Instruct-2503
- cogito-v2-preview-llama-70B
- Llama-3.3-70B-Instruct

The top performing model, along with the third, are Deep Cogito models. These are models using Llama as a base and is then fine-tuned by the Deep Cogito team, based out of San Francisco. The llama-based models can "self-reflect" which allows for the model to have chain-of-thought prompting [11, 12], that can increase the performance of the answers [46]. While version 1 of the Cogito model has a better general knowledge of Swedish, there might still be reasons to choose version 2. Although cogito-v1-preview-llama-70B had the greatest performance in Swedish, differences in performance among all of the presented models are likely to be minute, possibly unnoticeable. Version 2 however have been claimed to have improved thinking mode, and perform better in multiple general and math benchmarks [11, 12]. Based on this, the cogito-v2-preview-llama-70B model could be a good choice for more complex prompts. While the deep Cogito models seem like good choices, the far more popular choice is Llama-3.3-70B-Instruct, a further developed model of the model that the Deep Cogito models are based on. The Cogito models both have about 250 downloads on huggingface in the last month, meanwhile the Llama model has 400,000 downloads (as of December 2025) on the same site. The low usage of Deep Cogito models may mean that eventual problems with the model is may not have solutions online, whereas the larger usage of the Llama model likely leads to any problems with it having solutions. The Llama model performs slightly lower on certain benchmarks like MATH and MMLU [27], in part due to not having an explicit ”think-mode”, as it has not been fine-tuned to the same extent. The llama model could then be seen as a safer choice, but with the trade-off of having lower performance. The Llama model is developed by Meta, with the Deep Cogito team being based out of San Francisco, meaning that both of these models are heavily US based LLMs.

Both of the Deep Cogito and Llama models are 70 Billion parameters large, and as previously mentioned, using a model with 70 billion parameters can be costly and often unnecessary for most prompts. Therefore, lower parameter models can be used to reduce cost of computation. For this, a model like Mistral-Small-3.1-24B-Instruct-2503, with a size of 24 billion parameters, can be used. Mistral is the eight highest rated LLM on EuroEval Swedish [17], despite its comparatively small size. Thus, Mistral-Small-3.1-24B-Instruct-2503 could be a good choice for a smaller LLM.

In order for these LLMs to provide accurate information from Forsmark’s documentation, The use of RAG is proposed. To embed data an embedding model is used, a suitable candidate is BGE-M3, which is a very popular embedding model, with over 7,300,000 downloads in last month (December 2025) [2]. It supports over 100 languages [2, 8], and holds a high place on embedding leaderboard MTEB [28, 29]. BGE-M3 was selected based on the ability to to be run locally on-premise without any API calls, with good Swedish capabilities. To further improve accuracy, it can also be paired with re-ranker bge-reranker-v2-m3 that is trained on BGE-M3 [4, 3] by the same team.

## 5.5 Cost estimation

In order to understand the feasibility of the proposed LLM, a cost estimation has been done. Due to it being based on multiple assumptions, the choice was made to call it an cost estimation rather than a cost analysis. Pricing of hardware can vary depending on procurement agreements and factors that are not fully known before deployment. Thus, this chapter aims to identify the major cost drivers and approximate the magnitude of annual costs.

Some of the following assumptions were assumed in collaboration with respondent R5 during a short discussion after their interview, and may not reflect the real values. Still, it could be a good assessment of the approximate cost of building the proposed system.

**Initial assumptions:**

- 500 people total will be using the the system at Forsmark.

- 20% of them, i.e. 100 might use the system any given day.
- These 100 will send 10 queries each per day, resulting in 1000 total queries.
- The cost of each employee (salary, employer's fee and pension) is 900,000 SEK per year, 75,000 SEK per month.
- The split between the three tiers is 50/40/10, i.e. 50% of traffic going to semantic search only, 40% going to the smaller LLM, and 10% going to the larger.
- Forsmark would have to purchase the necessary hardware to run the LLM but have the necessary facilities to store it, and the cost of nvidia L40S is around 100,000 SEK/unit.
- After four years, the system might need to be upgraded or require new hardware. Depreciation time is thus four years.

