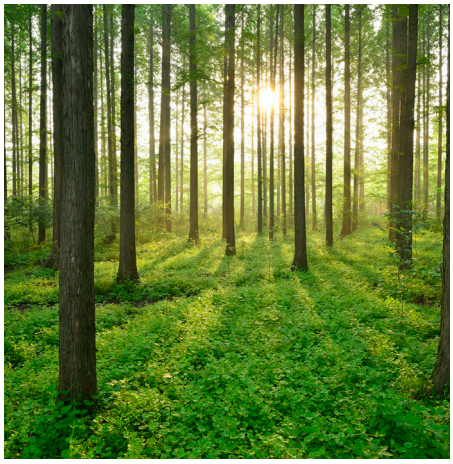
ON-PREMISE AI SOLUTIONS FOR NORDIC NUCLEAR APPLICATIONS

RAPPORT 2025:1093



DIGITALISATION
IN NUCLEAR POWER

Energiforsk

On-Premise AI Solutions for Nordic Nuclear Applications

Report by RISE Research Institutes of Sweden

LEON SÜTFELD
ANDREAS THORE

ISBN 978-91-89919-93-8 | © Energiforsk March 2025

Energiforsk AB | Phone: 08-677 25 30 | E-mail: kontakt@energiforsk.se | www.energiforsk.se

Foreword

Generative AI is revolutionizing industries by automating complex tasks, uncovering insights, and driving innovation. In the nuclear sector, it offers transformative potential—from predictive maintenance to enhanced safety analysis—addressing critical challenges while improving efficiency and reliability.

However, the adoption of generative AI in this field demands rigorous data security. On-premise solutions provide a robust answer, ensuring sensitive information remains within secure infrastructure. By avoiding reliance on external cloud platforms, these solutions align with stringent regulatory requirements and cybersecurity standards, enabling the safe and effective integration of AI into nuclear operations.

The study evaluates the needs and feasibility of on-premise AI solutions for the Nordic Nuclear Power Plants (NPPs). Examples of applications: LLM-based chatbots and computer vision. Additionally, it examines the legal, security, and connectivity constraints specific to the NPP domain and explores the technical and operational viability of tailored AI solutions, including hardware requirements and available options.

The project was executed and led by RISE Research Institutes of Sweden and conducted in collaboration with a reference group of stakeholders in the Nordic nuclear energy sector; Vattenfall, Fortum, OKG, and TVO. The project is part of the Digitalisation in Nuclear Power program and financed by Vattenfall, Uniper, Fortum, TVO, Skellefteå Kraft and Karlstads Energi.

Summary

This report explores the feasibility and requirements for implementing on-premise AI solutions in the Nordic nuclear energy sector.

The investigation focuses on natural language processing (NLP) and computer vision (CV) technologies, aiming to evaluate AI/ML-based systems for data analysis, prediction, and decision support, while addressing specific data handling and security restrictions inherent to nuclear power plants (NPPs). It provides an overview of the Nordic nuclear energy landscape, regulatory environment, and the technical challenges of deploying AI solutions within these constraints. The report highlights the potential applications of large language models (LLMs) for tasks such as document search, analysis, and summarization, and explores the use of retrieval-augmented generation (RAG) systems to enhance LLM performance. Additionally, it discusses the use of computer vision (CV) for monitoring, anomaly detection, and quality inspection tasks. Ongoing AI initiatives within the Nordic nuclear industry are reviewed, identifying current projects and future needs. A proposed pilot study aims to develop a proof-of-concept semantic search engine for large document collections, leveraging LLMs and RAG systems, while ensuring secure data handling and compliance with industry regulations.

Keywords

Artificial Intelligence, Machine Learning, Nuclear Power Plants, Nordic, Large Language Models, LLM, Retrieval-Augmented Generation, RAG, Computer Vision, CV, Monitoring, Anomaly Detection, Quality Inspection, Chat Bots, Document Search

Artificiell intelligens, Maskininlärning, Kärnkraftverk, Norden, Large Language Models, LLM, Retrieval-Augmented Generation, RAG, Computer Vision, CV, Övervakning, Anomalidetektering, Kvalitetsgranskning, Chat Bots, Dokumentsökning

Sammanfattning

Denna rapport undersöker genomförbarheten och kraven för att implementera lokala AI-lösningar i den nordiska kärnenergisektorn.

Undersökningen fokuserar på teknik för behandling av naturligt språk (NLP) och datorseende (CV), och syftar till att utvärdera AI/ML-baserade system för dataanalys, förutsägelser och beslutsstöd, samtidigt som man tar hänsyn till specifika datahanterings- och säkerhetsrestriktioner som gäller för kärnkraftverk (NPP). Rapporten ger en översikt över det nordiska kärnenergilandskapet, regelverket och de tekniska utmaningarna med att implementera AI-lösningar inom dessa begränsningar. Rapporten belyser de potentiella tillämpningarna av stora språkmodeller (LLM) för uppgifter som dokumentsökning, analys och sammanfattning, och utforskar användningen av RAG-system (retrieval-augmented generation) för att förbättra LLM-prestanda. Dessutom diskuteras användningen av datorseende (CV) för övervakning, anomalidetektering och kvalitetsinspektion. Pågående AI-initiativ inom den nordiska kärnkraftsindustrin granskas, och pågående projekt och framtida behov identifieras. En föreslagen pilotstudie syftar till att utveckla en proof-of-concept semantisk sökmotor för stora dokumentsamlingar, med hjälp av LLM- och RAG-system, samtidigt som man säkerställer säker datahantering och efterlevnad av branschregler.

List of content

1 Introduction

This report was produced as part of the On-Premise AI Solutions for Nordic Nuclear Applications (AI SNAP) project, led by RISE Research Institutes of Sweden and conducted in collaboration with a reference group of stakeholders in the Nordic nuclear energy sector; Vattenfall, Fortum, OKG, and TVO. The main objectives of the report are:

1. To assess the stakeholder's needs for on-premise AI solutions, in particular along the following two threads: (1) LLM-based chatbots for general use by Nordic Nuclear Power Plant (NPP) office personnel, (2) other AI/ML-based systems for data analysis, prediction, decision support, etc.
2. To assess the specific restrictions and limitations for AI solutions in the NPP domain (e.g., legal and security-related restrictions to data handling and online-connectivity during both development and deployment).
3. To explore the viability of on-premise AI solutions tailored to the NPPs needs. This includes an assessment of technical feasibility, computational/hardware requirements, expected outcomes, and a comparison of the available options where multiple exist.

The project's mission was thus to identify use cases and lay the groundwork for one or more follow-up projects, in which AI solutions are to be implemented. The follow-up project should target enhancements in operational efficiency, safety, and decision-making with AI, by aiding workers in an office setting. Safety-critical applications were not to be considered as per the project's objective agreement. As part of this project, a series of semi-structured online interviews was conducted with a total of 9 employees of the stakeholder organizations. Three online workshops were held with all stakeholders involved, and e-mail communication was used to address additional questions and topics. A preliminary version of this report was provided to the stakeholders for review prior to the project's end, and the final version was delivered on December 16, 2024.

This report is structured as follows: Section 2 gives a brief overview of the nuclear energy landscape in the Nordics. Section 3 lays out and addresses the specific limitations for the training and deployment of AI in the Nordic nuclear sector. Section 4 introduces various relevant methods and techniques in the current AI landscape and illustrates the state of the art in AI for the nuclear energy sector. Section 5 highlights ongoing and planned AI projects within the Nordic nuclear energy industry, and outlines the needs and interests discussed during the interviews. Section 6 outlines the plans for a pilot study on AI-based document search as a follow-up on the AI SNAP project. It discusses the requirements for and viability of such an AI application as an on-premise AI solution for the nuclear sector. Finally, this report is concluded in Section 7 with a brief summary and highlighting of the primary findings.

2 The Nordic Nuclear Energy Sector

The Nordic nuclear energy sector is comprised of five power plants – three in Sweden and two in Finland – operated by four corporations (all stakeholders in the AI SNAP project). The 11 reactors in these plants have a combined nameplate capacity of 11,354 MW. In 2022, power consumption in Sweden was 124 TWh, with an additional 50TWh generated and exported. The six Swedish reactors contributed around 30% (51.9TWh) to the total electricity produced[1]. In Finland the newly commissioned Olkiluoto 3 reactor helped to increase the percentage of nuclear energy in the electricity mix to 41% of the total consumption of 79.8 TWh in 2023[2], up from 30% in 2022[3]. An overview of the Nordic nuclear facilities is given in Table 1.

Power Plant	Licensee	Reactors	Capacity	Commission Dates	Country
Forsmark	Vattenfall	3	3320 MW	1980, 1981, 1985	Sweden
Ringhals	Vattenfall	2	2190 MW	1981, 1983	Sweden
Oskarshamn	OKG	1	1450 MW	1985	Sweden
Loviisa	Fortum	2	1014 MW	1977, 1981	Finland
Olkiluoto	TVO	3	3380 MW	1979, 1982, 2023	Finland

Table 1: Overview of power plants in the Nordics, 2024.

As for the future, Finland is committed to nuclear power as part of its long-term energy strategy, aimed at achieving carbon neutrality by 2035 and significantly reduced energy import dependence. The commitment to nuclear energy is also evident in the country's plans to open a nuclear waste disposal facility (Onkalo), expected to start operating in 2025[4]. Sweden, originally planning to phase out nuclear power by 2040, has in June 2023 reversed course by changing its energy target from "100% renewable" to "100% fossil-free" electricity by 2040, enabling a long-term future for nuclear energy production. This change comes with an announcement of plans to construct at least two large-scale reactors by 2035 and the equivalent of 10 new reactors, including small modular reactors, by 2045, alongside a number of regulatory and policy changes to facilitate the construction of new nuclear reactors[5][6]. The long-term commitment and planned growth of the nuclear sector in both Sweden and Finland create a stable environment for investments in AI solutions for the industry.

[1] https://world-nuclear.org/information-library/country-profiles/countries-o-s/sweden
[2] https://www.motiva.fi/en/solutions/energy_use_in_finland/electricity_supply_and_demand
[3] https://www.treasuryfinland.fi/investor-relations/sustainability-and-finnish-government-bonds/data-and-facts-energy-transition/
[4] https://www.iea.org/reports/finland-2023/executive-summary
[5] https://world-nuclear-news.org/Articles/Roadmap-launched-for-expansion-of-nuclear-energy-i
[6] https://world-nuclear.org/information-library/country-profiles/countries-o-s/sweden

3 Assessment of Specific Restrictions and Limitations

Due to its vital role in energy infrastructure, the inherent risks of radioactive materials, the potential for severe accidents, and its attractiveness as a target for malicious actors, the nuclear energy sector faces significant regulation. Regulatory oversight is conducted in Sweden by Strålsäkerhetsmyndigheten[7] (SSM) and in Finland by STUK (Säteilyturvakeskus)[8], while the relevant legislation is provided by the European Union (EU), as well as the Swedish and Finish states. Regulations and laws on the national and European level, however, are unspecific with regard to the technologies used in the context of nuclear power plants. Instead, it is the responsibility of the licensees to establish IT security teams or councils that evaluate proposals for the introduction of new IT-related technologies and decide over their approval based on legal demands for safety and security.

An important question for the introduction of AI-based technologies in nuclear plants is the physical location of both the compute hardware and data. Based on the conducted interviews, the permissible locations for server installations are in all cases restricted to company premises. This includes in principle both the power plants themselves, as well as office buildings, with the caveat that the communication between any end-user compute device with compute servers is restricted to intranet of each individual company and cannot happen via the public internet. Somewhat more complicated is the question of permissible locations of data. In nuclear power plants, vast amounts of data, such as operational and maintenance logs, are created and stored, and security classification systems determine precisely which data can be accessed by whom and where and how it must be stored. In our interviews, we found potential discrepancies regarding the details of these data security classifications; in some cases, any site-specific data or data generated within the premises of a power plant must remain on-site, while in other cases the processing and storing of such data on remote premises within the same company is likely to be permissible.

