Enhanced LLM for smart Knowledge Management in nuclear industry

Frédéric Godest[1], Mouna El Alaoui[1], Victor Richet[1], Robert Plana[1], Lies Benmiloud-Bechet[1], Jean-François Bossu[1], Olivier Malhomme[1]

[1] *Assystem Engineering and Operation Services, 9/11 Allée de l'Arche, 92610 Courbevoie, France*

Abstract

The nuclear industry, characterized by its large-scale projects and intricate processes, requires robust knowledge management (KM) strategies. Traditional KM approaches, such as training, documentation, and expert networks, have been employed to address this need . Documentation, though, has challenged by its volume, the outcome of this approach. However, the advent of AI and Large Language Models (LLMs) has opened new avenues for KM innovation. This paper explores the integration of CurieLM, a domain-specific LLM, with a nuclear ontology to improve the quality of knowledge retrieval and answers generation. By automatically expanding the input context through ontology-driven enrichment, our approach aims to address the shortcomings of existing KM methods, offering a scalable and efficient solution for the nuclear industry's unique challenges.

Keywords

Large Language Model, Ontology, Knowledge Management, Nuclear Industry

1. Introduction

Nuclear power plants are highly knowledge-intensive facilities [1] structured by long-term projects involving multiple disciplines and complex technologies. Alongside this growing complexity, the nuclear domain is facing actual challenges of growing decarbonization demand and replacement of the aging nuclear fleets [2].   The lifespan of power plants also stresses the importance of integration of heterogenous knowledge means, which can go back to 50 years. In order to meet the resulting high expectations and requirements of this critical domain, capturing and capitalizing on knowledge plays a crucial role. In fact, Knowledge Management (KM) has been identified as one of the key enabling discipline for distributed engineering enterprises such as nuclear power plant projects in the 21st century [1].

Knowledge Management is defined by [3] as the practice of selectively applying knowledge from previous experiences of decision making to current and future decision-making. To successfully achieve this knowledge capitalization and application in the future decisions, it is essential to consider the capture, storage, retrieval, and reuse of knowledge [3]. One of the main KM's objectives is providing appropriate information for the appropriate resource at the right time [1][4] which makes it essential in the context of the nuclear project.

Recent advancements in data-related digital technologies, particularly Artificial Intelligence (AI) led to the development of Large Language Models (LLMs). Training LLMs on numerous, diverse texts results in the integration of extensive knowledge, interpretation of complex information, general reasoning and aiding knowledge-intensive decision-making [5]. The recent evolution of LLMs have enabled the creation of Generative AI, a technology can understand and generate human-like text. Few experiments have been made in the nuclear domain, NukeBERT [5] is a pre-trained language

EKAW 2024: EKAW 2024 Workshops, Tutorials, Posters and Demos, 24th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2024), November 26-28, 2024, Amsterdam, The Netherlands.

fgodest@assystem.com (F. Godest); melalaoui-ext@assystem.com (M. El Alaoui); vrichet@assystem.com (V. Richet); rplana@assystem.com (R. Plana); lbenmiloud@assystem.com (L. Benmiloud-Bechet); jfbossu@assystem.com (J.-F. Bossu) ; omalhomme@assystem.com (O. Malhomme)

0009-0007-0395-5207 (V. Richet); 0000-0003-2378-8331 (R. Plana)

model based on BERT [6] that has exhibited significant performance improvements over the original BERT. Another is NuclearQA [7] that introduced a human-made benchmark aimed at assessing language models in the nuclear domain.

2. LLM limitations in Knowledge Management

Despite their undeniable benefits, the responses generated by LLMs present two main limitations: (i) outdated information potentially originating from the model's training date, and (ii) inaccuracies in factual representation, also known as "hallucinations" [8]. LLM 'hallucinations' designates the LLM's generation of incorrect results and answers. It is either due to the fact that the requested information has not been retrieved, or the fact that LLM tries to provide a creative answer.

Additionally, hardware is also an important limitation regarding LLM, especially in the nuclear industry where, most of the time, a local LLM is required so that the confidentiality and the security of the sensitive information is guaranteed. This constraint induces purchasing and installing Graphics Processing Unit (GPU) to specific servers. These GPU are rather expensive due to their use of more electrical resources than a regular computer. It also echoes on the IT infrastructure, such as installing GOPUs.

These limitations must be addressed to achieve a KM tool with a reliable and energy-efficient LLM developed for use in the nuclear industry.

