EPJ Web of Conferences **302**, 17006 (2024) https://doi.org/10.1051/epjconf/202430217006
SNA + MC 2024

CurieLM: Enhancing Large Language Models for Nuclear Domain Applications

Zakaria Bouhoun, *Ahmed* Allali, *Riccardo* Cocci, *Mohamad Ali* Assaad[1,*], *Alexandra* Plancon, *Frederic* Godest, *Kirill* Kondratenko, *Julien* Rodriguez, *Francesco* Vitillo, *Olivier* Malhomme, *Lies* Benmiloud Bechet[2,**], and *Robert* Plana[3,***]

[1]Assystem EOS, 92400 Courbevoie, France

Abstract. Large Language Models (LLMs), such as the Mistral model, have exhibited remarkable performance across diverse tasks. However, their efficacy in nuclear applications remains constrained by a lack of domain-specific knowledge and an inability to effectively leverage that knowledge. Nuclear-related tasks, including safety assessments and requirement analyses, pose unique challenges due to the intricate domain expertise and diverse constraints involved. To address these limitations, we introduce CurieLM, an LLM specifically tailored for the nuclear domain. CurieLM builds upon the Mistral model, enhancing its capabilities through domain-specific fine-tuning. Our team of nuclear engineers overcame the initial hurdle of accessing high-quality nuclear data, enabling CurieLM to comprehend and accurately respond to nuclear-specific instructions. This manuscript outlines the development and optimization process of CurieLM, marking a significant step toward enhancing nuclear-related natural language processing tasks. Experimental results demonstrate a 13% performance improvement over base LLMs, underscoring the effectiveness of our approach. Domain-specific LLMs like CurieLM hold a great potential across various applications, and this study sets the stage for further exploration in this emerging field.

1 Introduction

In the rapidly evolving landscape of Natural Language Processing (NLP) and artificial intelligence, Large Language Models have emerged as transformative tools, reshaping the way machines understand and generate human-like text. Pioneered by companies such as OpenAI with models like GPT-3.5 [1] and GPT-4 [2], as well as the availability of open-source alternatives like Mistral [3] and Llama [4], these LLMs signify a monumental leap in computational linguistics.

These models are built upon a unique paradigm. It involves pre-training on extensive corpora and then fine-tuning the model to follow specific instructions [5]. Instruction fine-tuning is a process that helps the model better understand and follow the instructions given in the input. It involves further training the model on a smaller, more specific dataset containing examples of the kind of instructions the model will need to follow.

*e-mail: maassaad@assystem.com
**e-mail: lbenmiloud@assystem.com
***e-mail: rplana@assystem.com

Another key component of this paradigm is Reinforcement Learning from Human Feedback (RLHF) [6]. RLHF is a method used to improve the model's performance. After the model generates a response, humans rate the quality of the response, the model then uses this feedback to learn and improve, reinforcing the behaviours that led to good responses and discouraging those that led to poor responses.

LLMs have become behemoths with billions of parameters. Their unprecedented scale enables them to exhibit human-like traits, including step-by-step reasoning, coherent communication, and nuanced contextual interpretation. This prowess has found applications across diverse domains, from healthcare [7, 8] to finance [9, 10] and law [11, 12], demonstrating their utility and impact in tackling complex tasks.

Despite these advancements, challenges persist when applying LLMs to specific domains, particularly in fields with stringent requirements like nuclear engineering. General-purpose LLMs, trained on diverse internet data, may lack the requisite understanding of nuclear engineering concepts, potentially leading to inaccuracies or misinterpretations.

These challenges are compounded by concerns regarding data security, privacy, and the need for regulatory compliance. Nuclear data, highly sensitive and subject to strict regulations, presents a formidable barrier to the adoption of external LLM services. The use of such services might expose this data to third parties, posing significant risks to data security and privacy.

Moreover, in safety-critical industries like nuclear engineering, the reliability of information provided by LLMs is paramount. Trust in the model's responses becomes challenging if it makes errors or provides inaccurate information.

To address these challenges, we introduce CurieLM, a domain-specific LLM designed for the nuclear field. CurieLM is a novel contribution to the field. However, it is worth noting that its predecessor, CurieGPT, has been briefly discussed for a particular use case for requirement management during IAEA conference [13], providing initial insights into the application of LLMs in nuclear safety, setting the stage for the more advanced and specialized CurieLM.