The outlined restrictions have consequences for the development, training, and deployment of potential AI applications, depending on the targeted kind of AI model and its data requirements. Here, model development describes the setup of data pipelines, implementation of often various models and model architectures for comparison, and iterative improvements and addition of features to the model, until the code basis for the model is complete. Model training happens repeatedly throughout the development process, and can be repeated or extended (”fine-tuning”) even after the core model development phase is complete. Finally, Model deployment describes the process of integrating the trained model into a production environment, enabling its practical application and utilization by the end-user. Larger models are typically trained on compute clusters as the training process can require vast amounts of compute power and memory capacity, while deployment can happen on much leaner and inexpensive hardware. Hardware

[7] https://www.stralsakerhetsmyndigheten.se/regler/
[8] https://stuk.fi/en/nuclear-safety

solutions to run an AI model on site would thus primarily be dimensioned for the use (also called ”inference”) of the model, not for its training. However, if the training or fine-tuning of a model requires the use of on-site data, the outlined restrictions may necessitate it to be trained on site, impacting hardware requirements. While the problem of data being bound to a particular site may also extend to model development, this can in some cases be circumvented by relying on similar, less safety-critical or even publicly available datasets during the development phase.

Another important consideration is that neural networks may in some cases store parts of the training data within their network weights, allowing the data to be recovered from the network weights alone. This may make it reasonable to extend any restrictions pertaining to training data also to models trained on this data, which in turn may hinder the creation of individual models trained on data from multiple sites or companies. A possible remedy to this issue are federated learning approaches with data privacy guarantees, enabling the training of a single model simultaneously on multiple sites, without the need to collect data on a central server. A different approach to tackle issues related to dataset size is transfer learning, where models are trained on large, often publicly accessible datasets, before being fine-tuned to specific tasks with limited amounts of task-specific data. Foundation models such as Large Language Models (LLMs) build upon this idea. These are typically very large models trained on large amounts of publicly available data that are often capable of solving specific tasks even without additional fine-tuning on task-specific datasets. These examples show the difficulty of making general statements about the feasibility of AI-based solutions in the nuclear energy sector. High demands for security of data and compute systems in the nuclear sector pose a challenge, but various techniques and technologies exist that allow these challenges to be addressed. We thus conclude that any AI application for the nuclear industry must be conceived of and evaluated on an individual basis given the outlined constraints to hardware and data handling.

3.1 SAFETY-CRITICAL AI APPLICATIONS IN THE NUCLEAR ENERGY SECTOR

While explicitly exempt from consideration for the proposed follow-up study – and thus not the primary focus for the more detailed discussions in sections 4 and 5 – AI-based systems for safety-critical applications in nuclear power plants are being pursued in the industry for the purpose of cost saving and improved safety.

It should be noted that some interviewees brought up considerable skepticism to the use of AI systems in safety-critical applications, citing a lack of trust towards AI as decision-makers, but also the risk of automation-induced complacency, i.e., humans supervising or interacting with AI systems becoming negligent and over-relying on the AI system. Indeed, problems in human-automation-interaction have in the past contributed to severe accidents in the nuclear energy industry [1].

The problem of automation-induced complacency has previously been examined in various contexts, for example intensive care units [2], aviation [3], maritime operations [4], and partially automated vehicles [5]. While some psychological research on the topic exists [6]–[8], the relatively small body of available research

does unfortunately not appear to match the importance and scope of the issue as AI and automation support become more and more commonplace in safety-critical operations of various sectors. However, research efforts directly connected to automation in the nuclear energy sector exist, for instance in the form of case studies [1], [9], a literature review on automation trustworthiness in nuclear power plants [10], research on safety-focused design approaches [11] and methods for the adoption of advanced automation [12]. These references may serve as a starting point for a deeper analysis of the scientific state-of-the-art in automation for nuclear power plants and automation-induced complacency.

As the authors of this report, we do not take a general stance for or against the use of AI and automation in safety-critical areas of nuclear power plants. However, for the exploration, planning, and design of such solutions we strongly advise the careful consideration of safety and security related risks not only on the technical side, but also on the human factors side, including the review of research from psychology, human-automation-interaction, and human-centered design.

3.2 FURTHER LIMITATIONS TO THE USE OF AI MODELS IN THE NUCLEAR INDUSTRY

Finally, it should be noted that some developers of pre-trained AI models limit the scope of tasks for which their models may be used to exclude applications in the nuclear industry. An example of this can be found in Meta's Llama 2 family of LLMs, whose license agreement states that "You agree you will not […] 2. Engage in, promote, incite, facilitate, or assist in the planning or development of activities that present a risk of death or bodily harm to individuals, including use of Llama 2 related to the following: a. Military, warfare, nuclear industries or applications, espionage, use for materials or activities that are subject to the International Traffic Arms Regulations (ITAR) maintained by the United States Department of State" [74]. However, specific prohibitions towards use in the nuclear industry are rare, and are not contained in common licensing agreements such as MIT [75] or Apache 2.0 [76].

4 State of the Art in the Nuclear Energy Sector

Two fields in AI with a particularly strong potential to significantly impact operational efficiency, safety and uptime in NPPs are natural language processing (NLP) and computer vision (CV). In this section we will discuss some of the most interesting recent developments in these fields, and how they can be – and in some cases already are – applied in NPPs. This will be done to some extent by looking at other industries, where AI, due to more relaxed safety and security restrictions, is further along in terms of utilization.

4.1 NATURAL LANGUAGE PROCESSING

Processing of text documents comprises a set of daily NPP office tasks that include everything from search and analysis to classification, sorting, summarizing, and editing. Common types of NPP documents include, e.g., incident and inspection reports, operational and maintenance logs, manuals, and regulatory documents. As discussed during the interviews, the number of internal documents is often massive; one of the licensees mentioned that they store about 650 000 documents in various formats and with different security classifications on their servers, as well as 210 000 digital drawings. Another stated that they store around 4 000 000 documents, ranging from a couple of pages to hundreds of pages each. This means that in total, Nordic NPPs are likely storing over 10 million multi-page documents, many of them not possible to be found elsewhere. Even in cases where handling of documents only up to a certain security classification is permitted, the number of documents is often still large enough that just using manual analysis and keyword-based search tools such as those included in most operating systems is not practically feasible. Algorithms with the ability to quickly read, understand and reason about the content in the documents are therefore highly desirable.

4.1.1 Large Language Models

Up until the first half of 2022, state-of-the-art in the field of natural language processing (NLP) had for several years been BERT and its variants [13], a set of transformer-based large language models (LLMs) known for their ability to understand text on a semantic level [14]. BERT is short for *Bidirectional Encoder Representations from Transformers*, where *bidirectional* indicates that it ”understands” the meaning of a given word by taking into consideration its context on both sides, i.e., it looks at the surrounding words in both directions simultaneously. *Encoder* indicates that it is an *encoder-only* model, utilizing only the encoder part of the transformer architecture (Fig. 1), while leaving out the decoder. The encoder transforms input text into contextualized *representations* (in the literature the term used is more commonly *embeddings*), which are vectors (essentially lists of numbers), where each vector represents a word or part of a word in the text, called a *token*.

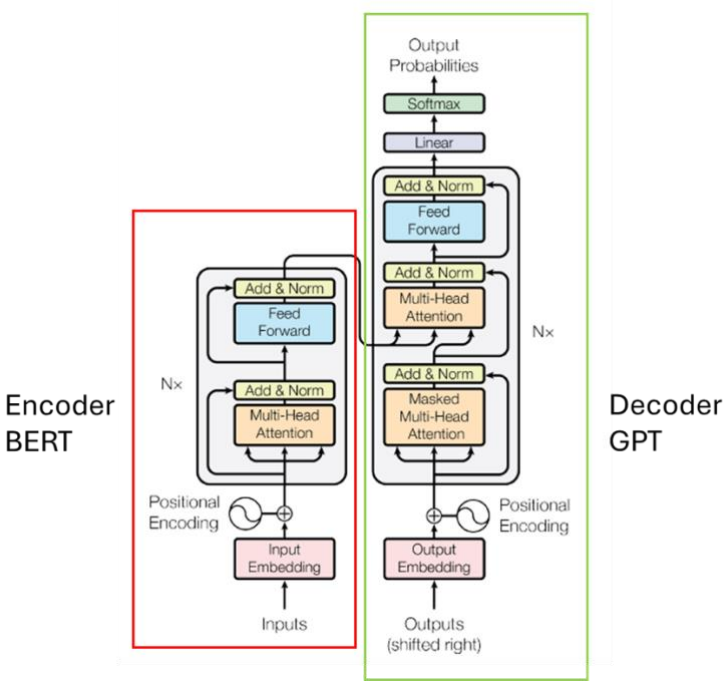


Figure 1: The transformer architecture, with the encoder and decoder parts in colored rectangles. Figure adapted from Fig. 1 in Vaswani *et al.* [14].

BERT is trained on vast amounts of text to create a multidimensional space of such vectors, clustered based on a mathematically defined measure of similarity. During inference, i.e., when the trained model is used for textual analysis, the text fed to BERT is divided into tokens whose vectors are then mapped to locations in latent space according to their contextual meaning. In this way, it is possible to use BERT in combination with additional models or node layers for analytical output like, e.g., sentence classification (whether an utterance expresses a positive or negative sentiment about something, for example) or natural language answers, like chats.

In November 2022, generative AI entered public consciousness through OpenAI's release of ChatGPT, a conversation-tuned variant of the GPT-3.5 architecture. ChatGPT was one of the first LLMs to be easily accessible to anyone with an internet connection, while at the same time being powerful enough to be useful, rather than merely a fun but quickly forgotten gimmick.

GPT is short for *generative pre-trained transformer*, and while GPT models are thus based on the transformer architecture just like BERT, they differ from BERT in that they are *decoder*-only models (Fig. 1) that work by predicting the next word (or token) in a sequence based on the previous words, i.e., they are natively *generative*. LLMs can be prompted to generate new, grammatically impeccable and perfectly coherent text in almost any style and in many different languages, based on conversational context. This context contains part or all the history within the ongoing conversation, system instructions, and other text elements. The context length, or size of the *context window*, which can be regarded the LLM's working memory, is often a significant limitation in current LLMs and an active area of research [15].

	GPT 4o	o1-preview	Gemini 1.5 Pro	Claude 3.5 Sonnet	Llama 3.2 405B	Mistral Large 2
Model size (no. parameters)	N/A*	N/A	N/A	N/A	405B	123B
Developer	Open AI	Open AI	Google	Anthropic	Meta	Mistral
Software License	Closed Source	Closed Source	Closed Source	Closed Source	Open weights (Meta Llama 3 Community License Agreement)	Open weights (Mistral Research License)
Modalities (input and/or output)	Text, images, audio	Text	Text, images, audio	Text, images	Text, images	Text, images
Context window (no. tokens)	128 000	128 000	128 000	200 000	128 000	128 000

Table 2: Top performing LLMs from different developers as of October 2024. *ChatGPT* is the name given to conversation-tuned versions of GPT-4o and most recently OpenAI o1-preview. A 128 000 token context window, which only Claude 3.5 Sonnet exceeds, corresponds roughly to a 300-page book. *Open weights* is different from *open source* in that the former license gives access to and allows training of the neural network parameters of the model, but not the full source code. *It is widely believed that the original GPT-4 contains 1.75T parameters, and that GPT-4o is smaller than that. However, this has not been confirmed by OpenAI.

Pre-trained GPT-based LLMs are trained without human supervision on enormous amounts of data to gain a broad understanding of language and the many concepts and relations it encodes, before they are then fine-tuned to specific tasks. Pre-training is done on public datasets [16], and proprietary data that has either been bought from external parties or created in-house. Because of this, LLMs are likely to have at least *some* knowledge in almost every domain, even before fine-tuning, and can thus in many cases be used out-of-the-box to do a lot of office-related work, including document search and analytics. However, for domains with little data accessible for pre-training, domain-specific fine-tuning may be beneficial.