3. Addressing LLMs limitations

This work aims to mitigate the previously described limitations through the following approaches.

LLM generated answers quality:

1. Larger LLM could prevent LLM hallucinations since it has much more trained parameters and could also better "*understand*" the questions and the overall context. Larger LLM also need more computer power to operate and, sometimes, the necessary power can only be provided by third party companies (ex: OpenAI, Microsoft, AWS, etc.).
2. Retrieval-Augmented Generation (RAG) architecture is expected to provide precise context to the LLM when the documentary corpus quality is adequate. RAG architecture can work with "*small*" LLM (7b to 14b parameters) and can provide an equivalent quality of generated answer as larger LLM.
3. Fine tuning a LLM enables improvement of the generated answers in a specific domain. In this case, training instructions related to nuclear domain has been provided to a LLM for its fine tuning.
4. Increasing the context sent to the LLM; The more the LLM has context, the better. This context can be increased by prompt engineering

This last aspect is the one addressed in our proposal. The idea is to use ontology-enriched request, or in other words, on the basis of an 'elementary' request, expand it using ontology. This means on the basis of the concepts used in the request, use the closest concepts in the ontology as concept to further enhance quality of the answers. This approach has not been implemented for the generation of the results below.

LLM energy efficiency:

1. The best way to reduce the energy consumption is to get a "*tiny*" LLM (<7b parameters). But this solution also comes with a downside; Smaller LLM are in fact less "intelligent" and will, by definition, provide poorer quality responses.
2. A common way to reduce energy consumption without drastically degrading the quality of the generated answer is to get a "*small*" LLM (7b to 14b parameters) and to quantize it. LLM quantization is a way of compressing LLM by changing the variables type of the parameters (ex: from float to int8) so it will need less memory to operate

Those two approaches were successfully implemented in the previous version of CurieLM, enabling significant increase of the performance, most noticeably the time to generate an answer.

3. Lastly, LLM are commonly used within Python projects. Some of the LLM have been redesigned so they can work on C++ program which is way more efficient. The major downside of this method is the integration within existing tools and larger projects since, nowadays, the community is mainly working on Python.

These solutions highlight the fact that there is a trade-off between computer power and generated answers quality but some of the technical levers might be used to optimize this trade off to our need of efficiency.

4. Methodology

Our first approach consists in combining a fine-tuned "small" LLM (7b parameters) which is then quantized within a RAG architecture implemented in a Knowledge Management solution. We called this project CurieLM (which stands for Marie Curie). This model has been fine-tuned over 25 000 instructions from internal dataset and open-source dataset related to nuclear domain.

Models have been compared on 5 main characteristics, which have been human-evaluated by a pool of experts. Those characteristics are described below:

- Accuracy: Capability of the model to provide an accurate answer, without using vague words, or concepts
- Synthesis: Capability of the model to provide a short answer, expressed as much as possible in the shortest way possible
- Quality: Capability of the model to provide an answer using the relevant elements, up to date
- Exhaustivity: Capability of the model to answer to the whole question, without shadowing or omitting some aspects of the question
- Clarity: Capability of the model to provide structured and organised answer, easy to read and understand

5. Results

We tested GPT4 and our CurieLM RAG architecture through 56 technical questions related to nuclear safety. After human evaluation, we observed a significant augmentation of the quality of the answer (higher notes), and a constant consistency (lower standard deviation) with the RAG technology. This evaluation was specifically performed on the exhaustivity, accuracy, overall quality, clarity and synthesis of the answer (Figure 1) which were expert-evaluated.

We also tested our fine-tuned CurieLM (with more than 23 000 instructions) model with a custom benchmark dataset related to nuclear domain to compare with Mistral 7b (our base model) and other models like GPT4. (Figure 2). This comparison will be evaluated thanks to evaluation dataset composed of questions and answers related to nuclear domain. Due to data safety and confidentiality constraints, this datasets cannot be disclosed in details but are described in [10]. Those datasets have been built under the form of Q/A sets defined by internal experts group.