CurieLM is currently an R&D effort within Assystem, progressing at the prototype stage between 4 & 5 Technology Readiness Level (TRL). We are validating its applicability through increasing industrial specific use cases, aiming for it to become an industry-ready tool.

Designed as a standalone tool, CurieLM can interoperate seamlessly via API with existing software such as MBSE engineering software. This interoperability ensures CurieLM can complement and enhance current engineering tools.

Our approach involves creating a high-quality dataset sourced from reputable nuclear authorities, validated by domain experts to ensure accuracy and relevance. The model, built upon Mistral-7B, undergoes fine-tuning using instruction tuning techniques to enhance its performance on nuclear-related tasks. In summary, our contributions include:

- Data Framework: We present a robust framework for constructing high-quality datasets tailored specifically to the nuclear domain.
- Dataset Validation: A curated dataset, meticulously vetted by nuclear experts, forms the basis of CurieLM's fine-tuning.
- LLM Fine-Tuning: By fine-tuning Mistral on nuclear data, we demonstrate a 13% improvement in its performance on tasks related to the nuclear field, emphasizing the efficacy of domain-specific fine-tuning.

Our work on CurieLM represents a significant step towards leveraging the power of LLMs in the nuclear field. By addressing the unique challenges of this domain, we aim to provide a tool that can help nuclear engineers improve the quality and security of their projects.

This paper will delve into the specifics of our approach, detailing the rigorous process of data collection and preparation, the process of instruction tuning, and the evaluation of

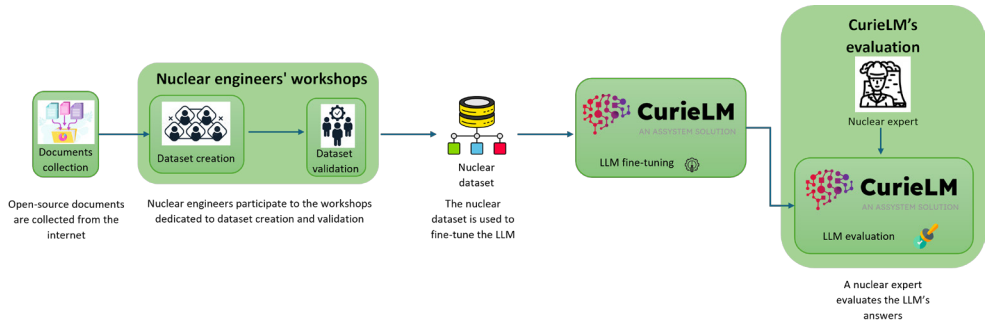

Figure 1. CurieLM Overview

CurieLM's performance. We believe that our work sets the stage for further exploration in this emerging field and underscores the potential of domain-specific LLMs in complex fields like nuclear engineering.

2 Related Work

Large Language Models have demonstrated immense potential in automating various tasks across different domains. OpenAI's proprietary models, such as GPT-3.5 [1] and GPT-4 [2], have set the benchmark for their versatility in understanding and generating human-like text. These models have been pivotal in automating a wide array of tasks, showcasing the transformative power of LLMs.

However, the landscape of LLMs extends beyond proprietary models. There are notable alternatives that offer powerful capabilities and permissive licenses. For instance, Mistral [3] and Llama [4] are two such models that have made significant contributions to the growing ecosystem of accessible and effective language models.

The research focus has recently shifted towards domain-specific fine-tuned models. Instead of using generic LLMs like GPT-4 or Mistral, several teams have started developing LLMs that are fine-tuned on specific domains to enhance their performance on tasks related to those domains.

In the realm of finance, BloombergGPT [9] stands out as an LLM that has been pre-trained from scratch on financial data. Another noteworthy model is the one proposed by [14], which continues the training on financial data after transforming the data into a specific format starting from a pre-trained LLM. FinGPT [10] is another LLM that has been fine-tuned on instruction data and employs RLHF to understand and adapt to individual preferences.

