Large Language Models for Isotope Separation

Endes
Department of Computer Science
Rochester Institute of Technology
Rochester, NY, United States
lge9081@rit.edu

Soto
Computer Science Initiative
Brookhaven National Laboratory
Upton, NY, United States
csoto@bnl.gov

Hatcher-Lamarre
Collider-Accelerator Department
Brookhaven National Laboratory
Upton, NY, United States
JHatcherLamarre@bnl.gov

***Abstract*—The Brookhaven National Laboratory Medical Isotope Research and Production group researches and produces experimental isotopes that are used in nuclear medicine using the Brookhaven Linac Isotope Producer. Before producing isotopes, scientists must go through the process of analyzing research papers to identify ideal separation conditions of these isotopes. This study examines the use of open-source large language models to extract separation conditions of isotopes from scientific literature using prompt engineering. To test the models, prompts are engineered with different techniques to find the most effective way to extract information from papers. It is found that the best approach is Chain-of-Thought prompting with LLaMA-2-13b. This will support research of automated chemical extraction using artificial intelligence.**

Keywords*—large language models, Llama-2, Galactica, Falcon, separations, isotopes, artificial intelligence, natural language processing

I. Introduction

A. Background

At Brookhaven National Laboratory (BNL), the Medical Isotope Research and Production group (MIRP) prepares commercially unavailable radioisotopes, like Actinium-225, that are used for the imaging and treatment of diseases like cancer. To do this, they utilize the Brookhaven Linac Isotope Producer (BLIP) that irradiates targets with protons at an intensity of 165 μA with an energy of up to 202 MeV originating from Brookhaven's linear accelerator. Since Brookhaven produces isotopes that are in great demand, researchers at BNL pursue the production pathway of other radioisotopes as well as alternative methods for the production of known isotopes. The problem with this is that the production of isotopes is a time-consuming process in which much of the time is spent identifying chemical separation conditions from previously published scientific literature. This study proposes the use of open-source large language models (LLMs) to extract information on the separation conditions of isotopes from the scientific literature using inference-only prompt engineering in order to accelerate the production of chemical isotopes.

LLMs: Large language models are machine learning algorithms designed to process text in a word-by-word generative mode. With the revelation of the transformer architecture from the 2017 paper, "Attention is All You Need" [1], large language models became more attainable, as transformer models are much more parallelizable, require much less time to train, and are generally superior in quality to the old norm of using recurrent neural networks. Recently LLMs became a hot topic in both the computing space and public media alike with the release of ChatGPT, a highly advanced, though closed-source, LLM optimized for chat-like dialogue that became renowned for its capabilities and confidence.

Fig. 1. Recently produced Actinium-225 depicted in a long-exposure image[1]

B. Relevant Literature

"Attention is All You Need" [1] proposed the transformer architecture for machine learning models, which is based entirely on attention mechanisms consisting of matrix multiplication, scaling, masking, and softmaxing of matrices comprised of queries (Q), keys (K), and values (V).

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

Transformers are made up of what is known as "multi-head attention" layers, which are basically the concatenation of multiple attention mechanisms.

$$MultiHead(Q,K,V) = Concat(head_1,...,head_i) \quad (2)$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (3)$$

This project was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internships Program (SULI).

[1] https://www.bnl.gov/newsroom/news.php?a=221204

Where W^{Q}_{i}, W^{K}_{i}, W^{V}_{i}, are projections represented as parameter matrices.

One way that transformers use this multi-head attention is in encoder-decoder attention layers in which "queries come from the previous decoder layer, and the memory keys and values come from the output of the encoder."

Generative LLMs take the encoder-decoder attention layers and utilize the decoder aspect of that attention to create a decoder-only architecture which allows for the function of LLMs to perform text generation as a type of auto-complete.

"Automated Extraction of Molecular Interactions and Pathway Knowledge using Large Language Model, Galactica: Opportunities and Challenges" [2] investigates the use of Galactica in life science research related to protein interactions. Ultimately, the researchers at BNL's Computer Science Initiative concluded that Galactica performs best when contextual text is provided, indicating a relevant connection to the use of Galactica for use in isotope production based on previous research.