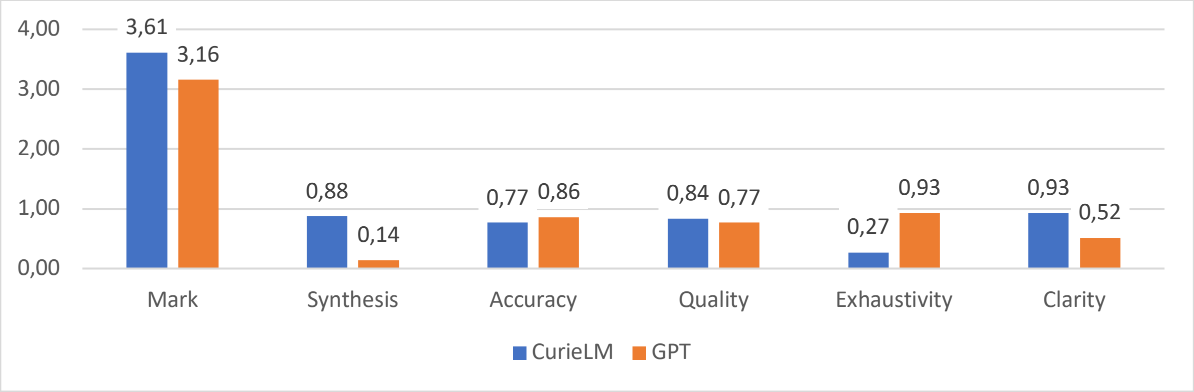

Figure 1: Score comparison between GPT4 and RAG architecture.

Explanation for the KPI used in this chart is provided in paragraph 4. For each question of the dataset considered, every generated answer will be then compared to the correct answer in the dataset. The ratio between the correct and incorrect generated answer will give us the accuracy of the model in Figure 2.

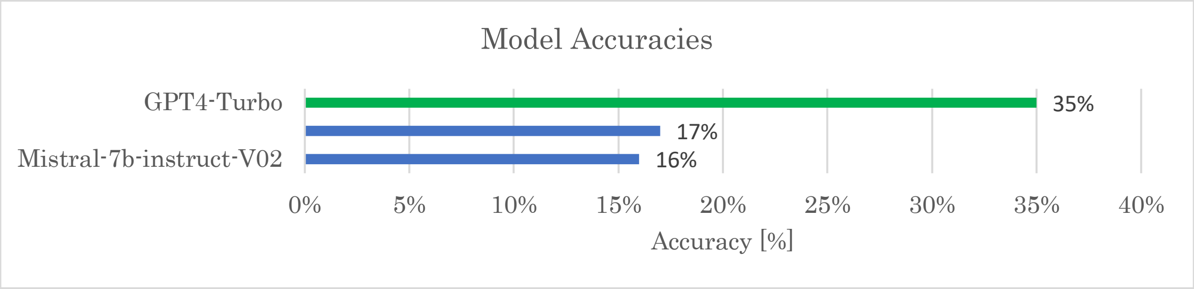

Figure 2: Model accuracies comparisons

We can observe that GPT4 provided correct answers only with 35% of accuracy even though GPT4 is one of the largest LLM available on the market. We can also observe that Mistral 7b (our base model) provided correct answers with 16% of accuracy and, when fine-tuned, this accuracy is slightly better with only 17% of accuracy.

These results highlight the fact that for very specific domains, LLMs have not great accuracy in terms of answers quality. We can draw another graphic if we define LLM efficiency as such:

