TOWARDS UNLOCKING INSIGHTS FROM LOGBOOKS USING AI

A. Sulc, G. Hartmann, HZB, Berlin, DE
J. Maldonado, BNL, New York, USA
V. Kain, F. Rehm, CERN, Meyrin, CH
A. Eichler, J. Kaiser, T. Wilksen, F. Mayet, R.Kammering, H. Tuennermann, DESY, Hamburg, DE
J. St. John, H. Hoschouer, K. J. Hazelwood, FNAL, Batavia, IL, USA
T. Hellert, LBNL, Berkeley, CA, USA
D. Ratner, W.-L. Hu, A. Bien, SLAC, Menlo Park, CA, USA

Abstract

Electronic logbooks contain valuable information about activities and events concerning their associated particle accelerator facilities. However, the highly technical nature of logbook entries can hinder their usability and automation. As natural language processing (NLP) continues advancing, it offers opportunities to address various challenges that logbooks present. This work explores jointly testing a tailored Retrieval Augmented Generation (RAG) model for enhancing the usability of particle accelerator logbooks at institutes like DESY, BESSY, Fermilab, BNL, SLAC, LBNL, and CERN. The RAG model uses a corpus built on logbook contributions and aims to unlock insights from these logbooks by leveraging retrieval over facility datasets, including discussion about potential multimodal sources. Our goals are to increase the FAIR-ness (findability, accessibility, interoperability, and reusability) of logbooks by exploiting their information content to streamline everyday use, enable macro-analysis for root cause analysis, and facilitate problem-solving automation.

INTRODUCTION

While large language models (LLMs) have aided operators [1–3], particle accelerator electronic logbooks (eLogs) remain underutilized due to their non-standard nature, employing highly technical language with content hidden in metadata, and privacy concerns. This challenges traditional LLMs in processing eLogs' valuable operational insights.

In this work [4], we aim to leverage Retrieval Augmented Generation (RAG) models, tailored for eLogs, to enhance the usability and make the first steps towards automation capabilities of these highly technical data sources across participating facilities and summarize their current progress.

RELATED WORK

In [1] the authors leverage notes from the fusion experiments DIII-D and Alcator C-Mod to develop a prototype "copilot" using RAG and off-the-shelf LLMs like `llama2-70b`. Their work demonstrates advantages in semantic search of experiments and assisting with device-specific operations compared to base language models and keyword search, mostly thanks to RAG combined with powerful `llama-2-70b`. Multiple particle accelerator facilities are currently working on their solution for how to lift their language sources. We are adopting a similar approach as [1], although it turned out that many eLogs have some specific challenges discussed later.

CERN is developing AccGPT [5], an ongoing project to integrate AI assistants into particle accelerator operations, including control room assistance, coding aid, documentation, and knowledge retrieval enhancement while streamlining FAQs. A prototype capable of retrieving knowledge from various CERN internal documentation is already in production.

At DESY, Mayet developed a RAG [2, 3] using `mixtral-8x7b-instruct-v0.1` for the SINBAD-ARES facility. The system further enhances user experience with the eLog with a chain of thought [6] and ReAct prompting [7]. The system uses various data sources including eLog, DESY DOOCS control system, machine-specific documentation, and even facility chat groups.

The work of [8] showed a logbook analysis that benefits from multimodal sources by enhancing the eLog with text extracted from images via OCR. The main novelty of this work is the combined use of images and topic models to tailor and enhance the eLog system at BNL.

PACuna [9] mines accelerator data to generate Q&A pairs for fine-tuning LMs, yielding tailored models for accelerator queries general LLMs can't handle, enabling facility-specific intelligent assistants.

RAG

The RAG pipeline is an approach that combines retrieving relevant documents from a large corpus with language generation models.

First, given an input query q, an ordered list of N relevant documents $\mathcal{D} = (d_1, \ldots d_N)$ are retrieved using dense vector indexing. In the second step, $\mathcal{D}$ and q are then passed to a generator (an LLM), which generates the final answer a with q and the retrieved documents $\mathcal{D}$ plugged into a predefined query template,

$$q \rightarrow \underbrace{\text{retrieve}(q)}_{\mathcal{D}} \rightarrow \underbrace{\text{generate}(q, \mathcal{D})}_{a} \rightarrow a. \tag{1}$$

This allows language models to hallucinate less and utilize real-world knowledge present in the retrieved documents $\mathcal{D}$ instead of internal (trained) model parameters.

Document Retrieval - retrieve(q)

Embedding is crucial for retrieval, but off-the-shelf models lack domain knowledge. We fine-tuned on domain data for SimCSE [10] but struggled with longer texts even after fine-tuning `sci-bert-{cased, uncased}`. The cased version performed better, but `all-MiniLM-L6-v2` outperformed while balancing size and accuracy.