Several LLMs have been developed for medical tasks. Med-PaLM-2 [15], a renowned model, combines a base LLM, medical instruction fine-tuning, and prompting strategies (Prompting strategies refer to the methods used to instruct the LLMs to perform specific tasks. For example, if the task is to answer a medical question, the prompt might be a question phrased in a certain way that guides the model to provide a detailed and accurate answer), achieving human expert level in answering USMLE-style questions. Clinical Camel [7], an open-source LLM, excels in clinical research, while ChatDoctor [8], adapted from the LLaMA, provides accurate medical advice using patient-doctor dialogues.

In the legal domain, ChatLaw [11] is an open-source legal LLM capable of generating legal documents and performing legal tasks based on integrated external knowledge bases.

LexGPT [12] is another LLM fine-tuned using Pile of Law dataset, making it specialized for the legal domain.

In the nuclear domain, NukeBERT [16] is a pre-trained language model based on BERT [17] and tailored for tasks within domains with limited datasets, such as nuclear science. It has exhibited significant performance improvements over the original BERT baseline during evaluations. In addition, NuclearQA [18] introduces a human-made benchmark aimed at assessing language models in the nuclear domain. Through this benchmark, a discernible gap in the scientific knowledge of current LLMs has been unveiled, highlighting the necessity for more specialized models in this field.

These domain-specific LLMs have shown significant potential in their respective fields, demonstrating the effectiveness of fine-tuning LLMs on domain-specific data. This sets the stage for the development of CurieLM, an LLM specifically tailored for the nuclear domain.

3 Datasets

In this section, we detail the rigorous process undertaken to build and prepare our dataset, designed specifically for fine-tuning and evaluating the performance of LLMs within the nuclear field. Our methodology was guided by three key principles: ensuring high-quality data, maintaining domain specificity, and involving domain experts in the process with a constant focus on copyright and confidentiality compliance.

- High-Quality Sources: We sourced our data from reputable and high-ranked non copyrighted documents within the nuclear domain (e.g. documents published by International Atomic Energy Commission, French Institute for Radiation Protection and Nuclear Safety, etc.). This ensured that the information used for fine-tuning our LLMs was not only relevant but also accurate and reliable.
- Avoidance of Generative AI Tools: To maintain the authenticity and reliability of our dataset, we refrained from using generative AI tools to generate new content. All data was sourced from existing, vetted and not-copyrighted materials, ensuring that the information reflects actual nuclear knowledge.
- Expert Involvement: Throughout the dataset construction process, we involved experts in nuclear engineering. These experts played a crucial role in filtering the data and ensuring that it met the stringent standards required in the nuclear field.

Despite the time-intensive nature of this approach, we believe it is necessary when working in the nuclear field to ensure the highest standards of quality and assurance. This meticulous approach is reflected in the quality of our datasets and subsequently in the performance of our fine-tuned LLMs.

3.1 SMLP dataset

The first dataset was created by the Supervised Machine Learning Program (SMLP) team, which is composed of nuclear professionals. The Assystem team's mission is to create datasets from actual knowledge, and they also work on the tasks of evaluation and validation of datasets and results of our models.

For the CurieLM project, we created the dataset following best practices to ensure high quality:

- Reference Creation: Several groups from the SMLP team identified a set of non-copyrighted references in the nuclear field. They then crafted high-quality question-answer pairs based on these references.

- Evaluation and Validation: Another team within the SMLP reviewed the work done by the first team. This checking process ensured that the dataset met high standards.

The total number of instructions created by the SMLP team for this project is 3,000. This rigorous process, while time-consuming, was necessary to ensure the quality and reliability of the data. The high standards required in the nuclear field necessitated this meticulous methodology.

3.2 DrQA dataset

The second dataset, called DrQA, was created by Assystem's dismantling experts in 2019 for the task of extractive question answering. The original DrQA dataset contains 13,000 entries.

However, the format of the DrQA dataset was not directly suited for our task. Therefore, we embarked on a process to adapt this dataset to an instruction format that would be more suitable for our project. We developed a script for conversion. The SMLP team then validated the output of this script, ensuring the converted data met the high standards required for nuclear dismantling-related tasks. Only 1,500 instructions from the DrQA dataset will be used for CurieLM's training.