"Med-HALT: Medical Domain Hallucination Test for Large Language Models" [3] looks into the challenges posed by hallucinations in LLMs. In the end, they found that while Falcon 40B is the most performant model of their suite that includes MPT-7B and all LLaMA-2 variations in information retrieval, all the models face a multitude of hallucination problems.

"Local Large Language Models for Complex Structured Tasks" [4] describes the "end-to-end process of training, evaluating, and deploying" local LLMs to perform domain-specific NLP tasks; concluding that LLaMA-based models perform much better than BERT-based models in all NLP tasks due to their size difference. This paper may provide insight on better ways to implement the models being used in the domain of chemical separations and isotope production.

"Llama 2: Early Adopters' Utilization of Meta's New Open-Source Pretrained Model" [5] provides a description of multiple different projects utilizing the new LLaMA-2 models. Of the described projects, "Document-based-question-answering-system-using-LLamaV2-7b"[2] and "H2oGPT"[3] seem to be the most relevant. The former is of great interest, as it is a question-answering system based on documents, which could almost certainly be adapted to papers on isotope production. The latter functions to query and summarize documents using new models like LLaMA-2 and Falcon which seems worthwhile to look further into for the future of this project.

C. *Scope*

This project was mostly limited by hardware. All models and tests run in this study were run on a NVIDIA RTX 4090 with 24 Gigabytes of VRAM. In addition to this limitation, this project is essentially an experimental proof-of-concept for later work to be built on. All work done during this study can be found at https://github.com/Log45/LLMs_For_Isotopes.

II. Methods

A. *Document Parsing*

A key part of drawing conclusions based on research papers is being able to access information from those documents. The majority of scientific literature is only accessible through PDF and HTML formats. While it is simple for humans to read papers from websites and PDFs, LLMs need access to the raw text of those documents in order to correctly tokenize the meanings of those words. To get access to the raw text of literature, both PDF and HTML files must be parsed.

PDF Parsing: In order to access text from PDFs, a Python library known as PDFQuery is used, which allows users to simply access different attributes from PDFs, of which build documents by placing text boxes and other features at specific coordinates. The most important attribute, and the one that this project must access, are the text elements of the PDFs which are then extracted into a text file. This text file is taken one step further and split into a list of context paragraphs which are split up based on the size of gaps between text in the file.

HTML Parsing: In order to access text from HTML files, a Python library known as BeautifulSoup is used, which allows users to search for key features like classes from HTML code. Since the needed text is nested behind specific classes in HTML, the relevant class of text is identified, and each separate paragraph associated with that class is returned in a list of contexts. A downside to using HTML is that each website that has relevant articles uses different class names, so the parser currently only works for papers from mdpi.com, though it would not be much more work to make it compatible with other websites like springer.com that have relevant papers.

B. *Context Filtering*

In an attempt to avoid extracting information from irrelevant paragraphs in papers, a variety of different filters that are applied to the data before the LLMs return conclusions based on those contexts were created.

Keyword Filter: The keyword filter (K) is defined well by its name, it sorts through each context paragraph and only generate responses with the LLMs if there are any keywords present in the paragraph. The keywords that it searched for were: separation, isolation, chromatography, ion exchange, eluted, elution, elute, fraction, resin, exchange, acid, and target.

Model Filter: The model filter (M) uses the LLM to predict whether or not each context contains a chemical extraction in it. The model is provided with the context and asked whether a chemical extraction is being described. If it returns "yes" as an answer, then the generation is carried on; if it does not answer "yes", then the paragraph is thrown out and it moves onto the next.

Keyword-Model Filter: The keyword-model filter (KM) combines both the keyword filter and model filter. If a paragraph contains a relevant keyword, then the LLM is asked whether there is a chemical separation in the paragraph, and it continues on to generate conclusions if it responds that there is a separation described.