Re-ranking [11] can improve performance by filtering and re-ranking based on relevance scores from a specialized model. Only top-ranked documents are passed to the generator, focusing on pertinent information.

CURRENT STATUS

This section will provide a concise overview of the existing eLog technology and outline the plans.

ALS

At the ALS, we've enhanced semantic logbook searches due to acronym/jargon use. Initial word2vec [12] attempts were unsatisfactory, prompting robust solutions. We're refining an embedding model and integrating `all-MiniLM-L6-v2` into the RAG pipeline with `Mistral-7B-Instruct-v0.2`.

Additionally, we've started recording weekly Operations Critic Meetings for RAG pipeline inclusion. Transcribing these faces hurdles due to poor Zoom audio quality, multiple in-room participants, and prevalent acronym/colloquial language use.

BNL

Techniques The BNL eLog system used at the Collider-Accelerator Department (C-AD) allows users to customize logbook settings, including the specification of favorite logbooks. Our goal was to use ML techniques to further personalize configurations and provide users with entries that match their specific interests. This work will benefit users who troubleshoot systems with a need for quick and accurate information to avoid delays in operational time.

We utilized Gensim [13] tools and the Doc2Vec [14] model to augment the search feature. The eLog data is stored in a MySQL database and imported using custom tools into a Pandas dataframe. The content of the eLog entries was transformed into vectors of words after tokenization, lemmatization, and removing punctuation. Initial tests showed the model was able to find similarities between common words like `polarization`, `beta`, `bunch`, `coherence`, and `emittances`. Classification techniques were explored to identify major topics in entries. These topics are useful as tags for grouping entries for users.

Web Application After testing, this search feature will be added to eLog. A temporary web demo [15] lets users test it by entering words/phrases to get similar entries. The text is processed through pretrained models, displaying search results with date, author, log name, similarity, and link.

Future Work For personalization, we aim to develop a reinforcement learning model using user feedback rewards. User feedback will also improve the workflow for eLog system implementation.

CERN

CERN is currently investing in building a pilot service with the name AccGPT [5] for hosting LLMs either fine-tuned or out-of-the-box with RAG based on CERN internal knowledge. For the time being, AccGPT will not incorporate logbooks due to concerns about their potentially lower data quality. With the idea that in the future, LLMs and AI, in general, could be used to fill next-generation logbooks and provide accelerator statistics. The team at CERN believes that this will make knowledge from logbooks and other accelerator data more exploitable. In the middle of 2024, AccGPT will go through its first accuracy tests with domain experts. At the same time benchmarking Q&A datasets are being curated.

DESY & BESSY

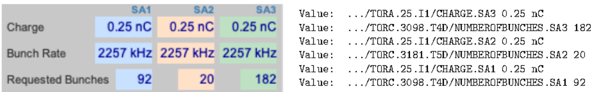
DESY and BESSY are developing new eLog simultaneously. DESY has already shown success in implementing RAGs at SINBAD-ARES [2, 3]. The implementation combines multiple sources such as the logbook itself, data obtained from the interaction with the control system, operations meeting slides, Mattermost chats, and the accelerator lattice (accelerator layout).

Embedding, Vectorstore and Generator For DESY European XFEL, we experimented with a fine-tuned embedding model as mentioned earlier, but smaller generic models like `all-MiniLM-L6-v2` [16], trained on supervised data still perform better. We also experimented with recent high-quality embedding models `mxbai-embed-large-v1` [17, 18] and `nomic-embed-text-v1` [19].

We use FAISS vectorstore [20] due to its simplicity and `Mistral-7B-Instruct-v0.2` as a generator.

Metadata We increased the verbosity of eLog entries by storing control system metadata from our control system when screenshots are made, see Fig. 1.

Future Work Recall is a bottleneck. Enhancing training data/benchmarks is needed. The next steps will lead to automating/augmenting eLog entries by processing screenshots ([21]), using achiever data, or enhancing from the machine/control system state. Use LLMs to enhance posts with specialized generators [9] or RAFT [22] (better RAG models).



	SA1	SA2	SA3
Charge	0.25 nC	0.25 nC	0.25 nC
Bunch Rate	2257 kHz	2257 kHz	2257 kHz
Requested Bunches	92	20	182

```
Value:  .../TORA.25.I1/CHARGE.SA3 0.25 nC
Value:  .../TORC.3098.T4D/NUMBEROFBUNCHES.SA3 182
Value:  .../TORA.25.I1/CHARGE.SA2 0.25 nC
Value:  .../TORC.3181.T5D/NUMBEROFBUNCHES.SA2 20
Value:  .../TORA.25.I1/CHARGE.SA1 0.25 nC
Value:  .../TORC.3098.T4D/NUMBEROFBUNCHES.SA1 92
```

Figure 1: Stored metadata from the control system.