Fine-tuning typically involves at least one of the following two methods. The first method is ”proper” training where the internal parameters of the neural networks constituting the LLMs’ architecture are updated, using small, curated datasets and at least some level of human supervision to ensure that the output of the model is in alignment with human values and desires, and as free of *hallucinations* (recently established LLM terminology for falsehoods) as possible. ChatGPT and other popular LLMs (see Tab. 2) have been fine-tuned in this way, in most cases with the objective of making them great at human-like conversation. The same method can be applied to NPP-specific data to make them native experts in the nuclear domain, but is, as we will discuss in Sec. 6, computationally rather expensive.

The second method is *in-context learning*, where the parameters are left untouched, and the model instead learns by example or by additional information being fed to

it as part of the conversation. Examples can be prompts in the form of task-solution pairs that are provided to the model, which teaches it how to solve similar tasks [17]. Additional information can be documents on a particular topic, which enhances the LLM's knowledge and understanding of this topic (and, as a bonus, enables discussion about the contents of the documents). In-context learning is not limited by the amount of compute as much as it is limited by the model's context window, which currently only allows for a few hundred pages of text in the most advanced LLMs. However, there are methods to circumvent this limitation, as we will discuss in Sec. 4.1.3.

Finally, we note that, although GPT-4o and its counterparts are still most commonly called large *language* models – the reason why this report will stick with this term – many of these models are now actually large *multi-modal* models, since they can take as input and/or produce as output data of other types, or *modalities*, than just text (the "o" in GPT-4o stands for *omni*, to signify its multi-modal capabilities). In future iterations of these models, it is conceivable that the number of modalities they can handle will increase and thereby unlock completely new use cases.

4.1.2 Impact of Model Size and Language on Performance

At the time of writing, the LLMs in Tab. 2 are those with the highest average performance scores taken over the rather extensive set of public LLM benchmarking datasets. However, according to the interviewed licensees, closed source models such as the ones in this table, which can only be accessed over the public internet, are allowed to handle only a small subset of the data in NPPs. Furthermore, they are too large (in terms of neural network parameters) to run cheaply and efficiently on local servers. An alternative is to use smaller open source or open weights (referred to only as "open" henceforth) LLMs, a small selection which is listed in Tab. 3. More extensive lists over open LLMs can be found online.[9]

A decrease in model size generally leads to lower performance on the benchmarking datasets. This is exemplified by Fig. 2, which shows the accuracy of the models in Tab 3 on the *Measuring Massive Multitask Language Understanding* (MMLU) benchmarking dataset [18], which is one of several common LLM benchmarks. MMLU is a general reasoning dataset where the models are tasked with answering multiple-choice questions in all kinds of subjects, including civics, economics, math, and physics. It is intended as a test of textual understanding, and how well the models can utilize their innate world knowledge gained from pre-training to answer the questions. The creators of the MMLU benchmark estimate that human expert level accuracy is about 90%, which is not much higher than the best <100 billion parameter LLMs in Fig. 2, and about the same as the largest state-of-the-art LLMs: Llama 3.2 405B scores about 87% and GPT-4o 89% [19]. However, it should be noted that parts of these benchmarks, despite their creators' best efforts, may have ended up in the training data of the models, which would make their performance scores more difficult to interpret[10]. Moreover, the accuracies in

[9] https://github.com/eugeneyan/open-llms

[10] An improved version of the MMLU benchmark, the MMLU-Pro benchmark [15], has recently been released and may soon make benchmarking on the MMLU obsolete; however, at the time of writing, its novelty means that only a few of the most recent LLMs have been tested on it.

Fig. 2 come from the developers themselves; independent test results would be preferable but are currently quite difficult to find. Nevertheless, on some benchmarks, in particular the recently instituted ARC-AGI[11] and SimpleBench[12], the best LLMs are still quite far from reaching even human *non-expert* level performance[13]. Another thing to note is that designing an LLM – which often means more than just fine-tuning it – for a specific task, like programming, can help bridge the performance gap between smaller and larger models on this task; for example, according to Mistral, their 22B parameter LLM Codestral, which has been designed for programming, beats out the general purpose model Llama 3 70B on several programming benchmarks[14].

	Llama 3	**Mistral NeMo**	**Qwen 2.5**	**OLMoE***	**Phi-3.5-MoE***
Model size (no. parameters)	1B (Llama 3.2), 3B (Llama 3.2), 8B (Llama 3.1), 11B (Llama 3.2), 70B (Llama 3.1), 90B (Llama 3.2)	12B	0.5B, 1.5B, 3B, 7B, 14B, 32B, 72B	1B active, 7B total	7B active, 42B total
Developer	Meta	Mistral & NVIDIA	Alibaba Cloud	Ai2	Microsoft
Software License	Open weights (Meta Llama 3 Community License Agreement)	Open weights (Apache 2.0)	Open weights (Apache 2.0, Qwen Research (3B, 72B))	Open source (Apache 2.0)	Open weights (MIT License)
Modalities (input and/or output)	Text, images (11B and 90B versions only)	Text	Text	Text	Text
Context window (no. tokens)	128 000**	128 000	128 000**	4096	128 000

Table 3: Some notable smaller open LLMs as of October 2024. *A so-called *Mixture-of-Experts* model, where only a fraction of the total number of parameters are activated per input token. In the literature, these models are usually compared to non-MoE LLMs with a similar number of parameters as their active ones. **True only for some of the larger versions. Smaller model versions generally have smaller context windows.

An additional aspect of LLM benchmark performance is how it varies with respect to language. Perhaps not surprisingly, there seems to be a strong tendency for performance to be higher for high-resource languages (i.e., languages for which there is a lot of online text), as well as for lower-resource languages with strong similarities to high-resource languages, such as Afrikaans [22]. In the context of the Nordic NPPs, this is important as many of their documents are in either Swedish

[11] https://arcprize.org/arc
[12] https://simple-bench.com/index.html
[13] Coming up with benchmarks for LLMs that capture all aspects of their intelligence is quite difficult, for several reasons. This is reflected in the steady stream of new benchmark proposals, coming both from the AI model creators themselves, and from independent researchers.
[14] https://mistral.ai/news/codestral/

or Finnish, which are classified as mid-resource EU languages in [23][15]. The authors point to two open source LLMs created specifically to perform well in Nordic languages: the Finnish-specialized Poro 34B, created by Silo AI in collaboration with University of Turku and the Horizon Europe funded High Performance Language Technologies (HPLT) project, and the same group's subsequent Viking model family[16],16 with 7B, 13B, and 33B parameters. Besides Finnish, the Viking models have been trained for improved performance in Swedish, Norwegian, Danish, and Icelandic.

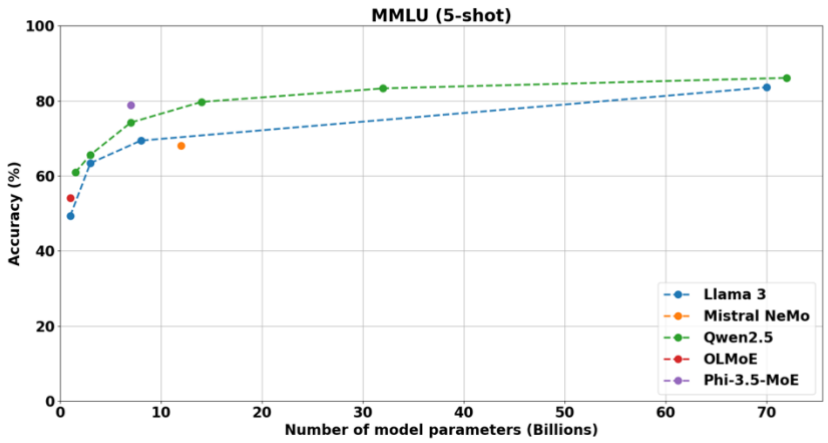


Figure 2: Accuracy on the MMLU benchmarking dataset for LLMs with less than 100 billion parameters (scores for Llama 3.2 11B and 90B could not be found). *5-shot* means that 5 question-answer pairs are fed to the LLM before testing starts, as examples of how to do the test. This in-context learning often increases the performance compared to *0-shot*, where no examples are provided. Sources: Llama [19],Meta blog; Mistral Mistral blog; Qwen Qwen blog; OLMoE [20]; Phi [21].

On average, Poro 34B outperforms other small open models, including the original Llama 33B, on several different common benchmarks that have been translated into Finnish, as well as on English-to-Finnish translation tasks [24]. While it is not clear how Poro compares to the latest small open models, its performance increase compared to the next best LLM in the study (Llama 33B), combined with the fact that the other LLMs perform worse in Finnish than in English, points towards research into LLM multilingualism being a worthwhile endeavor.

For the Viking models, no benchmark scores could be found, but Silo AI claims state-of-the-art performance compared to other small open models with respect to the five Nordic languages mentioned above. In their claim, they include GPT-SW3, which is a GPT-3 based model trained on a text dataset called *The Nordic Pile*[17], which is a Swedish-heavy dataset collected by AI Sweden.

Besides national language, some researchers have suggested that industry and even plant specific language may require some consideration. In 2021, the Electronic Power Research Institute (EPRI) set out to collect words and phrases in

[15] The authors of the study confusingly calls Swedish both mid-resource and high-resource
[16] https://www.silo.ai/blog/viking-7b-13b-33b-sailing-the-nordic-seas-of-multilinguality
[17] https://www.ai.se/en/project/gpt-sw3

four technical areas within the nuclear domain, to create a dictionary for use in nuclear NLP applications[18]. The current status of the project is unclear, however. Moreover, it is also unclear whether such a dictionary is still needed, considering how much more capable current LLMs are compared to the BERT models that were state-of-the-art when the project started.

Finally, understanding how the performance on the many benchmarks relate to a given use case, and which LLM to go for – if any – is not entirely easy. For use cases where long and nuanced conversations on a broad set of topics are expected, a large general-purpose model with a large context window may be the only type of model that works well enough. For other use cases, a smaller and possibly task-designed model may work just as well from a performance standpoint, while at the same time saving the user a lot of money on hardware and other resources necessary for running the model. A good rule of thumb could be to look at benchmarks that seem most relevant to the use case, choose the best performing, regulations compliant model that the available hardware can handle, and then subject it to careful evaluation on use case data. To maximize performance, language may be a factor to consider as well.

4.1.3 Retrieval-Augmented Generation

A retrieval-augmented generation (RAG) system typically combines retrieval-based models such as Dense Passage Retriever [25] (in which BERT serves as an embedding model) with generative LLMs to enhance the quality and accuracy of the generated responses, while reducing the need for a large context window[19]. For these reasons, RAG systems may be particularly useful when dealing with large knowledge bases or answering fact-based questions that require highly domain-specific knowledge that may not be very well represented in the training data.

A basic (or *naive*) RAG system [26] first uses an embedding model to vectorize the input question, to enable a similarity search of an external database containing vector representations of text chunks from various types of text documents, to find the chunks most relevant to the question. An LLM then receives both the question and the retrieved chunks as input to generate an answer. The idea is that this will allow for more accurate and up-to-date answers.

Since new data can be added to the database at any time, RAG systems, which can have many different architectures and different levels of complexity [26]–[29], seem especially well-suited for domains where data changes frequently, such as in an NPP environment. Moreover, by not relying only on the frozen data used for pre-training of the LLM, the level of hallucinations can potentially be significantly reduced, and the answers be more specific.

[18] https://eprijournal.com/a-dictionary-to-help-ai-tools-understand-the-language-of-the-electric-power-industry/, https://www.epri.com/research/products/000000003002023822
[19] Knowledge graphs can also be used as retrieval model, but this seems to be less common.