3.3 NucBank dataset

NucBank is the name for Assystem proprietary documents and all non-copyrighted documents collected by Assystem. These documents has been used according to the novel approach described in [19], which proposes a method to create large-scale instruction datasets for fine-tuning LLMs without relying on large amounts of human-annotated data. This process involves two main steps:

- Self-Augmentation: We started with an instruct LLM (i.e. Mixtral-8x7B [20], a LLM fine-tuned on the task of instruction following) and used it to generate instructions from unlabelled nuclear documents. It's important to note that the LLM was not generating instructions based on its own knowledge but on vetted and selected documents. Instead, it was given a document in the nuclear field as context and was asked to generate instructions based on that document.
- Self-Curation: Given that the self-augmentation approach can generate a high volume of data, some of which may be of poor quality, a subsequent step of self-curation is necessary. This step involves filtering out low-quality data to ensure the final dataset maintains high standards.

To further ensure the quality of our dataset, we involved the SMLP team in the validation process. The team reviewed the data generated in the self-augmentation phase and validated it against the high standards we set for our dataset. After validation by the SMLP team, this dataset comprises 1,000 instructions.

4 Instruction Tuning

Transitioning from the discussion on data collection and preparation, we now come to the essential process of instruction tuning. This process is integral to enhancing the performance of LLMs on domain-specific tasks, particularly in complex fields like nuclear engineering.

LLMs, such as GPT-4 , LLaMA2 and Mistral , have been trained on extensive datasets, providing them with a general understanding of language. However, their performance on domain-specific tasks can be significantly improved through a process known as instruction

tuning. This process involves fine-tuning the LLM on a set of domain-specific instructions, enabling it to improve its performance on related tasks.

Despite the benefits of instruction tuning, the process of fine-tuning LLMs is computationally intensive and requires significant resources and time. This has traditionally limited the applicability of fine-tuning, particularly for larger models that barely fit into GPU memory.

The advent of Low-Rank Adaptation (LoRA) [21] and Quantized LoRA (QLoRA) [22] has revolutionized the fine-tuning process. These techniques reduce the computational requirements of fine-tuning by breaking down complex structures, called weight matrices, into simpler and smaller, lower-rank forms This allows for efficient training of custom LLMs, even on consumer-grade GPUs.

The efficiency provided by LoRA and QLoRA not only makes fine-tuning more accessible but also enhances the feasibility of instruction tuning. This paves the way for the development of domain-specific LLMs like CurieLM, which are tailored to perform optimally on nuclear-related tasks.

4.1 Model selection

The model to be fine-tuned with the nuclear datasets must have three main features:

- To grant data confidentiality, the model should not be cloud-based but could be deployed locally on-premise;
- To grant data confidentiality, the model should be able to be hosted by the user, without passing by third-party services (e.g. OpenAI servers for GPT's server);
- The model's licence conditions should not restrict the use of the LLM in a nuclear context.

Online services like GPT-3.5 or GPT-4 by OpenAI or Gemini by Google shall then be excluded by the selection.

To facilitate on-premise deployment, the target model should not be too large and heavy. Thus, the two candidates left are Mistral-7B by Mistral and Llama-2-13B & Llama-2-34B by Meta. However, models from Llama-2 family cannot be used for nuclear applications (i.e. they do not respond to the third requirement).

Mistral-7B is then retained as our foundational LLM. This model, despite its relatively smaller size of 7 billion parameters, has demonstrated exceptional performance [3] (i.e. outperforming Llama-1-34B in areas of reasoning, mathematics, and code generation).

After examining the capabilities and advantages of Mistral-7B, we shifted our focus to the specific variants of this model that we selected for our experiments. Our choice was guided by the requirement for a model that could not only effectively comprehend and generate text but also tackle the challenges associated with complex tasks characteristics of nuclear engineering. This led us to two variants of Mistral-7B that have been specifically fine-tuned for instruction following tasks: Mistral-7B-Instruct and Dolphin-Mistral-7B-Instruct.

4.2 Experiments

In our experiments, we fine-tuned both Mistral-7B-Instruct and Dolphin-Mistral-7B-Instruct on the nuclear dataset presented in Section 4. The dataset was split into training and testing sets, with a sequence length of 2048 tokens (one token generally corresponds to ¾ a word).