[2] https://github.com/10deepaktripathi/Document_based_question_answering_system_using_LLamaV2-7b

[3] https://github.com/h2oai/h2ogpt

Keyword-Model-Check Filter: The keyword-model-check filter (KMC) combines the keyword filter and model filter as previously described, but takes the filtering a step further by asking whether or not the generation is a truthful statement in regards to the original context. If the model responds with "yes," then it allows the generation to be added to the list of all generations. This filter uses a second pass of the LLM to refine results generated in the first pass with the goal of more accurate generations.

Keyword-Model-Expert Filter: The keyword-model-expert (KME) is less of a filter for the context and more of an attempt at manipulating the LLM to respond more accurately by telling it to act like an expert in chemistry whereas the keyword-model filter included with it filters the contexts used for conclusion generation.

Keyword-Model-Expert-Check Filter: The keyword-model-expert-check filter (KMEC) combines all of the previously described methods to filter as much irrelevant data as possible while also attempting to generate the most accurate responses at the cost of generation efficiency.

C. *Prompting Techniques*

Zero-Shot: Zero-shot prompting [6] is the simplest prompting technique for LLMs done by either asking the model a question without any context to look for a response, or by providing a context paragraph and asking the model to draw a conclusion based on that. For our purpose, the latter is a relevant technique to try to generate chemical separation suggestions as simply as possible, though the zero-shot capabilities of smaller parameter models like the ones used in this study are not generally as great as those of models with tens of billions of parameters like ChatGPT.

Few-Shot: Few-shot prompting [7] is not much more complex than zero-shot prompting, but allows for a much greater potential for the LLM to generate a desired output by giving at least one example of what the user expects in the context given to the model. This study observes the use of few-shot prompting with one example as well as two examples.

Chain-of-Thought: Chain-of-Thought (CoT) prompting [8] is much more performant than standard few-shot and zero-shot prompting for certain tasks as it allows for LLMs to use reasoning in their response generations. CoT prompting can be either few-shot or zero-shot. For few-shot CoT, the user must give an example answer in the context that the model reads to recognize the logical steps to lead to a conclusion. In doing so, the model is more inclined to use reasoning when generating conclusions due to its nature of imitation. For zero-shot CoT, the user must instill reasoning in the question itself or start off the response for the LLM to finish. The most popular zero-shot CoT phrase is "Let's think step by step," [9]. In our experiments, asking the model to act as an expert in chemistry functionally emulates this instillation of reasoning.

Automatic Prompt Engineer: Automatic Prompt Engineer (APE) [10] is a framework that uses an LLM to generate and select an instruction to give based on a specific task. Essentially, a model is given a list of input-output pairs and is asked to give an instruction to induce a language model to generate those pairs. After that, the same LLM is given those generated tokens and returns a logit-loss score to grade each instruction and choose the best one. The issue with APE is that the original framework is built to interface with OpenAI's API, which is closed-source and therefore out of scope due to the open-source requirement of this project. A localized version of APE was implemented, currently only supporting Meta's OPT models, though it will be easy to generalize that to HuggingFace's transformers in the future.

D. *Models*

Galactica: Galactica [11] is a model trained for the primary purpose of performing scientific and technical tasks trained on a corpus consisting of scientific papers, reference material, knowledge bases, and other sources that outperforms many models in LaTeX equations, mathematical reasoning, and even general tasks.

LLaMA-2: LLaMA-2 [12] is the second iteration of Meta AI's LLaMA (Large Language model Meta AI) [13] pretrained language model. While this second iteration focuses more on overhauling chat dialogue in their model, it is generally more performant than their original model which was already known as one of the most performant LLMs in 2023. While LLaMA-2 did not reveal its training corpus, it is very capable at a variety of tasks and stands as one of the best models that could be used in this study.

Falcon: Falcon[4], a new LLM by Technology Innovation Institute has shown impressive performance with both its 40B and 7B parameter variants, with its 40B variant outperforming LLaMA-65B in some tests. This model was trained on a corpus of 79 percent RefinedWeb [14] dataset and 2 percent of scientific data from different sources like arXiv, PubMed, USPTO, and more. Based on claims by the creators, Falcon is a great candidate for the application of information extraction from research papers.