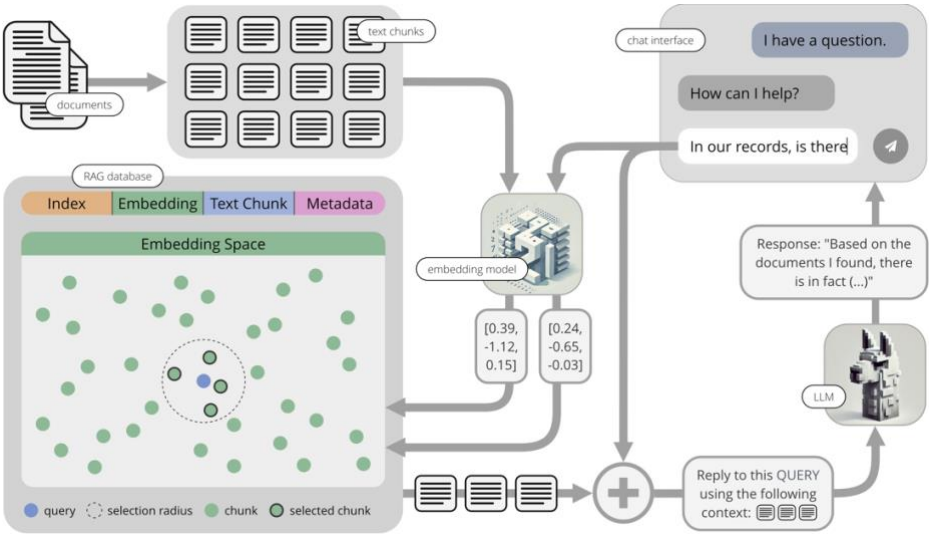


Figure 3: A basic RAG system.

Examples of LLM-based RAG for the nuclear domain are still scarce, but in a small study by Anwar *et al.* [30], the authors use ChatGPT-3.5 out-of-the-box in a basic RAG system for retrieval of information from a textbook on CANDU-type (CANada Deuterium Uranium) reactors, *The Essential CANDU*, and compare the performance of this system to direct responses from ChatGPT-3.5 (i.e., without RAG). The RAG system performs significantly better than the LLM alone on all evaluation criteria except one. It should be noted that ChatGPT-3.5 has since been far surpassed in performance several LLMs, including ChatGPT-4o and even the much smaller version ChatGPT-4o Mini. However, if the RAG system is set up in a sufficiently modular way, the LLM can quickly be replaced by more powerful models as they are released; how to modularize more complex RAG systems is discussed in [31].

While RAG systems hold a lot of promise, there are still challenges and limitations to overcome. Barnett *et al.* [32] conducted three case studies with RAG, and list seven different failure points they encountered, stemming from different RAG system components. They also briefly discuss the need for proper RAG system metrics. Ru *et al.* takes this discussion further by proposing the RAGCHECKER evaluation framework [33], which calculates three different metrics: the response quality of the RAG system (Overall Metrics, in terms of precision and recall), how good the retrieval process is at finding exactly the right information needed to generate a correct answer (Retriever Metrics), and, essentially, how well the LLM behaves with respect to hallucinations and four other factors, all of which the authors define in precise mathematical terms (Generator Metrics).

However, even a RAG system boosted by a powerful, but passive LLM may not by itself always be enough for the level of automation many companies envision for their document processing pipelines. For very complex multi-step tasks, additional components may be needed, such as AI *agents*.

4.1.4 Agents

An AI agent is an AI model that can solve long-horizon tasks that require planning and the ability to break the tasks down into subtasks, use tools to carry out these subtasks, and then memorize the subtasks' outcomes so that they can be used for subsequent subtasks. AI agents have recently become a very popular topic of discussion in relation to LLMs. While current base LLMs have vast world knowledge and can solve complex problems in one go, they are not yet able to function very well as agents, especially not on their own. One of the first attempts at creating an LLM-based agent was the open-source project AutoGPT[20], which uses GPT-4 API calls in a self-prompting loop to reason about a (user-prompted) task. From this reasoning, it creates a step-based plan, criticizes the plan to potentially improve upon it, takes an action based on the plan, and then uses the outcome of this action to update the plan and carry out the next step. This loop then runs until the main task has been solved. An action in this context can be anything from requesting information via an API to sending a control signal to a robot. Moreover, multi-agent collaboration is an active field of research that clearly points to a possible future enhanced through agents with the ability to self-organize [34], [35].

One or more agents integrated into a RAG system could potentially provide enhanced functionality compared to a RAG system with a non-agentic LLM. For example, an agent could interact with the retrieval model iteratively, to refine questions based on previous responses. The agent could prompt follow-up retrievals or adjust the questions if conflicting documents or unclear information is retrieved, thus potentially increasing the precision and relevance of the retrieved information. Agents would also open for actions to be taken on the retrieved information, such as maintenance scheduling, sending notifications, or generating maintenance reports.

A major problem with agents such as AutoGPT, including multi-agent systems, is that they often break down well before they have finished the main task, mostly due to compounding hallucinations. This makes current agents too unreliable for many tasks, especially more complex ones. On the other hand, as LLMs continue to improve, utilizing agents will become increasingly feasible. Eventually, agentic behavior will likely become an ability innate to many LLMs. This was first hinted at by OpenAI's latest models o1 and o1-mini, that can solve complex problems through chain-of-thought (or stepwise) reasoning and has now started becoming reality after Anthropic's release in October 2024 of a new version of Claude 3.5 Sonnet, with the ability to control a computer[21]. However, while Claude provides a new state-of-the-art on the OSWORLD benchmark [36] – a real computer environment for benchmarking of multi-modal agents – it only scores a 14.9 % success rate on tasks in the screenshot-only setting, compared to 72.36 % for humans. In this human-like setting, the agent decides based on screenshots of the OS environment how to move the mouse cursor and click on icons in the environment.

[20] https://github.com/Significant-Gravitas/AutoGPT
[21] https://www.anthropic.com/news/3-5-models-and-computer-use

4.1.5 Applications of LLMs in the Nuclear Industry

Since LLMs are a fairly new technology, many companies and institutions interested in them still need to find out how to leverage them to best benefit their businesses. Due to the LLMs' ability to extract relevant information from all kinds of documents, summarize the most salient points in a text, draw conclusions, or even make suggestions or recommendations based on texts, LLMs have many potential use-cases. For many real-world scenarios however, LLMs are becoming part of larger tool chains, requiring additional development beyond the mere deployment of chat-bots to create task-specific solutions. Many of the applications discussed with the NPP licensees fall within this category. Below we look at some of these applications, as well as a few that were not brought up but that may still be relevant in an NPP context.

Operator Training Provided an LLM fine-tuned on relevant NPP specific documentation, or one that is part of a RAG system that has access to such documentation, NPP operators could potentially use the LLM as a teacher, instructor or teaching assistant when learning how to work with new hardware or software in the plant. Millions of users are already using LLMs for informal teaching at work or in their spare time, and some online academies have started integrating LLMs into their teaching software; Khan Academy's GPT-4-powered assistant Khanmigo is probably the most well-known example of such integration[22]. However, in industrial settings – perhaps NPPs in particular – where requirements on safety and security are often very strict, partially or fully automating operator training with LLMs must be preceded by careful evaluation of the models.

Operational Support and Troubleshooting Operational support and troubleshooting is a natural extension of an LLM tailored to function as an instructor for operators. Again, through proper fine-tuning, RAG, or both, it would be possible to instill into the LLM plant specific knowledge down to individual machines and processes. If the LLM is then coupled to a user-friendly user interface – based on chat, voice, or video, or on a combination thereof – it may be able to provide support on, e.g., how to adjust the parameter settings of a machine or help troubleshoot it when it malfunctions. This is an active area of research, but papers focused on NPPs are still rare.

In [37], the authors discuss a demonstrator they created to showcase how LLMs can be integrated into fault diagnostics systems to provide explainability to sensor signals indicating faults in, e.g., nuclear power plants. The demonstrator used GPT-4 to enable operators to ask questions and receive answers about diagnoses in natural language, with the answers containing information about the origins of the faults as well as what effects they may have.

In a recent study by Freire *et al.* [38], they developed an LLM-based RAG system for operational support and troubleshooting in a detergent factory, to better understand the benefits, usability, risks, and barriers to adoption of such a system. Several different commercial and open (non-fine-tuned) LLMs were evaluated, including GPT-4 and Mixtral 8x7b (two models that performed comparably on

[22] https://www.khanmigo.ai/

most measures, with a slight edge to GPT-4). The data used were factory documents and issue analysis reports, the latter which is a type of report that is constantly incoming and would thus require regular fine-tuning to be incorporated into the LLM, which is difficult to accomplish in practice. The evaluation of the system was done outside of production, but with factory operators. While the system showed a lot of potential, operators expressed worry about risks and still preferred to talk to human experts, thus highlighting the need for further development of various aspects of the system, including the user interface.

Commercial LLM-based solutions that the vendors claim can do both operator training and operational support and troubleshooting are already on the market, but how well they work is unclear as independent tests are hard to find[23].

Finally, it is important to note that there is a lot of risk involved in introducing automation for applications that many times fall in the safety-critical category. As we discuss in Sec. 3.1, even if no decisions are ever made directly by the LLM and all it ever outputs are either suggested actions or just explanations of what might be wrong, automation-induced complacency could lead to highly negative outcomes of the model hallucinates.

Regulatory Compliance Since nuclear power plants are heavily regulated, plant operators frequently retrieve, read and interpret information in regulatory documents to understand whether plant operation, equipment, and processes are in compliance with these regulations. This often quite labor-intensive task may be well-suited for a RAG system where an LLM has access to all regulatory documents as well as documents on all the aspects of the plant that are covered by these regulations (e.g., all technical documents). To minimize the risk of inadvertent non-compliance, a mechanism could be set up that ensures that the system is updated as soon as something in the plant changes, or when there are changes to the regulations. Moreover, it may be possible to configure the system so that an automatic compliance check by the LLM is triggered every time such changes occur.

Using LLMs for regulatory compliance verification has been explored for other industries and domains. Fuchs *et al.* [39] evaluated GPT-3.5 for automated compliance checking of buildings designs, using in-context learning and two other techniques. Another example is Berger *et al.* [40], who compare different LLMs (two versions of GPT-3.5, GPT-4, and Llama2 7B, 13B and 70B) with respect to their ability to verify that corporate financial documents comply with regulations. The studies share the conclusion that, while the outcomes are promising, fine-tuning on domain-specific data should be explored to boost performance. The Berger study discusses language as one of the reasons to do fine-tuning; part of the financial document dataset they use for testing is in German, and the LLMs perform much worse on this part than the English one.

Report Generation Writing reports is an important task in an NPP, as it helps maintaining things like operational efficiency, safety, and regulatory compliance. Many types of reports are written in NPPs, such as reports on plant operations, incidents, events, safety issues, and maintenance. While much of this reporting is

[23] https://www.symphonyai.com/industrial/industrial-llm/

already automated using various digital systems, large language models can potentially handle cases where more complex and detailed reporting is needed, or, as discussed under ”Operational support and troubleshooting” above, provide explainability to reports.

Studies on report generation come mostly from other domains, such as medicine and engineering. Nakaura *et al.* assess the use of GPT-2, GPT-3.5, and GPT-4 for generation of radiology reports, by comparing the GPT-generated reports to reports written by human radiologists [41]. They find the reports by GPT-3.5 and GPT-4 to be very convincing from a linguistic standpoint, but the radiologists are still more accurate when it comes to the actual diagnostics, and the authors therefore recommend use of LLMs only as an aid.

Colverd *et al.* present an LLM-based RAG system for generation of flood disaster impact reports and compare their quality to that of human-written reports [42]. They test Google’s older model PaLM-Text-Bison, GPT-3.5, and GPT-4, and conclude that GPT-4 yields the best results. However, they too advice that the LLM is not to be left to work completely on its own.

Lastly, in [43], the authors introduce a RAG system for synthesis of comprehensive work session safety and security reports, from operational logs and session descriptions. They evaluate the quality of reports generated from data from the Aviation Safety Reporting System (ASRS) database, using different combinations of LLMs and embedding models, specifically different sizes of the original Llama, and different versions of BERT. On this dataset, which they only use as an example, the authors see good performance, in particular from Llama 70B combined with AeroBERT.

Requirements Engineering Requirements engineering includes several different activities, such as elicitation (i.e., gathering of requirements from stakeholders), analysis, refinement, specification, and management of requirements for a product or a system, such as an NPP. While it is conceivable that large language models could help with all these activities, research into requirements engineering specifically for NPPs seems to be mostly focused on how the requirements are represented to the engineer, and on requirements management [44]; according to the authors ”...it is relatively clear what a nuclear power plant should and should not do.”