Both experiments were conducted with the same hyperparameters for LoRA and training. Interestingly, we observed that both experiments followed the same pattern of training. Initially, the evaluation loss decreased rapidly, indicating effective learning. However, after a period of stability, the evaluation loss started to increase.

To prevent overfitting and catastrophic forgetting [23] (a common phenomenon in LLM fine-tuning where the model forgets previously learned information), we decided to stop training before the evaluation loss started to increase. This strategy ensured that the models did not lose their already developed capacities, such as reasoning.

These experiments underscore the potential of fine-tuning LLMs for domain-specific tasks, despite the challenges. The next section will delve into the evaluation of these models and the results obtained.

5 Evaluation

In evaluating CurieLM's performance, we adopted a dual approach. The first is an automatic evaluation using Mixtral-8x7B [20], a high-quality sparse mixture of experts model (SMoE). A Mixture of Experts model is a type of model that combines several smaller 'expert' subnetworks. Each of these subnetworks specializes in handling different types of tasks, allowing the model to leverage the strengths of each expert for different parts of the input space. This leads to more accurate and robust overall performance. The second approach involves a human evaluation conducted by an expert in nuclear engineering at PhD-grade.

For evaluation, we used four models: Mistral-7B-Instruct (not fine-tuned), CurieLM-Mistral-7B-Instruct (fine-tuned), CurieLM-Dolphin-Mistral-7B-Instruct (fine-tuned) and GPT-4 (not fine-tuned).

These models were chosen to compare the performance of a base model (Mistral-7B-Instruct) with that of the same model after fine-tuning (CurieLM-Mistral-7B-Instruct and CurieLM-Dolphin-Mistral-7B-Instruct). Even if it does not have two main features listed in the model selection section (i.e. granting data confidentiality through on-premise deployment and avoiding third-party services), we included GPT-4 in the evaluation as it is the state-of-the-art model. This model is much bigger than the other cited models and has been trained on way more data. Thus, it is expected to overperform smaller models, even fine-tuned. Its performances will be used as reference and goal for smaller fine-tuned models.

5.1 Automatic Evaluation

The automatic evaluation was performed on a dataset of question-answer pairs that were unseen by all three models during their training phase. This ensured an unbiased evaluation of the models' ability to generalize to new data.

We employed Mixtral-8x7B for performance evaluation due to its exceptional capabilities. This model not only outperforms Llama-2-70B on most benchmarks with an inference speed that is six times faster, but it also matches and even surpasses the performance of GPT3.5 [20], demonstrating its superior effectiveness.

To evaluate the correctness of the answers provided by each LLM, we prompted Mixtral-8x7B with the question, the LLM's answer, and the true answer. We then asked Mixtral-8x7B to evaluate the correctness of the LLM's answer based on the true answer. After several trials, we found the best prompt that yielded the most accurate evaluations.

The results of the automatic evaluation were quite revealing. The base model, Mistral-7B-Instruct, achieved a correctness score of 47.13%. The fine-tuned models, CurieLM-Mistral-7B-Instruct and CurieLM-Dolphin-Mistral-7B-Instruct, achieved scores of 55.75% and 60.34% respectively. This represents a gain of 13.21% for the CurieLM-Dolphin-Mistral-7B-Instruct model compared to the base model, underscoring the effectiveness of fine-tuning LLMs on nuclear data.

We also observed that the base model provided answers in English even when the prompt and question were in French. In contrast, the fine-tuned models correctly responded in the

Table 1. Main characteristics of CurieLM's datasets

Criterion	1	2	3	4	5	Total
Mistral-7B-Instruct	3	3	2	6	4	18
CurieLM-Mistral-7B-Instruct	6	6	3	7	6	28
CurieLM-Delphin-Mistral-7B-Instruct	**8**	**7**	**3**	9	**8**	35
GPT-4	7	7	3	**10**	8	35

language of the prompt. Furthermore, the fine-tuned models provided more concise and helpful answers, while the base model provided either longer answers with details that were not related to the nuclear field or hallucinations (i.e. completely invented answers with false information).