$$LLM_{efficiency} = \frac{LLM_{accuracy}}{LLM_{nb\_parameters}} \qquad (1)$$

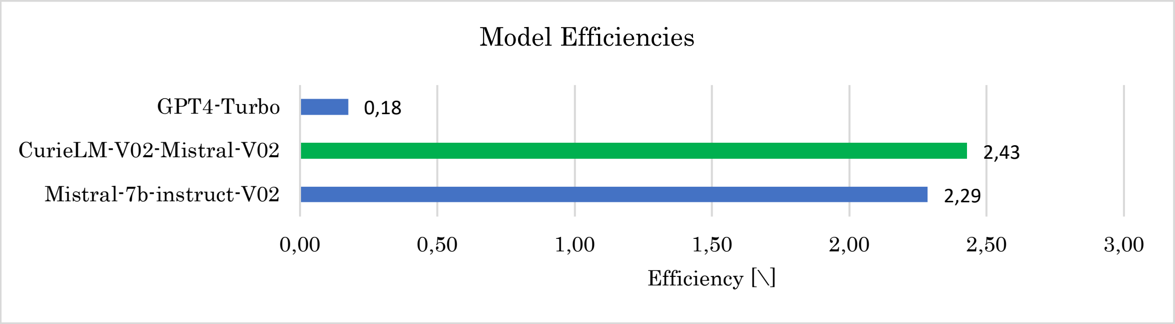

Figure 3: Model efficiencies comparisons

We can now observe that our CurieLM model is far more efficient than GPT4 and significantly more efficient that Mistral 7b (our base model). Please note that for this chart we assume that GPT4 have 200b parameters which is a low estimation. One key point in the explanation of this higher efficiency is to notice that even though GPT provides answer with higher accuracy, it does so using a much larger number of parameters, leading to an overall ratio being lower. In other words, CurieLM presents a higher 'per parameter' accuracy. We can go further in this improvement by increasing the input context sent to our CurieLM model. This input context can be dynamically modified with user information and with a nuclear ontology which will enrich the context of the question. This context enrichment by ontology will consist of extracting keywords from the question to filters relevant branch and nodes to a nuclear ontology (Table 1).

Table 1

Question (CurieLM model)	Question + RAG (CurieLM model + KG tool)	Question + RAG + user information. (CurieLM model + KG tool)	Question + RAG + user information + ontology enrichment. (CurieLM model + KG tool)
What role does civil engineering play in the optimization of thermal efficiency in a nuclear power plant?	What role does civil engineering play in the optimization of thermal efficiency in a nuclear power plant? Context: [document 1, document 2, document 3]	I'm a civil engineer working in nuclear industry for an electrical French provider. What role does civil engineering play in the optimization of thermal efficiency in a nuclear power plant? Context: [document 1, document 2, document 3]	I'm a civil engineer working in nuclear industry for an electrical French provider. What role does civil engineering play in the optimization of thermal efficiency in a nuclear power plant? Similar concepts: Reactor Core Design, Heat Transfer Systems, Cooling System Design, Waste Heat Recovery. Context: [document 1, document 2, document 3]

This method will automatically increase the input context without changing the user's experience. Larger input context will significantly improve the quality of the generated answer.

6. Conclusion

The CurieLM project has shown that operating a fine tuning of a "small" LLM, by using a RAG architecture and by extending the input context sent to a LLM might be an interesting way to optimize the trade-off between computer power and generated answers quality. In other words, in a computing resource-constrained paradigm which is often the case in engineering, using fine-tuning constitutes a way to further improve overall quality, and accuracy without increasing the computing power.

The overall results of these earlier stage results are quite promising but can be improved by:

1. Improving the fine-tuning of the CurieLM model to increase the accuracy and, by extension, the efficiency of the model.
2. Improving the RAG architecture with a nuclear oriented Knowledge Graph.
3. Improving the ontology enrichment of the input prompt so it can give more relevant context for the RAG architecture to find relevant documents.

References

[1] Minglu Wang, Mingguang Zheng, Lin Tian, Zhongming Qiu, Xiaoyan Li, (2017) *A full life cycle nuclear knowledge management framework based on digital system*, Annals of Nuclear Energy, Volume 108, 2017, Pages 386-393, ISSN 0306-4549, https://doi.org/10.1016/j.anucene.2017.04.047.

[2] Samuel Carrara (2020) Reactor ageing and phase-out policies: global and regional prospects for nuclear power generation, Energy Policy, Volume 147, 2020, 111834, ISSN 0301-4215, https://doi.org/10.1016/j.enpol.2020.111834.

[3] Jennex, M. E. (2005). What is KM? International Journal of Knowledge Management, 1(4), i-iv.

[4] Zhao, J., Pablos, P.O.D., Qi, Z., 2012. Enterprise knowledge management model based on China's practice and case study. Comput. Hum. Behav. 28 (2), 324–330.

[5] Jain A., Meenachi D.N., Venkatraman D.B. (2020), arXiv preprint arXiv:2003.13821

[6] Devlin J., Chang M.W., Lee K., Toutanova K. (2018), arXiv preprint arXiv:1810.04805

[7] Acharya A., Munikoti S., Hellinger A., Smith S., Wagle S., Horawalavithana S. (2023), arXiv preprint arXiv:2310.10920

[8] Kernan Freire, S., Wang, C., Foosherian, M., Wellsandt, S., Ruiz-Arenas, S., & Niforatos, E. (2024). Knowledge sharing in manufacturing using LLM-powered tools: user study and model benchmarking. *Frontiers in Artificial Intelligence*, *7*, 1293084.

[9] Bouhoun, Z., Allali, A., Cocci, R., Assaad, M. A., Plancon, A., Godest, F., ... & Plana, R. (2024). CurieLM: Enhancing Large Language Models for Nuclear Domain Applications. In EPJ Web of Conferences (Vol. 302, p. 17006). EDP Sciences.