Efficient and accurate requirements classification is a time- and cost-saving measure that makes it a lot easier to ensure that the project stays aligned with the requirements as it proceeds. In work from 2019, Myllynän in collaboration with Fortum trained an NPP requirements classifier based on a feed forward network coupled to an NLP model consisting of a recurrent neural network with long short-term memory cells [45]. The training dataset consisted of regulatory guides (YVL Guides) from the Finnish Radiation and Nuclear Safety Authority (STUK), while evaluation was done on requirements from both Finnish and UK nuclear regulatory authorities. Although the model achieved promising results, generative LLMs may be even better suited for this type of application due to their vastly improved textual understanding compared to any models available in 2019.

Predictive maintenance Predictive maintenance is a highly active field of research that has the potential to significantly increase operational efficiency, uptime, and safety in nuclear power plants. Predictive maintenance commonly entails AI-based analysis of real-time data originating from plant sensors, to find trends or anomalies known from training on historic data to correlate with subsequent failures or even plant outages. On the nuclear side, predictive maintenance research is still mostly carried out using various types of traditional machine learning algorithms and models based on, e.g., long short-term memory (LSTM) networks, which, in contrast to LLMs, are designed for direct processing of time-series data [46], [47]. However, there are still ways to leverage the power of LLMs for predictive maintenance. One is to combine LLMs with models that can pre-process the time-series data. In this hybrid approach, a model such as an LSTM could be used to transform the time-series data into statistics like averages, standard deviations, and max and min values, which can then be presented to the LLM in natural language format for further analysis. Another and potentially more powerful approach is to use prompting techniques such as in-context learning and chain-of-thought prompting to essentially elicit in the LLM a latent ability to do time-series analysis, as suggested in [47]. The authors show that providing GPT-4 with prompts containing both the time-series data and additional questions, instructions, or information about this data, it is possible to make the model process at least shorter time-series data with good results. They also evaluate Llama 3 8B, which does not perform well out-of-the-box due to its much smaller parameter size but improves significantly on several measures with fine-tuning on prompts similar to those used with GPT-4, but with the answers included.

It is worth noting that the new (at the time of writing) OpenAI-o1 series models are chain-of-thought reasoners innately, which could mean that they are even better suited for time-series processing than one-forward-pass LLMs like GPT-4. However, their much longer inference times may be disqualifying in many cases.

Programming Two surveys[24][25] from 2024 indicate a widespread adoption by software developers in several countries of LLMs as tools for programming. Considering the overall very positive sentiment expressed by the surveyed developers, this trend seems likely to follow in all industries where programming is part of the job, including the nuclear industry.

In addition to code generation, which includes generation of code from scratch as well as code completion, LLMs can do bug fixing, commenting, unit testing, and documentation. LLMs are commonly proficient in all major programming languages, but just as with natural language, where most LLMs tend to score highest when both input and output is in English [22], there is likely a strong correlation between proficiency in a particular programming language and the amount of training data that exists for it. While the context windows of the most powerful models (see Tab. 2) are often too small to generate an entire code base, the models can be very useful for generating smaller chunks of code or scripts for, e.g., data processing and visualization.

[24] https://www.bairesdev.com/blog/72-software-engineers-genai-productivity/
[25] https://github.blog/news-insights/research/survey-ai-wave-grows/

4.2 COMPUTER VISION

AI-based computer vision is likely the most industrially mature type of AI, with many commercial actors already selling computer vision systems for applications such as quality inspection and various kinds of real-time monitoring. These systems have started to find their way into not only the manufacturing and process industries, but the nuclear industry as well. As stated in the introduction of this section, computer vision came up as another interesting AI technology track during the interviews with the NPP licensees, albeit to a much lesser extent than LLMs. Here we will give a brief overview of the field of computer vision and then discuss a few NPP-relevant applications in little more detail.

4.2.1 Computer vision models and techniques

Most industrial computer vision systems that use AI rather than rule-based vision algorithms[26] still rely on AI architectures that can only be trained on and process image data. However, in the last couple of years, transformer-based multi-modal vision models have started making inroads into the field, with expanded capabilities like visual-question answering, where the user can converse with the model about the contents of the images (GPT-4o is an example of such a model), and image editing, where the model can be asked to add, modify, or remove elements in images. However, both the older and newer architectures can handle a lot of the most industrially important computer vision techniques: *image classification*, *object detection*, *image segmentation*, *anomaly detection*, *optical character recognition*, and *pose estimation*.

Image classification Image classification is the most basic computer vision technique, where the model is trained to classify images according to their content, i.e., it can tell the user whether an image contains a cat, a car, a person, etc (see Fig. 4). Training is done by feeding the model images together with their content labels. AlexNet, the model that kick-started the deep learning revolution in 2010, is an image classification model [48].

Object detection Object detection is slightly more advanced than image classification (Fig. 4). Object detection models can not only classify the objects that an image contains but also locate them in the image by placing a (usually rectangular) bounding box around them. Training of object detection models can be a bit more laborious than image classification models, since each instance of an object in every image must be labeled. One of the most popular object detection models is the YOLO model family [49], which is open source and relatively easy for someone even with only rudimentary Python programming skills to train and apply.

Image segmentation Image segmentation is the technique of doing pixel-level classification of objects in an image (Fig. 4). This technique is more advanced than

[26] Rule-based computer vision uses human-defined image features such as edges, shapes, and color gradients to determine the contents of an image. This type of computer vision is quite rigid in terms of the patterns it can recognize, but for some industrial vision tasks, where the patterns have simple shapes that do not vary a lot, and the features are therefore easily defined, they can still be very useful. An example of such a task is verifying the presence of boreholes.

object detection, since it localizes objects more exactly. Segmentation models can be trained to understand which pixels belong to which individual object even when two or more objects partially overlap. Training a segmentation model is a slightly more involved procedure than training an object detection model, since the bounding box around each object in the images cannot just be a simple shape, like a rectangle, but has to be a filled polygon that closely follows the outline of the object – a so-called *segmentation mask*[27]. There are many segmentation models to choose from, but some of the most powerful ones are the open source models

Segment Anything and Segment Anything 2 from Meta [50], [51]. Moreover, the latest iterations of YOLO can now also do image segmentation[28].

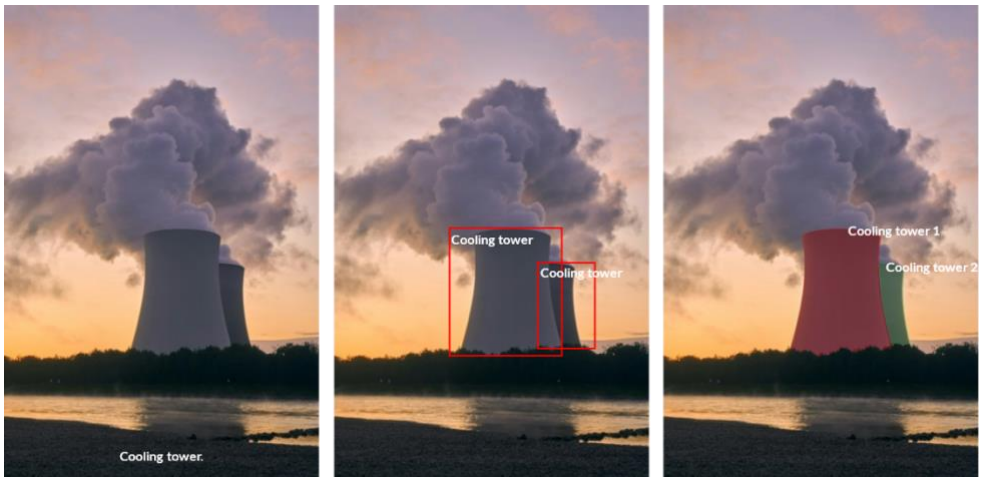


Figure 4: Illustrations of different computer vision techniques: image classification (left), object detection (middle), and image segmentation (right). Source: Wikimedia commons (modified).

Anomaly detection Anomaly detection is a popular technique for quality inspection, where the task is to find, e.g., surface defects or deviations on products or structures (Fig. 5). This technique is different from those previously mentioned in that the training is done on images that *do not* contain the objects that the model is supposed to detect or classify. For example, if the purpose of the anomaly detection model is to detect cracks in walls, then it should be trained on a representative distribution of crack-free walls, in order to learn what such walls look like. Images of walls *with* cracks will then be difficult for the model to process, and the output will look very different – anomalous – compared to the output from processing of an image of a crack-free wall.

Anomaly detection models are suitable in cases where the captured images are very similar to each other (due to fixed camera angles and lightning etc.), and when the anomalies are rare. Anomaly detectors are also less labor-intensive to train than both object detectors and image segmentation models, since very little labeling is usually needed; the images just have to be sorted according to whether they contain anomalies or not. One big downside is that anomaly detection models

[27] This process can be partially automated, however, which reduces the need for very precise labeling.
[28] https://github.com/ultralytics/ultralytics

cannot by themselves classify the anomalies. The ability to classify anomalies is sometimes desirable as it gives the ability to gather statistics over occurrences of different kinds of anomalies, which in turn can lead to better root cause analyses. On the other hand, anomaly detectors can be used to collect data for training of models than *can* do classification. Anomaly detection models can be found as open source, e.g., the Anomalib library [52], which contains several anomaly detection models based on different architectures, where the most powerful ones utilize transformers.

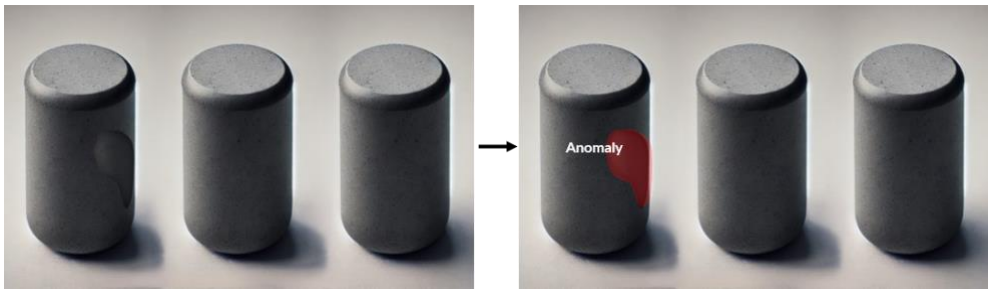


Figure 5: Anomaly detection. The direct output is commonly a heat map that is transformed into a segmentation mask (red field in the right image) based on a threshold value.

Optical character recognition Optical character recognition (OCR) is a technique to detect and label text in images. Many popular OCR models rely on a combination of different models, such as an image segmentation model to find individual characters and words, and a sequence prediction model like a long short-term network (LSTM) to predict the resulting word sequence. EasyOCR is an example of an open-source OCR framework which allows for different combinations of models to be used for the OCR pipeline[29]. However, multi-modal OCR models like CLIP4STR [53], which is based on OpenAI's CLIP model [54], seem poised to become the new standard due to their performance advantage. Large language models with multi-modal capabilities can also do OCR and are powerful enough out-of-the-box to be useful for OCR tasks that do not require near-perfect accuracy[30], but for tasks that do, fine-tuning is necessary. Due to LLMs' much higher computational cost with respect to fine-tuning and inference, models like CLIP4STR will likely still be preferable for OCR for some time.

Pose estimation Pose estimation is the ability to determine the pose of an object, i.e., its orientation in space. Pose estimation models commonly track the coordinates of points on the object. Pose estimation is very important for applications such as bin picking, where a robot has to determine on-the-fly how to grip an object in order to safely pick it up. The latest YOLO models can do pose estimation.

The field of computer vision is developing quickly, and it is becoming increasingly easy for companies to use in-house competence to experiment with different vision techniques. One important factor that is still being researched, however, is generalizability of the models. A common problem when a company wants to

[29] https://github.com/JaidedAI/EasyOCR
[30] https://blog.roboflow.com/best-ocr-models-text-recognition/

implement computer vision for a task is that the models often become quite inflexible in terms of their ability to handle objects and anomalies that they have not been trained on.

Nevertheless, in contrast to LLMs, computer vision systems are already used in several applications in NPPs. Below we will look at a few examples, some of which were mentioned during the interviews.