We're exploring approaches like topic modeling [23], root cause analysis [24, 25], and using large language models for autonomous accelerator tuning [26].

Fermilab

ADEL The Fermilab Accelerator Directorate Electronic Logbook (ADEL) [27] is the primary record of accelerator operations at Fermilab. Everything from machine failures to study notes is recorded and categorized there. Since its creation in 2013, ADEL has recorded nearly one million user-generated entries, comments, and files. ADEL data may be filtered by ID, date, text, subject, log, category, and user. All data is saved in a relational database and a rich HTTP client and program API exists to read or extract data.

Past Work There was some effort in the past to introduce AI/ML into ADEL. The first attempt focused on using Optical Character Recognition (OCR) to capture text from the many thousands of controls system screenshots that are often posted in the eLog during tuning and failure diagnosis. The other use of AI/ML for ADEL was to categorize images to improve search [28]. There was moderate success training a model to classify images by which controls system GUI they originated from or if the image was of a photo or a screenshot. There were plans to run the trained image classifier model on the eLog data and save the categorization to a new database table field, but lack of resources and time hampered the work and it has not yet been implemented.

Current Work Fermilab has developed a semantic search prototype capable of searching ADEL quickly and retrieving relevant results. Accelerator Operators frequently use the Elog when troubleshooting everyday issues and have found the semantic search results to be more useful and relevant than lexical search results.

Embedding, Vectorstore, Re-Ranking and Recommendations For embedding entries, `all-mpnet-base-v2`[16] proved to be very powerful even compared to the unsupervised SimCSE fine-tuned model which is expected to be biased towards shorter log entries.

Qdrant[29] was used as the vector store. It is an open-source, locally hosted solution that allows for storing vectors alongside associated metadata, providing useful filtering options such as date, topic, and so on.

Additionally, a re-ranking function was implemented using cross-encoding models. Tested models for the re-ranking include `mxbai-rerank-base-v1`[11] and `ms-marco-MiniLM-L-12-v2`[30]. The re-ranking produces very relevant results, even without filtering, allowing retrieval of notes that may have been mislabeled. Qdrant also supports a recommendation feature, allowing it to push the query closer to relevant points.

Future work: The Fermilab team plans to fully integrate the search function with the Elog (ADEL) soon and also to incorporate image-similarity search.

SLAC

SLAC is investigating both text generation and natural-language queries (NLQ). Within text generation, LLMs can assist in generating routine reports, e.g. shift summaries, as well as assist in creating entries for future queries, e.g. by identifying ambiguous terms or suggesting keywords.

NLQ will help filter ambiguous terms (e.g. `beam energy`), select related terms (e.g. multiple names for an RF station), or identify historical entries that are similar to a target entry. More ambitiously, queries could perform calculations, such as computing variance in repeated measurements or total time spent on a task category.

Datasets Natural language datasets comprise eLogs for major accelerator facilities for both physicists and operators. Operator manuals (e.g. the ‘wiki’), technical documents and user-side eLogs provide additional context. Databases of GUI measurements and auto-archiving of diagnostic measurements, accelerator setpoints, and high-level analysis, can be linked to logbook entries through time stamps. Concerns exist that personally identifiable information (PII) in logbooks may require scrubbing or access limitation. As a precautionary measure, we have established a data processing workflow and are applying for an Institutional Review Board (IRB) review to ensure that our research complies with regulations and protects PII. User-side eLogs require permission from authors and are excluded from early work.

Initial work The first application of LLMs at SLAC augments logbook interpretation using an internal operations Mediawiki. Whitelisted Wiki articles support a RAG pipeline generating operations-backed context. A ’cleaning’ stage prompts an LLM to read an article, generate Q&A, and prompt staff for validation. The process updates the wiki and produces question-answer pairs for fine-tuning the LLM.

CONCLUSION

This collaborative work demonstrated initial progress in implementing RAG models to unlock insights from particle accelerator eLogs across multiple institutes with various logbooks. While challenges with a technical language remain, which cause issues in document retrieval, the findings highlight the potential of leveraging AI to enhance the usability and accessibility of eLog data. Further research building on this foundation is crucial for realizing the full impact of NLP and the use of metadata.

ACKNOWLEDGEMENT

We acknowledge DESY (Hamburg, Germany) and HZB (Berlin, Germany), a member of the Helmholtz Association HGF, for their support in providing resources and infrastructure. This work has in part been funded by the IVF project InternLabs-0011 (HIR3X). This manuscript has been authored in part by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics. SLAC work is supported by the Department of Energy, Office of Science, Office of Basic Energy Sciences under contract DE-AC02-76SF00515.