### 5.5.1 Current costs

Assuming each employee costs 900,000 SEK/year, the total cost of labor can be calculated as:

$$900,000 * 500 = 450,000,000 \text{ SEK/year}$$

LLMs have been shown to potentially increase the productivity of users, with the users completing tasks up to 25% quicker [13]. This could mean that utilizing AI tools could yield an extra 25% more deliverables per employee, and could reduce project time by 25% in certain phases. Under conservative assumptions, a third of that value is used for the calculations, thus AI could be assumed to reduce time spent on some tasks by 8%. As respondents estimated that half of their time was spent handling documentation, 0,5*0,08 yields 4% of total time within the construction process could be saved. This results in a final cost saved:

$$450,000,000 * 0,04 = 18,000,000 \text{ SEK/year}$$

### 5.5.2 Construction cost

The cost of constructing the system is estimated using a person-month approach. Due to the large amount of unknown variables, such as document and metadata quality, chunk sizes, database structures, etc, any estimated time plan would be highly fictional. Instead a contingency-heavy assumption of 40 person-months is used. This makes the total cost of construction

$$75,000 * 40 = 3,000,000 \text{ SEK}$$

Depreciated over four years yields a cost of 750,000 SEK/year.

### 5.5.3 CAPEX/Hardware costs & possible value

One of the major upfront costs is the cost of purchasing hardware, and one of the major cost drivers are GPUs. Based on Nvidias own tests, Llama 3.1 70B Instruct (earlier version of the proposed version 3.3) can be optimally run on eight nvidia L40S GPUs, and models in the 8 - 70 billion parameters range (the proposed mistral model having 24B) running optimally on 2 [32]. Under the assumption that L40S costs 100,000 SEK, this gives a total cost of 1,000,000 SEK. The number of GPUs could be scaled up to lower latency during more active times. It could also be scaled down, for example for a pilot study, though this would increase latency times (and depending on the magnitude of the scale-down require quantization of the model which would affect performance).

Another large cost is the server platform (CPUs, chassis, RAM, etc) which for 8 GPUs can cost upwards of 200,000-3,000,000 SEK depending on technical needs [6]. As this is a comparatively small system, while still being conservative, a cost of 1,000,000 SEK is assumed to encompass both the server platform all smaller hardware cost. This totals the CAPEX costs at 2,000,000 SEK. If depreciated over four years, the annual CAPEX is 400,000 SEK/year.

### 5.5.4 OPEX/Operational costs

Staffing is another major cost. Once again, to maintain safe margins the equivalent of one full time employee is assumed to to be needed for upkeep, for a cost of 900,000 SEK yearly.

Energy is also a cost driver. The L40S GPUs can consume max 350W per unit [31], for a total for 3,5 kW for 10 GPUs. Using a high estimate of 1,5kW for the rest of the system such as CPUs, storage, networking and cooling yields a total of 5 kW. The average datacenter in the EU have a power utilisation effectiveness of 1,6 [18], meaning that the total consumed power is 5*1,6=8 kW. For every hour of the year this yields

$$365 * 24 * 8 = 70,080 \text{ kWh.}$$

Using the average price of electricity in SE3 (zone of where the datacenter would be located) during 2025 (0,5517 SEK/kWh) [15] the total cost of electricity can be calculated:

$$0,5517 * 70,080 = 38,663 \approx 40,000 \text{ SEK}$$

A maintenance cost of 10% of hardware value is also added, as there may be need of upkeeping. This totaled at 200,000 SEK.

### 5.5.5 Total costs

Based on this the total costs of the identified major cost drivers have been calculated and presented in table 2.

| **Cost driver** | **Amount** |
|---|---|
| Construction | 750,000 |
| Hardware | 400,000 |
| Staff | 900,000 |
| Energy | 40,000 |
| Maintenance | 200,000 |
| **Total** | **2,290,000** |

Table 2: Total costs in SEK.

The total cost was calculated to be 2,290,000 SEK/year. Put into perspective, this is 2,290,000/5000=4580 SEK per employee. Under these assumptions the total profit per year from the LLM-system is :

$$18,000,000 - 2,290,000 = 15,710,000 \text{ SEK.}$$

Of course, the estimated cost is likely to be different from the actual cost. Still, under these assumptions the costs could be multiplied 7 times and the project would still be profitable, meaning that it satisfies the requirement of affordability.