GPT-4 was also included in the evaluation. The model achieved a remarkable score of 72%, outperforming all three Mistral-based models. This performance- is explained by GPT-4's larger size and training data. However, comparing GPT-4 with Mistral-based models may not be entirely fair due to its larger scale. Nonetheless, GPT-4 serves as a benchmark for performance and underscores the need for further improvements in domain-specific LLMs. Moreover, GPT-4 could not be employed in the nuclear domain because of the lack in granting data confidentiality through on-premise deployment and avoiding third-party services.

5.2 Human Evaluation

The human evaluation was performed on ten question-answer pairs randomly selected from the dataset used for the automatic evaluation. These allowed a senior nuclear expert to review the nuclear-specific capabilities of CurieLM's models with respect to the base model Mistral-7B-Instruct.

The nuclear expert has been asked to evaluate the three LLMs answers based on the following five criteria:

- The answer deeply covers all the aspects requested in the question;
- The answer is clear and precise, making it easy to understand the details and explanations provided;
- The given examples in the answers are specific, relevant and detailed, reinforcing the credibility of the answer;
- The answer is well structured and logically organized;
- The answer demonstrates solid expertise in the nuclear field and the subject threated by the question.

The nuclear expert gave the LLM's answer as much points as met criteria. Thus, an LLM can have up to five points (i.e. the LLM's answer meets all the criteria) per question, up to a total of 50 points (10 points per criterion). In Table 1, the evaluation results are shown and then commented.

Firstly, the three first scores (respectively 18, 28 and 35 points) given by the nuclear expert are coherent with the automatic evaluation performed by Mixtral-8x7B (i.e. respectively 47.13% 55.75% and 60.34%). The last score from GPT-4 (35 points) is equal to CurieLM-Delphin-Mistral-7B-Instruct's, showing an incoherency with the evaluation performed by Mixtral-8x7B (72%). This could be explained by the fact that the nuclear expert evaluated the models on 10 answers and Mixtral-8x7B on the whole dataset. This incoherency will require more investigations. Here below, we comment the main results of each LLM.

Mistral-7B-Instruct received a total score of 18 points out of 50, showing capabilities in providing well-structured answers but struggling to provide comprehensive, clear, and precise answers with relevant and specific examples, and to demonstrate solid expertise in the nuclear field. This result was expected since it has not been trained on nuclear-related data.

CurieLM-Mistral-7B-Instruct received a total score of 28 points out of 50, showing better performances in all criteria. It shows strong performance in the depth of coverage of the answers, the clarity and precision of the answers and for the expertise demonstrated in the nuclear field. However, it struggles in giving relevant and specific examples. These results suggest that CurieLM-Mistral-7B-Instruct is capable of providing comprehensive, clear, and precise answers with solid expertise in the nuclear field.

CurieLM-Delphin-Mistral-7B-Instruct received the highest total score of 35 points out of 50 (as GPT-4), with its strongest performance in the fourth criterion (i.e. the structure and logical organization of the answer).

GPT-4 received the highest total score along with CurieLM-Delphin-Mistral-7B-Instruct, with an outstanding performance in the structural and logical organization of the answer. However, it shows less ability in covering all the nuclear-related part of the answers.

These results suggest that CurieLM-Delphin-Mistral-7B-Instruct is capable of providing comprehensive, clear, precise, and expert answers, with a clear and coherent flow of ideas. Moreover, it shows a performance on nuclear-related question-answering comparable with GPT-4.

6 Conclusion

In conclusion, the introduction of CurieLM signifies a substantial advancement in the application of Large Language Models (LLMs) within the nuclear domain. The model's remarkable 13% performance enhancement over base LLMs underscores the potential of domain-specific fine-tuning, paving the way for more effective and specialized models in nuclear-related tasks.

CurieLM's development addresses the unique challenges of the nuclear field, such as safety assessments and requirements analyses, thereby bridging the gaps in domain-specific knowledge and application efficacy inherent in general LLMs. This tailoring to the nuclear domain marks a significant step forward in the revolutionization of nuclear-related natural language processing tasks.