REFERENCES

[1] V. Mehta *et al.*, "Towards llms as operational copilots for fusion reactors," in *NeurIPS 2023 AI for Science Workshop*, 2023.

[2] F. Mayet, *Building an intelligent accelerator operations assistant using advanced prompt engineering techniques and a high level control system toolkit*, 2024.

[3] F. Mayet, "Gaia: A general ai assistant for intelligent accelerator operations," *arXiv preprint arXiv:2405.01359*, 2024.

[4] FERMILAB-CONF-24-0237-AD.

[5] F. Rehm, J. M. Guijarro, N. Soufflet, and V. Kain, *AccGPT: A Vision for AI Assistance at CERN's Accelerator Control and Beyond*, 2024.

[6] J. Wei *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[7] S. Yao *et al.*, "React: Synergizing reasoning and acting in language models," *arXiv preprint arXiv:2210.03629*, 2022.

[8] J. Maldonado, S. Clark, W. Fu, and S. Nemesure, "Enhancing Electronic Logbooks Using Machine Learning," in *Proc. 19th Int. Conf. Accel. Large Exp. Phys. Control Syst. (ICALEPCS'23)*, Cape Town, South Africa, Oct. 9–13, 2023, 2024, pp. 382–385, `10.18429/JACoW-ICALEPCS2023-TUMBCMO15`

[9] A. Sulc, R. Kammering, A. Eichler, and T. Wilksen, "Pacuna: Automated fine-tuning of language models for particle accelerators," *arXiv preprint arXiv:2310.19106*, 2023.

[10] T. Gao, X. Yao, and D. Chen, "Simcse: Simple contrastive learning of sentence embeddings," *arXiv preprint arXiv:2104.08821*, 2021.

[11] A. Shakir, D. Koenig, J. Lipp, and S. Lee, *Boost your search with the crispy mixedbread rerank models*, `https://www.mixedbread.ai/blog/mxbai-rerank-v1`, Accessed: 2024-05-07, 2024.

[12] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[13] R. Rehurek and P. Sojka, "Gensim–python framework for vector space modelling," *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, vol. 3, no. 2, 2011.

[14] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International conference on machine learning*, PMLR, 2014, pp. 1188–1196.

[15] J. Maldonado, S. Clark, and S. Nemesure. "Improving electronic logbook searches using natural language processing." (2024), `https://www.indico.kr/event/47/contributions/514/`

[16] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019, `https://arxiv.org/abs/1908.10084`

[17] D. K. Sean Lee Aamir Shakir and J. Lipp. "Open source strikes bread - new fluffy embeddings model." (2024), `https://www.mixedbread.ai/blog/mxbai-embed-large-v1`

[18] X. Li and J. Li, "Angle-optimized text embeddings," *arXiv preprint arXiv:2309.12871*, 2023.

[19] Z. Nussbaum, J. X. Morris, B. Duderstadt, and A. Mulyar, *Nomic embed: Training a reproducible long context text embedder*, 2024.

[20] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2019.

[21] L. Zhang *et al.*, "Tinychart: Efficient chart understanding with visual token merging and program-of-thoughts learning," *arXiv preprint arXiv:2404.16635*, 2024.

[22] T. Zhang *et al.*, "Raft: Adapting language model to domain specific rag," *arXiv preprint arXiv:2403.10131*, 2024.

[23] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[24] D. Roy *et al.*, "Exploring llm-based agents for root cause analysis," *arXiv preprint arXiv:2403.04123*, 2024.

[25] T. Ahmed, S. Ghosh, C. Bansal, T. Zimmermann, X. Zhang, and S. Rajmohan, "Recommending root-cause and mitigation steps for cloud incidents using large language models," in *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, IEEE, 2023, pp. 1737–1749.

[26] J. Kaiser, A. Eichler, and A. Lauscher, "Large language models for human-machine collaborative particle accelerator tuning through natural language," *arXiv preprint arXiv:2405.08888*, 2024.

[27] K. J. Hazelwood, D. Finstrom, M. McCusker-Whiting, and L. G. Mills, "The fermilab accelerator division electronic logbook (adel) at 10 years," 2023, `10.2172/2246727`

[28] T. Njekeu, "(noice) neural optical image categorizer for the e-log," 2020, `https://www.osti.gov/biblio/1706139`

[29] Qdrant, *Qdrant vector database*, version 1.9.2, 2024, `https://https://qdrant.tech/`

[30] N. Reimers and I. Gurevych, "The curse of dense low-dimensional information retrieval for large index sizes," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2021, pp. 605–611, `https://arxiv.org/abs/2012.14210`