## 5.6 Final remarks

As the study is qualitative, the results and derived discussion is largely contextual. Still there might be room for generalization. For instance, the reducing of time spent on tasks such as searching for documents or writing slightly different stakeholder-specific documentation may be something that can be beneficial in workplaces guided by strict security and confidentiality policies. Similarly, the identified requirements are likely to be common in such environments. The proposed approach (with semantic retrieval, and gradually introduced LLM capabilities) and the high level model might then serve as a generally applicable implementation, or a start-of point for designing systems in similar cases the need for an LLM is previously identified.

# 6 Conclusions

In this thesis, the potential implementation of a LLM-system at Forsmarks construction process in order to improve productivity has been studied. This was done through a qualitative interview study based on theory regarding LLMs and their limits, retrieval-augmented generation (RAG) and change management. The result indicated that LLMs could viably be implemented, under the right conditions. Based on this a technical approach was proposed to satisfy these requirements, and an order-of-magnitude cost estimation was done.

## 6.1 Main findings

One of the main takeaways is that LLMs can be useful in the studied context, but only under strict conditions. Tasks that were identified that possibly could be assisted by a LLM were those that are classically associated with knowledge work such as writing reports and finding and compiling data. Multiple requirements for the system were identified, that construction of a LLM-system would need to adhere to for the implementation to viable, technically useful, and culturally adopted. These requirements were: on-premise deployment, access control, consistent generation, have clear referentials, work within current digital systems, extractive summarization modes, and be able to motivate the investment. Based on this a high-level architecture of the systems design was proposed, with three different modes depending on the use case and complexity of prompts. A few possible choices regarding which models to use was also recommended.

The results also indicate that the system should not be regarded solely as a technical one, but a socio-technical one as how the system is received by their intended user is a deciding factor of its effectivity. The results found that there were three points of skepticism towards AI, these were the fear of becoming too reliant upon it, fear of accuracy, and general fear of technological change. The integration of the LLM should then be made with these fears in mind, and proper education should be administered, in order to make the transition more successful.

## 6.2 Future Work

To base the findings in data, and move on from only feasibility, future work could focus on actual implementation. This could be done through a pilot study, where a retrieval only semantic system is initially built and evaluated for time savings, user satisfaction and success rate. The governance, ownership and policies regarding data handling could be examined in connection to this to determine how the system fits into current processes. This pilot study could also examine how implementation would affect social aspects in reality, as there might be a divide between employees thoughts, and their actions. This study could then be used as a basis for fully developing a LLM-system.

Further work could also target more precise cost modelings, based on measured data. This could then be used to more accurately calculate costs for further development of LLM-systems.

If Forsmark, or any functionally similar facility, were to build or implement LLMs later in time, they might want to reconsider the recommended models and proposed technical solutions as better alternatives are likely to be developed in the future.

# References


[1] O. E. Ademola, “Change management trends in the ai modern world: Adapting to the future of work,” *Journal of Behavioural Informatics, Digital Humanities and Development Research*, vol. 10, no. 1, pp. 41–50, 2024, hämtad 2025-11-11. [Online]. Available: https://www.researchgate.net/publication/379449370_Change_Management_Trends_in_the_AI_Modern_World_Adapting_to_the_Future_of_Work

[2] Beijing Academy of Artificial Intelligence (BAAI), “Bge-m3: Multi-lingual, multi-functionality, multi-granularity text embeddings,” https://huggingface.co/BAAI/bge-m3, 2024, model card on Hugging Face, accessed 2026-01-07.

[3] Beijing Academy of Artificial Intelligence (BAAI). (2024) Bge-reranker-v2 documentation. Accessed 2026-01-07. [Online]. Available: https://bge-model.com/bge/bge_reranker_v2.html

[4] Beijing Academy of Artificial Intelligence (BAAI). (2024) Bge-reranker-v2-m3: Multilingual cross-encoder reranker. Model card on Hugging Face, accessed 2026-01-07. [Online]. Available: https://huggingface.co/BAAI/bge-reranker-v2-m3

[5] V. Braun and V. Clarke, “Using thematic analysis in psychology,” *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77–101, 2006.