4.2.2 Applications of Computer Vision in the Nuclear Industry

Visual event monitoring In an NPP, visual event monitoring is an important safety measure. An event can be understood both as something expected happening that needs tracking, or as sudden occurrences of something unwanted or suspicious, e.g., a pipe that starts leaking, a machine that catches fire, or the presence of persons or objects that should not be in a certain room or on or near the NPP premises at all. Such monitoring has historically been done solely by human operators, but with computer vision it is possible to automate this task, or at least significantly reduce the burden on the operators by acting as a filter that alerts them for review of the recorded videos only occasionally.

The literature on computer vision for event monitoring in NPPs and NPP-related facilities is abundant, although, for security reasons, some studies base their results on simulated data, which may limit their interpretability. In [55], for example, the authors present a computer vision model for automated surveillance, based on anomaly detection. The authors train the model on video from a mock-up of a dry room for pyroprocessing of spent nuclear material, to monitor for deviations in the predetermined path of motion of the material via a gantry crane, and for unauthorized persons.

In [56], the object detection model YOLOv7 was trained on real surveillance camera data to look for and track the (usually planned) movement of spent nuclear fuel casks, with the purpose of reducing the amount of video that the operators then have to watch. The operators are tasked with manually inspecting all parts of the video covering the movement to look for anomalies, which is extremely time-consuming; the proposed solution could potentially cut out all parts that are not covering the moving casks.

Drone detection as a security measure for, e.g., NPPs, is discussed in a paper by [57], where YOLOv5 is used to differentiate drones from birds with high accuracy.

These and other studies on various forms of monitoring show the potential of current computer vision for applications in NPPs, but in particular data is still an issue in many cases. For example, how to capture a sufficient number of samples of normal events in, e.g., a room in an NPP may not always be entirely clear if the set of possible events classified as normal is very large. Similarly for detection models, if an object class has a lot of variation in terms of visual characteristics – which is the case for the classes ”drone” and ”bird” – it is important to capture this variation in the dataset. One solution here can be generation of synthetic data, which can increase the variation in the datasets significantly; another solution is to explore the use of multi-modal vision models, in which a richer ”understanding”

of the objects of interest may be instilled via training on additional data modalities that also describe aspects the objects, like their sound profiles.

Visual anomaly and quality inspection Another important safety measure in NPPs is visual anomaly and quality inspection. The computer vision techniques discussed in Sec. 4.2.1 enable automation of various NPP inspection processes, such as wear and tear of various components in the plant (e.g., pipes and turbines), the condition of welds and bolts, the structural integrity of walls and other parts of the plant building itself, and the quality of the nuclear fuel.

Cracks are a common type of wear or damage in NPPs, which is why a lot of research in computer vision for NPP applications is focused on crack detection. A recent example is Yu *et al.* [58], who proposed a crack detection model based on image segmentation for detection of cracks in nuclear containment buildings. The model, which uses a U-shaped neural network architecture with convolutional layers, was trained on 400 raw images of cracks and shows very good performance on the researchers' own dataset as well as on the public DeepCrack dataset[31]. It is also able to accurately measure the size of the cracks.

Image segmentation is also the basis of another recent crack detection model proposed by Li *et al.* [59], in this case for inspection of nuclear fuel pellets. Their work focuses on the data annotation process, which is particularly time-consuming for image segmentation, as discussed in 4.2.1. They integrate into their detection model a novel method for automatic transformation of simple bounding box annotations into the more complex segmentation masks that segmentation models require as training input. This method significantly reduces the annotation time, while still performing comparably to segmentation models trained on human-annotated segmentation masks.

A different kind of task that can also benefit from computer vision is reading of analog NPP equipment gauges to detect anomalous values. This is currently done manually as part of the frequent inspection rounds that operators are required to do in most, if not all, NPPs [60]. The readings could be automated either by mounting cameras in the NPPs that constantly monitor the gauges using an OCR model, or by mounting them on mobile robots or drones, as demonstrated by the American company Boston Dynamics[32]. However, security regulations may be a roadblock for either of these solutions.

Digitization and Digitalization of Engineering Drawings Many NPPs are in possession of old, hand-drawn engineering drawings (EDs) of plant facilities and components such as reactors, piping, and electrical systems (Fig. 6) that only exist as physical documents or as image documents in, e.g., PDF or TIFF format. While some of these drawings describe obsolete facilities and components or are prohibited from going through even the most basic level of digitization due to security concerns, other EDs are still both relevant and permitted to fully digitize.

The process of digitizing an ED involves scanning it into an image and from this image convert its elements, such as text, symbols, and lines, into a format that

[31] https://github.com/yhlleo/DeepCrack/tree/master/dataset
[32] https://bostondynamics.com/solutions/inspection/visual/

allows for manipulation of these elements as well as for text search. This essentially means converting the image into a format that can be interpreted by CAD software.

Software for digitizing EDs has been available for decades, but despite this the digitization process still requires a lot of manual work, especially in cases where the EDs are very complex, contain notes or densely packed elements (in particular notes and elements in non-standard fonts or shapes), or are damaged or covered in dirt. For a very long time the best ED digitization software relied heavily on hardcoded algorithms, but in the last few years AI-based computer vision as a tool for ED digitization has become a very active area of research and now seems likely to become the new standard. However, while many different computer vision models, including the YOLO model family and various segmentation models, have been evaluated for different aspects of the digitization process, such as symbol detection and recognition, challenges remain. As discussed in a recent review article [61], there is a general lack of public ED datasets, and even when data exists, it is often not annotated. Annotations are needed for supervised learning approaches, such as object detection. Synthetic datasets with automatically generated annotations is being explored as one way to overcome these challenges.

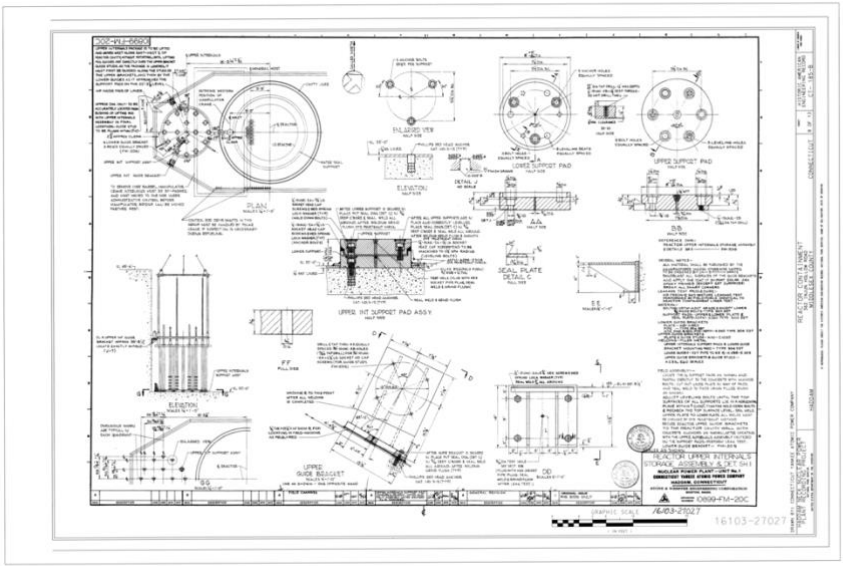


Figure 6: Mid 1900s engineering drawing of a reactor containment building and its internals. Source: Connecticut Yankee Atomic Power Company, creator, Public domain, via Wikimedia Commons.

In the last two years, research has also been published on ED digitization using vision capable multi-modal LLMs like GPT-4o and Claude 3 Opus. In one such study [62], the authors conclude that even the best performing model, GPT-4o, is still significantly limited when it comes to visual ED analysis.

5 AI in the Nordic Nuclear Industry: Current Initiatives and Future Needs

The interviewed NPP licensees have all been working with various types of AI over the years, including NLP (pre-ChatGPT) and computer vision models, for many different applications. Based on the interviews, we here briefly discuss some of their past and ongoing NLP and computer vision projects, as well as their future needs with respect to applications in these areas.

5.1 LLM PROJECTS IN THE NORDIC NUCLEAR INDUSTRY

Some of the interviewed NPPs have already started working with generative LLMs. Due to security reasons, installing local instances of LLMs is something they are looking into actively, but the hardware requirements make this challenging from a cost perspective, as discussed in Sec. 6.

One of the licensees is currently testing an offline, locally hosted small open weights LLM for document processing, but it can only handle a few documents at a time, which limits its usefulness. Two NPPs are evaluating Microsoft Copilot (a chatbot based on a licensed version of OpenAI's GPT-4) for the same purpose, but at least one of them is only allowed to do so on public data. The experiences from the Copilot evaluations have generally been positive.

Two of the interviewed NPPs stated that they are using ChatGPT and Copilot for programming; one example given was Visual Basic code generation for processing of Excel spreadsheets.

Lastly, one NPP mentioned that they have previously developed a document retrieval system based on a combination of BERT and clustering algorithms, for retrieval of fuel damage incident reports, to narrow down the number of reports that then have to be manually searched. However, this system is not currently in use.

5.2 LARGE LANGUAGE MODELS: NEEDS AND INTERESTS

The use cases for large language models in the nuclear industry are, as discussed in Sec. 4.1, numerous and potentially very impactful. This was also reflected in the interviews, which focused heavily on LLMs.

All interviewed NPP licensees are interested in and see a need for LLMs, but virtually all applications they would like to explore require that the LLMs can access classified data.

Since such data must not leave the NPP premises, this excludes the use of closed source models, which are accessible only via the cloud. As mentioned in the previous section, some of the licensees already have locally hosted, small LLMs up and running for experimentation, but both they and the other licensees see a need for more work on how to properly set up the infrastructure for these compute-intensive models, as well as a need to better understand which out of the plethora

of available open LLMs to go for. They also want to understand how they can create LLM-based systems that strictly adhere to their internal regulations that cover how documents should be handled. Operators that do not have clearance to access documents at a certain classification level should never be presented with these documents by the system, or with any content from these documents, other than at most their titles.

The licensees see the potential for many different applications. The central theme during the interviews was the many ways in which LLMs can interact with documents. Most document-related applications mentioned can be mapped to the ones listed in Sec. 4.1.5.

For example, it was brought up that project reports and reports on things that happen in the plant, like incidents and equipment failures, would be great to use for improvements of, e.g., work routines, root cause analyses, and maintenance plans, and for coming up with suggestions for future research projects. The challenge is that the extremely large number of such reports at each plant often makes it difficult to retrieve the relevant ones, and even more so to then read and comprehend their contents and use this comprehension to suggest solutions or potentially useful ideas. A RAG or RAG-like system utilizing the power of LLMs is viewed by the NPP licensees as a promising candidate for meeting this challenge.

5.3 COMPUTER VISION PROJECTS IN THE NORDIC NUCLEAR INDUSTRY

A few of the licensees are working or have worked on projects centered on computer vision. Some of these projects are connected to parts of the licensees' non-NPP related businesses but are still relevant as they address similar challenges as those encountered in NPPs and use NPP-applicable computer vision techniques. In one such project, camera-equipped drones were used for crack detection in a hydroelectric dam; in another, computer vision was used to monitor, count, and classify fish in the fish ladder by a hydroelectric plant.

Other projects are directly about NPPs. One licensee is considering investigating computer vision for crack detection in an ongoing project on long-term underground storage of nuclear waste. To better understand the potential risk of radioactive leakage from the storage facilities, holes are bored into the ground to map out cracks, and this mapping could benefit greatly from full automation using computer vision. In a related project, the task is to use LiDAR to scan the walls of the tunnels leading to the waste storage facilities, and then, potentially, use computer vision to map all rock bolts in the walls.