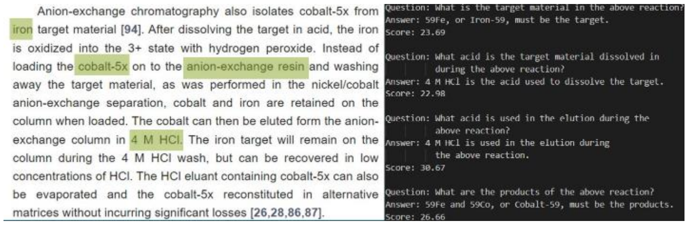
MPT-7B: Although MPT-7B [15] is a great candidate for this project and should be explored in the future, dependency issues impeded the use and testing of it in this work.

III. Results

Automatic Prompt Engineer: Although an implementation of APE that generated potential prompts was developed, it was dropped in favor of other techniques, as the initial results were very poor and the project was heading in the direction of context-filtering as opposed to automatic prompts. Despite this, there is definitely still potential in using APE if it is further developed and fine-tuned for smaller local models.

After giving each LLM context from different research papers and asking questions based on different aspects of chemical separations like the target material, the acid used to dissolve the target, the resin, the elution acid, and the products of the separation, they successfully generated suggestions for chemical separations, though with noted inaccuracies due to LLM hallucinations as shown in Figure 2.

[4] https://falconllm.tii.ae/



Anion-exchange chromatography also isolates cobalt-5x from iron target material [94]. After dissolving the target in acid, the iron is oxidized into the 3+ state with hydrogen peroxide. Instead of loading the cobalt-5x on to the anion-exchange resin and washing away the target material, as was performed in the nickel/cobalt anion-exchange separation, cobalt and iron are retained on the column when loaded. The cobalt can then be eluted form the anion-exchange column in 4 M HCl. The iron target will remain on the column during the 4 M HCl wash, but can be recovered in low concentrations of HCl. The HCl eluant containing cobalt-5x can also be evaporated and the cobalt-5x reconstituted in alternative matrices without incurring significant losses [26,28,86,87].

```
Question: What is the target material in the above reaction?
Answer: 59Fe, or Iron-59, must be the target.
Score: 23.69

Question: What acid is the target material dissolved in
          during the above reaction?
Answer: 4 M HCl is the acid used to dissolve the target.
Score: 22.98

Question: What acid is used in the elution during the
          above reaction?
Answer: 4 M HCl is used in the elution during
          the above reaction.
Score: 30.67

Question: What are the products of the above reaction?
Answer: 59Fe and 59Co, or Cobalt-59, must be the products.
Score: 26.66
```

Fig. 2. Human annotations (left) vs. AI generations (right) on a paragraph from “Production, Purification, and Applications of a Potential Theranostic Pair: Cobalt-55 and Cobalt-58m” [16]

Through a benchmark developed using two papers on isotope separation [16], [17], variations of each relevant model, filtering technique, and prompting technique were scored on accuracy, efficiency, perplexity, and wastefulness. Accuracy measures the percent of correct conclusions based on a human-made benchmark. Efficiency is the number of conclusions made per minute. Perplexity is a LLM score determined through logit loss. And Wastefulness is the percentage of generations made that are not included in the benchmark. These results can be seen in the figures below.

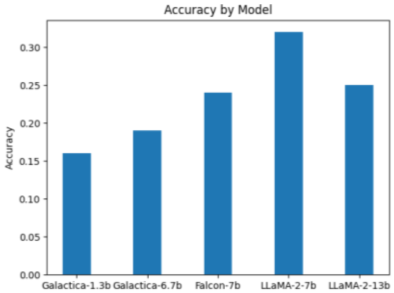


Fig. 3. Average accuracy of each model.