[6] Broadberry Data Systems, “Supermicro gpu superserver 821ge-tnhr (sys-821ge- tnhr),” https://www.broadberry.co.uk/gpu-based-supermicro-servers/sys-821ge-tnhr, 2025, product page, accessed 2026-01-09.

[7] A. Bryman, E. Bell, B. Harley, and L. Holmén, *Företagsekonomiska forskningsmetoder*, 4th ed. Liber, 2024.

[8] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, “Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation,” 2024. [Online]. Available: https://arxiv.org/abs/2402.03216

[9] L. Chen, M. Zaharia, and J. Zou, “Frugalgpt: How to use large language models while reducing cost and improving performance,” *arXiv preprint arXiv:2305.05176*, 2023. [Online]. Available: https://arxiv.org/abs/2305.05176

[10] M. Chui, E. Hazan, R. Roberts, A. Singla, K. Smaje, A. Sukharevsky, L. Yee, and R. Zemmel, “The economic potential of generative ai: The next productivity frontier,” Jun. 2023. [Online]. Available: https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier

[11] Deep Cogito, "Cogito v1 preview – llama 70b," https://huggingface.co/deepcogito/cogito-v1-preview-llama-70B, 2025, model card on Hugging Face, accessed 2025-12-16.

[12] Deep Cogito, "Cogito v2 preview – llama 70b," https://huggingface.co/deepcogito/cogito-v2-preview-llama-70B, 2025, model card on Hugging Face, updated 2025-07-31, accessed 2025-12-16.

[13] F. Dell'Acqua, E. McFowland III, E. Mollick, H. Lifshitz-Assaf, K. C. Kellogg, S. Rajendran, L. Krayer, F. Candelon, and K. R. Lakhani, "Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality," Harvard Business School, Working Paper 24-013, Sep. 2023, harvard Business School Technology & Operations Mgt. Unit Working Paper No. 24-013. [Online]. Available: https://www.hbs.edu/ris/Publication%20Files/24-013_d9b45b68-9e74-42d6-a1c6-c72fb70c7282.pdf

[14] P. F. Drucker, "Knowledge-worker productivity: The biggest challenge," *California Management Review*, vol. 41, no. 2, pp. 79–94, 1999.

[15] Elbruk.se, "Electricity price history 2025," https://www.elbruk.se/elpris-historik-2025, 2025, accessed 2026-01-09.

[16] Energimyndigheten, "Myndighetsgemensam uppföljning av samhällets elektrifiering, Huvudrapport 2024," Statens energimyndighet, Energimyndigheten, Box 310, 631 04 Eskilstuna, Rapport ER 2025:03, 2025, hämtad 2025-10-29. [Online]. Available: https://energimyndigheten.a-w2m.se/System/TemplateView.aspx?p=Arkitektkopia&id=04b90e827bd84cdea37ecdfabe133f93&q=myndighetsgemensam&lstqty=1

[17] EuroEval, "Euroeval monolingual swedish leaderboard," https://euroeval.com/leaderboards/Monolingual/swedish/, 2025, accessed 2025-12-13.

[18] European Commission, Joint Research Centre, "Eu code of conduct for data centres: towards more innovative, sustainable and secure data centre facilities," https://joint-research-centre.ec.europa.eu/jrc-news-and-updates/eu-code-conduct-data-centres-towards-more-innovative-sustainable-and-secure-data-centre-facilities-2023-09-05_en, 2023, accessed 2026-01-09.

[19] I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, and N. K. Ahmed, "Bias and fairness in large language models: A survey," *arXiv preprint arXiv:2309.00770*, 2023. [Online]. Available: https://arxiv.org/abs/2309.00770

[20] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, and H. Wang, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint*, 2023, arXiv:2312.10997. [Online]. Available: https://simg.baai.ac.cn/paperfile/25a43194-c74c-4cd3-b60f-0a1f27f8b8af.pdf

[21] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Transactions on Information Systems*, vol. 43, no. 2, pp. 1–55, 2025.