Moreover, computer vision is currently being used by two licensees to inspect the metal rods containing the nuclear fuel pellets for potential quality issues, and two licensees are using the computer vision-based commercial TrueFlaw system[33] for the inspection of cracks in welds and for gas leakage monitoring. The TrueFlaw system uses both radiology and ultra sonic testing for fault detection, and edge computing (the "TrueflawBox") that works completely disconnected from the internet. In a first field trial with EPRI, conducted in April 2022, it has shown the

[33] https://trueflaw.com/ml

potential for great time savings in the inspection of reactor pressure vessel head penetrations: The system was used to scan the entire dataset (a 7.1km trace) and perform anomaly detection, highlighting all areas in the data requiring more thorough review. This reduced the amount of data requiring manual review by humans to 140m, or 5% of the full data, leading to both time savings and potentially higher fault detection rates, as fatigue on the side of the inspectors from prolonged review of healthy sections is drastically reduced[34].

Lastly, in a rather different type of project, one of the licensees is looking at ways to use computer vision to convert video data to 3D environments for operator training in virtual reality. AI for 2D-to-3D image conversion and novel view synthesis (essentially filling in the empty space between images taken from different spatial viewpoints of the same scene) has become increasingly powerful in recent years, thanks to AI models like Neural Radiance Fields (NeRFs) [63]; however, in the aforementioned project, the investigated technique, Gaussian splatting, which is currently considered state-of-the-art, actually relies on handcrafted algorithms [64].

5.4 COMPUTER VISION: NEEDS AND INTERESTS

One of the licensees is very interested in digitization of their engineering drawings from the 1980s specifically into CAD drawings a topic discussed in Sec. 4.2.2.

Automatic monitoring for component degradation for predictive maintenance is something several licensees see as interesting, and computer vision could be one of possibly several AI technologies to provide such functionality. A specific use case that was mentioned was monitoring of fuel rod bending, which is

Finally, one licensee brought up the fact that some monitoring and inspection they would like to be able to do is dangerous to humans due to potentially high radiation levels, and here drones could be used, in combination with computer vision for automatic analysis of the video streams.

[34] https://www.epri.com/research/products/000000003002025510

6 Pilot Study: Document Discovery with On-Premise AI

This section will outline the pilot study recommended as a follow-up to the AI SNAP project. The request for this pilot study entailed a small target budget, and the objective and scope were designed to meet this target.

6.1 BACKGROUND AND MOTIVATION

Workers in nuclear power plants regularly search through large bodies of internal documents, such as incident and inspection reports, operational and maintenance logs, manuals, and regulatory documents, for example to prepare maintenance procedures. The total amount of these internal documents per power plant is often very large – in the scale of hundreds of thousands or even millions of documents – and ranging between one and several hundred pages in length. An automated way of finding documents related to specific topics or queries could therefore save workers substantial amounts of time spent with tedious search work, outlining a clear economic incentive for license holders to invest in the technology. Moreover, the topic would establish the use of on-premise LLMs in a precisely defined use case, and thus present a basis for later LLM-based AI solutions.

6.2 OBJECTIVE AND SCOPE

The task of finding the most relevant documents in a body of hundreds of thousands is challenging and requires a slightly different approach from typical RAG-based retrieval techniques. Multiple approaches, such as dense and sparse vector approaches, as well as approaches integrating databases, and combinations of the former are possible solutions. At the same time, the details and intricacies of the document searches carried out in nuclear power plants need to be well understood by the research and development team, requiring workshop sessions with licensee personnel actively involved in performing manual document searches.

The aspired project may therefore include work packages around problem definition, integration of AI tools into the operator's workflows, and at its core, AI development work packages. The primary objective of the proposed project is to explore the feasibility of LLM-based document search on a massive scale. The main deliverables would be a proof-of-concept semantic search engine for large bodies of documents (i.e., as a python application), as well as a report outlining the determined technical and user interface-related requirements, the results of the conducted experiments and a description of the solution implemented in the proof-of-concept solution.

6.3 TECHNICAL REQUIREMENTS: LOCALLY HOSTED LLMS

This section discusses hardware requirements for locally hosted LLM solutions – as planned for the outlined project – providing some technical background on this

topic, as well as illustrating various options and expected impact on model performance. This is relevant to the planned pilot project because, as outlined in section 3, any solution targeted at use inside a nuclear power plant will have to be hosted on a local server on the premises of the licensee, either in an office building, or inside the power plant itself. Owing to the technical nature of this topic, a TLDR summary is provided below.

Due to the size of LLMs, hardware requirements for locally hosted LLMs are dominated by the requirement for VRAM (video RAM, i.e., graphic memory) on the host system. While central processor (CPU), system memory (RAM), and storage (SSD) should fulfil certain minimum requirements, the graphics processing unit (GPU) is by far the most decisive factor for inference speed (i.e., speed of the model in production, in tokens or words per second) and the available video memory (VRAM), located on the GPU itself, is in almost all cases the central hardware-related bottleneck for the deployment of LLMs.

To date, most machine learning models, including LLMs, are both trained and run on GPUs, which requires the model and its parameters to be loaded into VRAM. LLMs are typically between 14GB and 800GB large, a result of~3-400 billion parameters stored, by default, at a floating point (FP) precision of 16bit (2 bytes) each. In other words, at identical parameter count and bit rate (FP precision), size differences between different releases (e.g., LLAMA2, LLAMA3, Mistral) will be negligible. A technique called ”offloading”, where only part of a model is loaded into VRAM at a time, allows for working with models larger than the system’s VRAM capacity, but leads to significant reduction in processing speeds, and should only be considered as a last resort.

A popular technique to reduce a model’s size and thus VRAM requirement is quantization. Here, the native 16bit FP precision is getting reduced to lower bit rates, or compression factors (8, 4, 3, 2, or even 1bit). The lower precision of stored model parameters leads to less precise computation and typically deteriorates model performance. However, LLMs process data in highly parallel signal paths, creating computational redundancy and therefore robustness against smaller errors. Moreover, in the process of quantization, various techniques can be used to further lessen the impact of the reduction in precision, and to maintain precision while reducing the required memory space [65]. In practice, even 2bit quantization has been shown to suffer losses of as little as 10-20% in relative task performance [66], while VRAM demand is dramatically reduced, and inference speed is often increased. We therefore favor quantized models for inference tasks for any VRAM-constrained system.

Fine tuning is a technique where the network’s parameters are updated in another round of training after the ”main” training (often referred to as pre-training) has concluded. This is typically used to make the model perform better on a specific task, and/or data within a specific domain. Traditionally, in fine tuning all parameters of a model are optimized. However, this comes with major additional VRAM requirements: The model needs to be represented in its native, typically 16 bit precision to make for a smooth error landscape, and for each trainable parameter the GPU needs to store the parameter itself, its gradient, and – with modern optimizers – moving averages over past gradients and squared gradients,

resulting in four 16 bit values to be stored per parameter, and thus 4x the VRAM-requirement compared to inference runs at the native bit rate. To alleviate this issue and to allow fine tuning even with quantized models, various so-called parameter-efficient fine tuning (PEFT) techniques have been developed in recent years. Typically, they leave the pre-trained model itself largely untouched (or "frozen") during fine-tuning, but insert trainable layers in between the existing ones [67], extend the embedding spaces with additional, trainable parameters [68], [69], or using smaller low rank matrices to approximate parameter updates for much larger weight matrices, accumulating these, and only updating the frozen pre-trained weight matrices once at the end of the process [65]. Overall, while traditional fine-tuning of all parameters yields the highest performance on average, PEFT techniques can often get close while keeping additional VRAM demand moderate.

	LLAMA3-70B 2bit + RAG	**LLAMA3-70B 3bit + RAG**	**LLAMA3-70B 4bit + RAG**	**LLAMA3-70B 16bit + RAG**
VRAM demand	21GB+	30GB+	40GB+	150GB+
Suggested GPU	4090	A6000	A6000	2xA100
GPU VRAM	24GB	48GB	48GB	2x80GB
GPU architecture	Ada Lovelace	Ampere	Ampere	Ampere
GPU release	2022	2020	2020	2020
Hardware price	~50k SEK	~80k SEK	~80k SEK	~500k SEK

Table 4: Overview of LLM quantization steps (for LLAMA3 models) and hardware requirements (as of June 2024).

Finally, Retrieval-Augmented Generation (RAG) techniques make it possible to improve the output of LLMs for specific domains (e.g., a company's internal documentation and laws applicable to the company's business areas), by changing the prompt input to the LLM, instead of the parameters of the LLM itself. For this, a separate language embedding model is used to find semantically related passages in a stack of external documents (e.g., PDFs), then adding these passages to the LLM's input prompt. While this does come with additional VRAM requirements for the embedding model, these models are typically small compared with LLMs, often in the range of 0.5-2GB [70]–[73].

Overall, a reasonable approach for any resource-constrained environment is to use quantized LLMs, add a RAG system if needed for the targeted use case, and evaluate the performance of this system before considering the more resource-intense and effort-laden step of fine-tuning the model.

An overview of exemplary model choices and appropriately matched AI workstation systems (as of September 2024) is provided in 4. This overview is meant to provide a rough point of orientation for the cost factor and options involved in AI workstation systems to date. Note that this overview is expected to be outdated with the arrival of the newest generation of NVidia GPUs and AI workstation cards expected in late 2024 to early 2025.

TLDR; The most important hardware specification for systems running large language models (LLMs) is the graphics card (GPU), and in particular its video memory (VRAM) capacity. LLMs are typically between~6GB and~800GB large but need to fit into VRAM to run at acceptable speed. Compression techniques can reduce the VRAM demand to a quarter of the original size but come with small to moderate reductions in model performance. Fine-tuning and retrieval augmented generation (RAG) techniques can be used to improve LLM performance in specific domains. Exemplary model choices and appropriate hardware configurations for LLMs are provided in Table 4.

Energiforsk is the Swedish Energy Research Centre – an industrially owned body dedicated to meeting the common energy challenges faced by industries, authorities and society. Our vision is to be hub of Swedish energy research and our mission is to make the world of energy smarter.
www.energiforsk.se

Energiforsk

References

[1] K. Schmitt, “Automations influence on nuclear power plants: A look at three accidents and how automation played a role,” *Work*, vol. 41, no. Supplement 1, pp. 4545–4551, 2012.

[2] M. Browne and P. Cook, “Inappropriate trust in technology: Implications for critical care nurses,” *Nursing in Critical Care*, vol. 16, no. 2, pp. 92–98, 2011.

[3] T. Evjemo and S. Johnsen, “Lessons learned from increased automation in aviation: The paradox related to the high degree of safety and implications for future research,” in *29th European Safety and Reliability Conference*, 2019.

[4] K. Pazouki, N. Forbes, R. A. Norman, and M. D. Woodward, “Investigation on the impact of human-automation interaction in maritime operations,” *Ocean engineering*, vol. 153, pp. 297–304, 2018.

[5] H. Muslim and M. Itoh, “A theoretical framework for designing human-centered automotive automation systems,” *Cognition, Technology & Work*, vol. 21, no. 4, pp. 685–697, 2019.

[6] S. M. Merritt, A. Ako-Brew, W. J. Bryant, *et al.*, “Automation-induced complacency potential: Development and validation of a new scale,” *Frontiers in psychology*, vol. 10, p. 225, 2019.

[7] K. Okamura and S. Yamada, “Adaptive trust calibration for human-ai collaboration,” *Plos one*, vol. 15, no. 2, e0229132, 2020.

[8] N. R. Bailey and M. W. Scerbo, “Automation-induced complacency for monitoring highly reliable systems: The role of task complexity, system experience, and operator trust,” *Theoretical Issues in Ergonomics Science*, vol. 8, no. 4, pp. 321–348, 2007.

[9] S. Olaveson, “Automation complacency on humans and cyber-physical systems in the energy sector,” 2023.

[10] M. H. Khalid, P. F. HB, A. Al Rashdan, and Z. Mohaghegh, “Automation trustworthiness in nuclear power plants: A literature review,” in *Probabilistic Safety Assessment and Management (PSAM) International Topical Meeting on Artificial Intelligence (AI) and Risk Analysis*, 2023.

[11] G. A. Boy and K. A. Schmitt, “Design for safety: A cognitive engineering approach to the control and management of nuclear power plants,” *Annals of Nuclear Energy*, vol. 52, pp. 125–136, 2013.