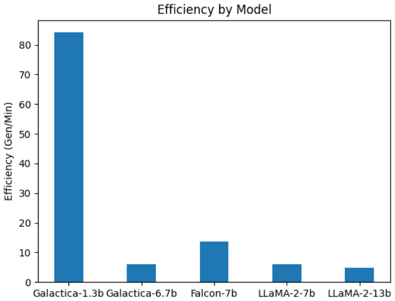


Fig. 4. Average efficiency of each model

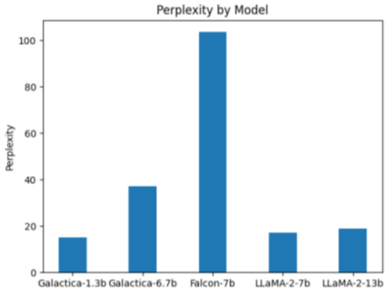


Fig. 5. Average perplexity score of each model

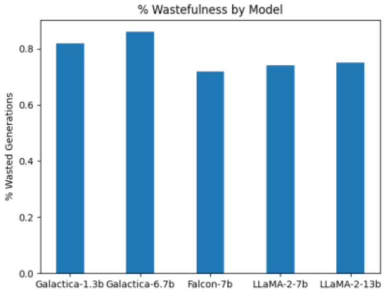


Fig. 6. Average wastefulness of each model

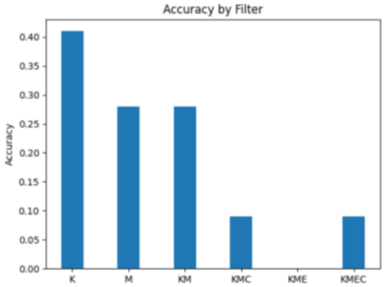


Fig. 7. Average accuracy of each filter. K=Keyword, M=Model, C=Check, E=Expert. The keyword-model-expert filter failed in all tests.

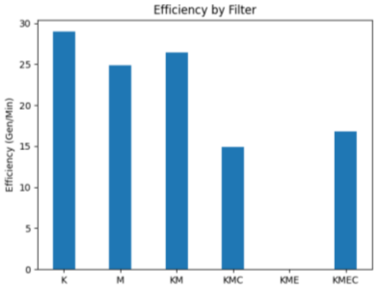


Fig. 8. Average efficiency of each filter

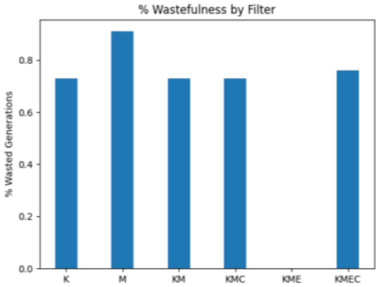


Fig. 9. Average wastefulness of each filter

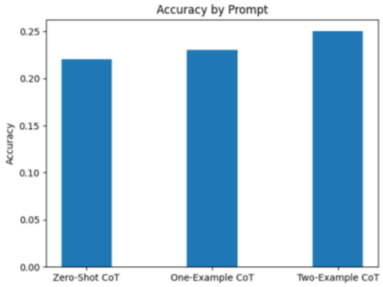


Fig. 10. Average accuracy of each prompt

Ultimately, it was found that LLaMA-2-13b using the keyword filter and two-example, few-shot, CoT prompting was the most accurate with an accuracy of 0.64, an efficiency of 7.58 gen/min, a perplexity score of 27.06, and a wastefulness of 0.67.

IV. Conclusion

This paper examined the use of LLMs and inference-only prompt engineering in conjunction with context-filtering techniques to aid in the research of chemical separations and isotope production. Document parsers for PDF and HTML file formats based on existing Python libraries were developed. The effectiveness of different filters in avoiding wasteful generations by language models was investigated – this is important as LLMs are costly to run, even in inference-only mode. The accuracy and efficiency of different prompting techniques in conjunction with different pretrained LLMs was compared. Ultimately, it was found that although language models continue to hallucinate, there is certainly potential for further development of this project to become a very helpful asset to researchers in the future.