The paper's comprehensive evaluation, employing both automatic and human assessments, further validates CurieLM's superiority. The model outperformed the base LLM in correctness, language adaptability, and the provision of concise, relevant, and domain-specific responses. This superior performance was corroborated by evaluations conducted by Mixtral-8x7B and a senior nuclear expert. The expert's evaluation led to state-of-the-art GPT-4 comparable results. This is a strong result that need further investigations in light of the limited validation dataset used by the expert with respect to the automatic evaluation's one used by Mixtral-8x7B.

The contributions of this paper extend beyond the development of CurieLM. It presents a robust framework for constructing high-quality, domain-specific datasets, provides a curated dataset validated by nuclear experts, and demonstrates the efficacy of LLM fine-tuning for domain-specific applications.

While CurieLM represents a significant stride, the paper acknowledges the need for continuous improvements, particularly in generating relevant and specific examples within answers. This sets the stage for future research and refinement in the field of domain-specific LLMs, emphasizing the dynamic and evolving nature of this exciting area of study applied to high-regulatory domains such as nuclear.

Acknowledgments

We extend our sincere gratitude to all members of the SMLP team who are nuclear professionals for their invaluable contributions to this project. Their expertise, insights, and feedback were instrumental in creating and validating the high-quality data that meets the rigorous standards of the nuclear industry. Their dedication and collaboration have been indispensable in ensuring the success of this endeavour.

References

[1] OpenAI, *Introducing chatgpt* (2022), `https://openai.com/blog/chatgpt`
[2] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F.L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat et al., arXiv preprint arXiv:2303.08774 (2023)
[3] A.Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D.S. Chaplot, D.d.l. Casas, F. Bressand, Lengyel et al., arXiv preprint arXiv:2310.06825 (2023)
[4] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale et al., arXiv preprint arXiv:2307.09288 (2023)
[5] OpenAI, *Aligning language models to follow instructions* (2022), `https://openai.com/research/instruction-following`
[6] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray et al., Advances in Neural Information Processing Systems **35**, 27730 (2022)
[7] A. Toma, P.R. Lawler, J. Ba, R.G. Krishnan, B.B. Rubin, B. Wang, arXiv preprint arXiv:2305.12031 (2023)
[8] Y. Li, Z. Li, K. Zhang, R. Dan, S. Jiang, Y. Zhang, Cureus **15** (2023)
[9] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, G. Mann, arXiv preprint arXiv:2303.17564 (2023)
[10] H. Yang, X.Y. Liu, C.D. Wang, arXiv preprint arXiv:2306.06031 (2023)
[11] J. Cui, Z. Li, Y. Yan, B. Chen, L. Yuan, arXiv preprint arXiv:2306.16092 (2023)
[12] J.S. Lee, arXiv preprint arXiv:2306.05431 (2023)
[13] B. Zakaria, D.T. Kien, A. Ahmed, P. Alexandra, C. Riccardo, R. Emir, Technical Meeting on the Safety Implications of the Use of Artificial Intelligence in Nuclear Power Plants (2023)
[14] D. Cheng, S. Huang, F. Wei, arXiv preprint arXiv:2309.09530 (2023)
[15] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, K. Clark, S. Pfohl, H. Cole-Lewis, D. Neal et al., arXiv preprint arXiv:2305.09617 (2023)
[16] A. Jain, D.N. Meenachi, D.B. Venkatraman, arXiv preprint arXiv:2003.13821 (2020)
[17] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, arXiv preprint arXiv:1810.04805 (2018)
[18] A. Acharya, S. Munikoti, A. Hellinger, S. Smith, S. Wagle, S. Horawalavithana, arXiv preprint arXiv:2310.10920 (2023)
[19] X. Li, P. Yu, C. Zhou, T. Schick, L. Zettlemoyer, O. Levy, J. Weston, M. Lewis, arXiv preprint arXiv:2308.06259 (2023)
[20] A.Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D.S. Chaplot, D.d.l. Casas, E.B. Hanna, F. Bressand et al., arXiv preprint arXiv:2401.04088 (2024)
[21] E.J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, arXiv preprint arXiv:2106.09685 (2021)
[22] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, Advances in Neural Information Processing Systems **36** (2024)
[23] Y. Luo, Z. Yang, F. Meng, Y. Li, J. Zhou, Y. Zhang, arXiv preprint arXiv:2308.08747 (2023)