[22] International Energy Agency, "Energy Policy Review: Sweden 2024," International Energy Agency, Tech. Rep., 2024, hämtad 2025-10-29. [Online]. Available: https://iea.blob.core.windows.net/assets/b80f421d-1e1c-4c73-bea4-acef5e60b3dd/Sweden2024.pdf

[23] International Energy Agency, "World Energy Employment 2024," Paris, 2024, Accessed 2025-11-08. [Online]. Available: https://iea.blob.core.windows.net/assets/2be1e935-a2e5-4535-9524-6f2610efd23f/WorldEnergyEmployment2024.pdf

[24] S. Johnson. (2023) Swedish parliament passes new energy target, easing way for new nuclear power. Reuters. Accessed 2025-12-02. [Online]. Available: https://www.reuters.com/sustainability/climate-energy/swedish-parliament-passes-new-energy-target-easing-way-new-nuclear-power-2023-06-20/

[25] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020. [Online]. Available: https://arxiv.org/abs/2001.08361

[26] J. M. Lichtenberg, A. Buchholz, and P. Schwöbel, "Large language models as recommender systems: A study of popularity bias," *arXiv preprint arXiv:2406.01285*, 2024. [Online]. Available: https://arxiv.org/abs/2406.01285

[27] LLM Stats, "Llama 3.3 70b instruct: Pricing, context window, benchmarks, and more," https://llm-stats.com/models/llama-3.3-70b-instruct, 2025, accessed 2025-12-18.

[28] MTEB Team. (2025) Mteb leaderboard – massive text embedding benchmark. Accessed 2026-01-07. [Online]. Available: https://huggingface.co/spaces/mteb/leaderboard

[29] N. Muennighoff, N. Tazi, L. Magne, N. Reimers *et al.*, "Mteb: Massive text embedding benchmark," *arXiv preprint arXiv:2210.07316*, 2022. [Online]. Available: https://arxiv.org/abs/2210.07316

[30] D. S. Nielsen, “Scandeval: A benchmark for scandinavian natural language processing,” *arXiv preprint arXiv:2304.00906*, 2023. [Online]. Available: https://arxiv.org/abs/2304.00906

[31] NVIDIA Corporation, “Nvidia l40s gpu,” https://www.nvidia.com/en-us/data-center/l40s/, 2025, product page, accessed 2026-01-09.

[32] NVIDIA Corporation, “Support matrix — nvidia nim for large language models (version 1.1.0),” https://docs.nvidia.com/nim/large-language-models/1.1.0/support-matrix.html, 2025, accessed 2026-01-08.

[33] Prosci, Inc., “The prosci adkar model overview ebook,” Prosci, Inc., Fort Collins, CO, Tech. Rep., 2021, © Prosci, Inc. All rights reserved. Hämtad 2025-11-12. [Online]. Available: https://www.prosci.com

[34] P. Pyöriä, “The concept of knowledge work revisited,” *Journal of Knowledge Management*, vol. 9, no. 3, pp. 116–127, 2005.

[35] M. Raza, Z. Jahangir, M. B. Riaz, M. J. Saeed, and M. A. Sattar, “Industrial applications of large language models,” *Scientific Reports*, vol. 15, no. 1, p. 13755, 2025. [Online]. Available: https://www.nature.com/articles/s41598-025-98483-1

[36] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Pearson Educational Limited, 2022.

[37] N. Sardana, J. Portes, S. Doubov, and J. Frankle, “Beyond chinchilla-optimal: Accounting for inference in language model scaling laws,” 2024. [Online]. Available: https://raw.githubusercontent.com/mlresearch/v235/main/assets/sardana24a/sardana24a.pdf

[38] T. Thesing, C. Feldmann, and M. Burchardt, “Agile versus waterfall project management: Decision model for selecting the appropriate approach to a project,” *Procedia Computer Science*, vol. 181, pp. 746–756, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050921002702

[39] R. Todnem By, “Organisational change management: A critical review,” *Journal of Change Management*, vol. 5, no. 4, pp. 369–380, 2005.