Discussion: The findings of this study are going to support the future development of an automated chemical separation system in which an LLM will be used to suggest components of different chemical separations that the automated system can then test, rapidly accelerating the slow process of isotope research and production. In support of this goal, we believe there are opportunities for both an LLM and a BERT model with name-entity recognition to be used together to produce suggestions for these separations. For example, the LLM may generate a suggestion for a target, acid, resin, elution, and product of a separation in natural language, and an NER model may be applied to those generations to return only the relevant parts of those generations such as the specific element, acid, concentration, etc.

Acknowledgment

Logan Endes would like to thank his mentor, Carlos Soto, and the originating project PI, Jasmine Hatcher-Lamarre, for the great experience throughout the course of this project, and the great insights that they have provided him. This paper is not export controlled.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” CoRR, abs/1706.03762, 2017.

[2] G. Park, B. Yoon, X. Luo, V. Lpez-Marrero, P. Johnstone, S. Yoo, F. Alexander, “Automated extraction of molecular interactions and pathway knowledge using large language model, Galactica: opportunities and challenges,” in The 22nd Workshop of Biomedical Natural Language Processing and BioNLP Shared Tasks, Association for Computational Linguistics, July 2023, pp. 255-264.

[3] L. K. Umapathi, A. Pal, M. Sankarasubbu, “Med-Halt: medical domain hallucination test for large language models,” arXiv 2307.15343, 2023.

[4] V. K. C. Bumgardner, A. Mullen, S. Armstrong, C. Hickey, J. Talbert, “Local large language models for complex structured medical tasks,” arXiv 2308.01727, 2023.

[5] K. I. Roumeliotis, N. D. Tselikas, D. K. Nasiopoulos, “Llama 2: early adopters’ utilization of meta’s new open-source pretrained model,” Preprints 202307.2142, August 2023.

[6] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lster, N. Du, A. M. Dai, Q. V. Le, “Finetuned language models are zero-shot learners,” CoRR, abs/2109.01652, September 2021.

[7] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, “Language models are few-shot learners,” CoRR, abs/2005.14165, 2020.

[8] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. H. Chi, Q. Le, D. Zhou, “Chain of thought prompting elicits reasoning in large language models,” CoRR, abs/2201.11903, 2022.

[9] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, “Large language models are zero-shot reasoners,” arXiv 2205.11916, 2023.

[10] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, J. Ba, “Large language models are human-level prompt engineers,” arXiv 2211.01910, 2023.

[11] R. Taylor, M. Kardas, G. Cucurull, T. Scialom, A. Hartshorn, E. Saravia, A. Poulton, V. Kerkez, R. Stojnic, “Galactica: a large language model for science,” arXiv 2211.09085, 2022.

[12] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khasba, I. Kloumann, A. Korenev, P. S. Koura, M. Lachaux, T. Lavirl, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, T. Scialom, “Llama 2: open foundation and fine-tuned chat models,” arXiv 2307.09288, 2023.

[13] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, “LLaMA: open and efficient foundation language models,” arXiv 2302.13971, 2023.

[14] G. Penedo, Q. Malartic, D. Hesslow, R. Cojocaru, A. Cappelli, H. Alobeidli, B. Pannier, E. Almazrouei, J. Launay, “The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only,” arXiv 2306.01116, 2023.

[15] MosaicML NLP Team, “Introducing MPT-7B: a new standard for open-source, commercially usable LLMs,” www.mosaicml.com/blog/mpt-7b, 2023.

[16] K. E. Barrett, H. A. Houson, W. Lin, S. E. Lapi, J. W. Engle, “Production, purification, and applications of a potential theranostic pair: cobalt-55 and cobalt-58m,” Diagnostics, vol. 11, number 7, article 1235, 10.3390/diagnostics11071235, 2021.

[17] Q. Xie, H. Zhu, F. Wang, X. Meng, Q. Ren, C. Xia, Z. Yang, “Establishing reliable Cu-64 production process: from target plating to molecular specific tumor Micro-PET imaging,” Molecules, vol. 22, number 4, article 641, 10.3390/molecules22040641, 2017.