[12] C. R. Kovesdi, Z. A. Spielman, J. D. Mohon, T. M. Miyake, R. A. Hill, and C. Pederson, “Development of an assessment methodology that enables the nuclear industry to evaluate adoption of advanced automation,” Idaho National Lab.(INL), Idaho Falls, ID (United States), Tech. Rep., 2021.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding*, 2019. arXiv: 1810 . 04805 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1810.04805.

[14] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, *Attention is all you need*, 2023. arXiv: 1706.03762 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1706.03762.

[15] X. Wang, M. Salmani, P. Omidi, X. Ren, M. Rezagholizadeh, and A. Eshaghi, "Beyond the limits: A survey of techniques to extend the context length in large language models," *arXiv preprint arXiv:2402.02244*, 2024.

[16] Liu, J. Cao, C. Liu, K. Ding, and L. Jin, "Datasets for large language models: A comprehensive survey," *arXiv preprint arXiv:2402.18041*, 2024.

[17] Q. Dong, L. Li, D. Dai, *et al.*, "A survey on in-context learning," *arXiv preprint arXiv:2301.00234*,
2022.

[18] D. Hendrycks, C. Burns, S. Basart, *et al.*, "Measuring massive multitask language understanding," *arXiv preprint arXiv:2009.03300*, 2020.

[19] A. Dubey, A. Jauhri, A. Pandey, *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[20] N. Muennighoff, L. Soldaini, D. Groeneveld, *et al.*, "Olmoe: Open mixture-of-experts language models," *arXiv preprint arXiv:2409.02060*, 2024.

[21] M. Abdin, S. A. Jacobs, A. A. Awan, *et al.*, "Phi-3 technical report: A highly capable language model locally on your phone," *arXiv preprint arXiv:2404.14219*, 2024.

[22] J. Achiam, S. Adler, S. Agarwal, *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*,
2023.

[23] W. Ali and S. Pyysalo, "A survey of large language models for european languages," *arXiv preprint arXiv:2408.15040*, 2024.

[24] R. Luukkonen, J. Burdge, E. Zosa, *et al.*, "Poro 34b and the blessing of multilinguality,"
arXiv preprint arXiv:2404.01856, 2024.

[25] V. Karpukhin, B. Oˇ guz, S. Min, *et al.*, *Dense passage retrieval for open-domain question*
answering, 2020. arXiv: 2004.04906 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2004.04906.

[26] Y. Gao, Y. Xiong, X. Gao, *et al.*, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, 2023.

[27] M. Fatehkia, J. K. Lucas, and S. Chawla, "T-rag: Lessons from the llm trenches," *arXiv preprint arXiv:2402.07483*, 2024.

[28] W. Fan, Y. Ding, L. Ning, *et al.*, "A survey on rag meeting llms: Towards retrieval-augmented large language models," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 6491–6501.

[29] T. Procko, "Graph retrieval-augmented generation for large language models: A survey," *Available at SSRN*, 2024.

[30] M. Anwar, M. de Costa, I. Hammad, and D. Lau, "Evaluating chatgpt on nuclear domain-specific data," *arXiv preprint arXiv:2409.00090*, 2024.

[31] Y. Gao, Y. Xiong, M. Wang, and H. Wang, "Modular rag: Transforming rag systems into lego-like reconfigurable frameworks," *arXiv preprint arXiv:2407.21059*, 2024.

[32] S. Barnett, S. Kurniawan, S. Thudumu, Z. Brannelly, and M. Abdelrazek, "Seven failure
points when engineering a retrieval augmented generation system," in *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI*, 2024, pp. 194–199.

[33] D. Ru, L. Qiu, X. Hu, *et al.*, "Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation," *arXiv preprint arXiv:2408.08067*, 2024.

[34] Y. Talebirad and A. Nadiri, "Multi-agent collaboration: Harnessing the power of intelligent llm agents," *arXiv preprint arXiv:2306.03314*, 2023.

[35] X. Li, S. Wang, S. Zeng, Y. Wu, and Y. Yang, "A survey on llm-based multi-agent systems: Workflow, infrastructure, and challenges," *Vicinagearth*, vol. 1, no. 1, p. 9, 2024.

[36] T. Xie, D. Zhang, J. Chen, *et al.*, "Osworld: Benchmarking multimodal agents for open-
ended tasks in real computer environments," *arXiv preprint arXiv:2404.07972*, 2024.

[37] A. J. Dave, T. N. Nguyen, and R. B. Vilim, "Integrating llms for explainable fault diagnosis in complex systems," *arXiv preprint arXiv:2402.06695*, 2024.

[38] S. Kernan Freire, C. Wang, M. Foosherian, S. Wellsandt, S. Ruiz-Arenas, and E. Niforatos, "Knowledge sharing in manufacturing using llm-powered tools: User study and model benchmarking," *Frontiers in Artificial Intelligence*, vol. 7, p. 1 293 084, 2024.

[39] S. Fuchs, M. Witbrock, J. Dimyadi, and R. Amor, "Using large language models for the interpretation of building regulations," *arXiv preprint arXiv:2407.21060*, 2024.

[40] A. Berger, L. Hillebrand, D. Leonhard, *et al.*, "Towards automated regulatory compliance verification in financial auditing with large language models," in *2023 IEEE International Conference on Big Data (BigData)*, IEEE, 2023, pp. 4626–4635.

[41] T. Nakaura, N. Yoshida, N. Kobayashi, *et al.*, "Preliminary assessment of automated radiology report generation with generative pre-trained transformers: Comparing results to radiologist-generated reports," *Japanese Journal of Radiology*, vol. 42, no. 2, pp. 190–200, 2024.

[42] G. Colverd, P. Darm, L. Silverberg, and N. Kasmanoff, "Floodbrain: Flood disaster reporting by web-based retrieval augmented generation with an llm," *arXiv preprint arXiv:2311.02597*, 2023.

[43] M. L. Bernardi, M. Cimitile, and R. Pecori, "Automatic job safety report generation us-
ing rag-based llms," in *2024 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2024, pp. 1–8.

[44] M. Raatikainen, T. M¨ annist¨ o, T. Tommila, and J. Valkonen, "Challenges of requirements engineering—a case study in nuclear energy domain," in *2011 IEEE 19th International Requirements Engineering Conference*, IEEE, 2011, pp. 253–258.

[45] S. Myllynen *et al.*, "Utilization of artificial intelligence in the analysis of nuclear power
plant requirements," M.S. thesis, 2019.

[46] H. A. Gohel, H. Upadhyay, L. Lagos, K. Cooper, and A. Sanzetenea, "Predictive maintenance architecture development for nuclear infrastructure using machine learn-

ing," *Nuclear Engineering and Technology*, vol. 52, no. 7, pp. 1436–1442, 2020.

[47] M. Dong, H. Huang, and L. Cao, "Can llms serve as time series anomaly detectors?" *arXiv preprint arXiv:2408.03475*, 2024.

[48] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[49] A. Vijayakumar and S. Vairavasundaram, "Yolo-based object detection models: A review and its applications," *Multimedia Tools and Applications*, pp. 1–40, 2024.

[50] A. Kirillov, E. Mintun, N. Ravi, *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF*
International Conference on Computer Vision, 2023, pp. 4015–4026.

[51] N. Ravi, V. Gabeur, Y.-T. Hu, *et al.*, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.

[52] S. Akcay, D. Ameln, A. Vaidya, B. Lakshmanan, N. Ahuja, and U. Genc, "Anomalib: A deep learning library for anomaly detection," in *2022 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2022, pp. 1706–1710.

[53] S. Zhao, R. Quan, L. Zhu, and Y. Yang, "Clip4str: A simple baseline for scene text recognition with pre-trained vision-language model," *arXiv preprint arXiv:2305.14014*, 2023.

[54] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, PMLR, 2021, pp. 8748–8763.

[55] S.-H. Park, B.-H. Won, and S.-K. Ahn, "Safeguards-related event detection in surveillance video using semi-supervised learning approach," *Nuclear Engineering and Technology*, 2024.

[56] M. Thomas, A. Pollack, R. Hofman, S. Rocchi, M. John, and M. Moeslinger, "Tracking spent fuel movements with a modular deep learning system for enhanced efficiency of safeguards surveillance data review,"

[57] A. Pandat, P. Rajasekhar, G. Aravamuthan, G. Joseph, R. Shukla, and G. Vinod, "Role of ai in anti-drone systems: A review," in *International Conference on Reliability, Safety, and Hazard*, Springer, 2024, pp. 29–39.

[58] J. Yu, Y. Xu, C. Xing, J. Zhou, and P. Pan, "Pixel-level crack detection and quantification
of nuclear containment with deep learning," *Structural Control and Health Monitoring*, vol. 2023, no. 1, p. 9 982 080, 2023.

[59] F. Li, B. Zhang, B. Zhang, *et al.*, "Implementation of surface crack detection method for nuclear fuel pellets by weakly supervised learning," *Journal of Nuclear Science and Technology*, pp. 1–12, 2024.

[60] M. Liinasuo, T. Passi, and S. Pakarinen, "Working with senses-visual inspection in a nuclear power plant," in *Proceedings of the European Conference on Cognitive Ergonomics 2024*, 2024, pp. 1–4.

[61] L. Jamieson, C. Francisco Moreno-Garc´ıa, and E. Elyan, "A review of deep learning methods for digitisation of complex documents and engineering diagrams," *Artificial*

Intelligence Review, vol. 57, no. 6, pp. 1–37, 2024.

[62] A. C. Doris, D. Grandi, R. Tomich, M. F. Alam, H. Cheong, and F. Ahmed, "Designqa: A multimodal benchmark for evaluating large language models' understanding of engineering documentation," *arXiv preprint arXiv:2404.07917*, 2024.

[63] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[64] B. Kerbl, G. Kopanas, T. Leimk¨ uhler, and G. Drettakis, "3d gaussian splatting for real-
time radiance field rendering.," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.

[65] E. J. Hu, Y. Shen, P. Wallis, *et al.*, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[66] Y. Xu, L. Xie, X. Gu, *et al.*, "Qa-lora: Quantization-aware low-rank adaptation of large language models," *arXiv preprint arXiv:2309.14717*, 2023.

[67] B. Newman, P. K. Choubey, and N. Rajani, "P-adapters: Robustly extracting factual in-
formation from language models with diverse prompts," *arXiv preprint arXiv:2110.07280*, 2021.

[68] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," *arXiv preprint arXiv:2104.08691*, 2021.

[69] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," *arXiv preprint arXiv:2101.00190*, 2021.

[70] Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., ... & Yih, W. T. (2020). Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

[71] Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[72] Reimers, N. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv preprint arXiv:1908.10084*.

[73] Sanh, V. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

[74] Meta AI, "Llama Use Policy," [Online]. Available: https://ai.meta.com/llama/use-policy/ [Accessed: Feb. 27, 2025].

[75] Open Source Initiative, "The MIT License," [Online]. Available: https://opensource.org/licenses/MIT [Accessed: Feb. 27, 2025].

[76] Apache Software Foundation, "Apache License, Version 2.0," [Online]. Available: https://www.apache.org/licenses/LICENSE-2.0 [Accessed: Feb. 27, 2025].

ON-PREMISE AI SOLUTIONS FOR NORDIC NUCLEAR APPLICATIONS

This report explores the implementation of on-premise AI solutions in the Nordic nuclear energy industry. It highlights the potential of natural language processing and computer vision technologies to enhance efficiency, safety, and decision-making in nuclear power plants. It addresses key challenges such as data handling and security, along with promising applications of large language models and retrieval-augmented generation (RAG) systems. The report also reviews ongoing AI initiatives and proposes a pilot study for developing a semantic search engine. This comprehensive analysis provides valuable insights for future AI projects in the nuclear sector..

Ett nytt steg i energiforskningen

Forskningsföretaget Energiforsk initierar, samordnar och bedriver forskning och analys inom energiområdet samt sprider kunskap för att bidra till ett robust och hållbart energisystem. Energiforsk är ett politiskt neutralt och icke vinstutdelande aktiebolag som ägs av branschorganisationerna Energiföretagen Sverige och Energigas Sverige, det statliga affärsverket Svenska kraftnät, samt gas- och energiföretaget Nordion Energi. Läs mer på energiforsk.se.

Energiforsk