[40] Vattenfall, “Vattenfall in brief,” https://group.vattenfall.com/about-us/vattenfall-in-brief, 2024, accessed: 2025-12-01.

[41] Vattenfall, “Vattenfall: Videberg kraft ansöker om statligt stöd för investering i ny kärnkraft,” Pressmeddelande (PDF), publicerat via Cision, Dec. 2025. [Online]. Available: https://mb.cision.com/Main/865/4286109/3857196.pdf

[42] Vattenfall AB, “Internal forsmark governing document,” confidential internal document, not publicly available.

[43] Vattenfall AB, “Annual and sustainability report 2023,” Vattenfall AB, Solna, Sweden, Tech. Rep., 2023, accessed 2025-12-02. [Online]. Available: https://group.vattenfall.com/globalassets/com/sustainability/vattenfall-annual-and-sustainability-report-2023.pdf

[44] Vattenfall AB, “Annual and sustainability report 2024,” Vattenfall AB, Solna, Sweden, Tech. Rep., 2024, accessed 2025-12-01. [Online]. Available: https://group.vattenfall.com/globalassets/com/sustainability/vattenfall-annual-and-sustainability-report-2024.pdf

[45] Vattenfall AB, “Our climate targets,” https://group.vattenfall.com/sustainability/climate-transition-plan/climate-targets, 2024, hämtad 2025-12-01.

[46] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou, “Chain-of-thought prompting elicits reasoning in large language models,” *arXiv preprint arXiv:2201.11903*, 2022. [Online]. Available: https://arxiv.org/abs/2201.11903

[47] S. Wu, Y. Xiong, Y. Cui, H. Wu, C. Chen, Y. Yuan, L. Huang, X. Liu, T.-W. Kuo, N. Guan, and C. J. Xue, “Retrieval-augmented generation for natural language processing: A survey,” *arXiv preprint arXiv:2407.13193*, 2024, version 3 (posted 1 Mar 2025). [Online]. Available: https://arxiv.org/abs/2407.13193

[48] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, F. Wu, and G. Wang, “Instruction tuning for large language models: A survey,” *arXiv preprint arXiv:2308.10792*, 2023. [Online]. Available: https://arxiv.org/abs/2308.10792

[49] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, and J.-R. Wen, “A survey of large language models,” *arXiv preprint arXiv:2303.18223*, 2023. [Online]. Available: https://arxiv.org/abs/2303.18223

# A Appendix

Interview guide, translated to english.

1. Explain the background for the thesis and purpose for the interview, along with my background.

*Hi! My name is Sebastian, and I'm currently writing my thesis at Uppsala University in the master's programme in sociotechnical systems engineering here at Vattenfall. The thesis is about integrating LLMs in the construction process att Forsmark. We've observed that some parts of the construction process could be improved or automated through building a LLM. This could, by being being connected to different databases reduce time spent searching, reading and writing documents. Ideally the system would also be able to do online searches to complement the information in the databases. I've prepared some questions regarding how you work in the construciton process, and how such an LLM-integration might affect you and your co-workers.*

2. What are your daily tasks? What does your day-today look like?

3. How do you think a LLM would affect your work?

4. Which parts and task of the construction process do you experience to be the most time consuming and repetitive?

5. When searching for technical information, how do you normally go at it?

6. Do you have any views of how such an LLM should be designed? Any specifications or functionalities that you feel would be important for your work?

7. What type of output would be most useful?

8. What proportion of your time do you spend writing and reading documents?

9. Do you use LLMs like ChatGPT, Gemini or Claude today?

10. Is there anything you absolutely do not want a LLM to do?

11. Do you feel you have enough knowledge today to start using a system like this?

12. How do you feel about implementing a system like this, or AI in general, in your day-to-day work?

13. How do you (and your co-workers) generally view using AI tools at work?

14. What do you think would be required for employees to feel comfortable starting to

use LLMs?

15. Do you have previous experience introducing new digital tools? How did it go? What went well and why?

16. What barriers do you think might exist to starting to use AI in projects (e.g., time, trust, training, etc.)?

17. How do you think this kind of change should best be implemented?