Potental Uses of (Large) Language Models for DESY

Antonin Sulc, Annika Eichler, Tim Wilksen Hamburg,





Rule 1: Do not talk about LLMs!



Rule 1: Do not talk about LLMs! unless you find them useful.





Source books, conference proceedings, and arxiv preprints as PDFs.

PACuna: Automated Fine-Tuning of Language Models for Particle Accelerators

Antonin Sulc* DESY, Hamburg, Germany antonin.sulc@desy.de Raimund Kammering DESY, Hamburg, Germany Annika Eichler DESY, Hamburg, Germany

Tim Wilksen DESY, Hamburg, Germany

Abstract

Navigating the landscape of particle accelerators has become increasingly challenging with recent surges in contributions. These intricate devices challenge comprehension, even within individual facilities. To address this, we introduce PACuna, a fine-tuned language model refined through publicly available accelerator resources like conferences, pre-prints, and books. We automated data collection and question generation to minimize expert involvement and make the data publicly available. PACuna demonstrates proficiency in addressing accelerator questions, validated by experts. Our approach shows adapting language models to scientific domains by fine-tuning technical texts and auto-generated corpora capturing the latest developments can further produce pre-trained models to answer some specific questions that commercially available assistants cannot and can serve as intelligent assistants for individual facilities.





- Source books, conference proceedings, and arxiv preprints as PDFs.
- Training a LLM without a human in the loop.

PACuna: Automated Fine-Tuning of Language Models for Particle Accelerators

Antonin Sulc* DESY, Hamburg, Germany antonin.sulc@desy.de Raimund Kammering DESY, Hamburg, Germany Annika Eichler DESY, Hamburg, Germany

Tim Wilksen DESY, Hamburg, Germany

Abstract

Navigating the landscape of particle accelerators has become increasingly challenging with recent surges in contributions. These intricate devices challenge comprehension, even within individual facilities. To address this, we introduce PACuna, a fine-tuned language model refined through publicly available accelerator resources like conferences, pre-prints, and books. We automated data collection and question generation to minimize expert involvement and make the data publicly available. PACuna demonstrates proficiency in addressing accelerator questions, validated by experts. Our approach shows adapting language models to scientific domains by fine-tuning technical texts and auto-generated corpora capturing the latest developments can further produce pre-trained models to answer some specific questions that commercially available assistants cannot and can serve as intelligent assistants for individual facilities.





- Source books, conference proceedings, and arxiv preprints as PDFs.
- Training a LLM without a human in the loop.
- Showing improved performance over general chatbots like ChatGPT and Falcon.

PACuna: Automated Fine-Tuning of Language Models for Particle Accelerators

Antonin Sulc^{*} DESY, Hamburg, Germany antonin.sulc@desy.de Raimund Kammering DESY, Hamburg, Germany Annika Eichler DESY, Hamburg, Germany

Tim Wilksen DESY, Hamburg, Germany

Abstract

Navigating the landscape of particle accelerators has become increasingly challenging with recent surges in contributions. These intricate devices challenge comprehension, even within individual facilities. To address this, we introduce PACuna, a fine-tuned language model refined through publicly available accelerator resources like conferences, pre-prints, and books. We automated data collection and question generation to minimize expert involvement and make the data publicly available. PACuna demonstrates proficiency in addressing accelerator questions, validated by experts. Our approach shows adapting language models to scientific domains by fine-tuning technical texts and auto-generated corpora capturing the latest developments can further produce pre-trained models to answer some specific questions that commercially available assistants cannot and can serve as intelligent assistants for individual facilities.



- Source books, conference proceedings, and arxiv preprints as PDFs.
- Training a LLM without a human in the loop.
- Showing improved performance over general chatbots like ChatGPT and Falcon.
- > Use: Search, Validation, Checking

PACuna: Automated Fine-Tuning of Language Models for Particle Accelerators

Antonin Sulc^{*} DESY, Hamburg, Germany antonin.sulc@desy.de Raimund Kammering DESY, Hamburg, Germany Annika Eichler DESY, Hamburg, Germany

Tim Wilksen DESY, Hamburg, Germany

Abstract

Navigating the landscape of particle accelerators has become increasingly challenging with recent surges in contributions. These intricate devices challenge comprehension, even within individual facilities. To address this, we introduce PACuna, a fine-tuned language model refined through publicly available accelerator resources like conferences, pre-prints, and books. We automated data collection and question generation to minimize expert involvement and make the data publicly available. PACuna demonstrates proficiency in addressing accelerator questions, validated by experts. Our approach shows adapting language models to scientific domains by fine-tung technical texts and auto-generated corpora capturing the latest developments can further produce pre-trained models to answer some specific questions that commercially available assistants cannot and can serve as intelligent assistants for individual facilities.



- Source books, conference proceedings, and arxiv preprints as PDFs.
- Training a LLM without a human in the loop.
- Showing improved performance over general chatbots like ChatGPT and Falcon.
- > Use: Search, Validation, Checking Challenge: Hallucation

PACuna: Automated Fine-Tuning of Language Models for Particle Accelerators

Antonin Sulc* DESY, Hamburg, Germany antonin.sulc@desy.de

Raimund Kammering DESY, Hamburg, Germany Annika Eichler DESY, Hamburg, Germany

Tim Wilksen DESY, Hamburg, Germany

Abstract

Navigating the landscape of particle accelerators has become increasingly challenging with recent surges in contributions. These intricate devices challenge comprehension, even within individual facilities. To address this, we introduce PACuna, a fine-tuned language model refined through publicly available accelerator resources like conferences, pre-prints, and books. We automated data collection and question generation to minimize expert involvement and make the data publicly available. PACuna demonstrates proficiency in addressing accelerator questions, validated by experts. Our approach shows adapting language models to scientific domains by fine-tung technical texts and auto-generated corpora capturing the latest developments can further produce pre-trained models to answer some specific questions that commercially available assistants cannot and can serve as intelligent assistants for individual facilities.





How many superconducting radio frequency (SRF) cavities does European XFEL have?







How many superconducting radio frequency (SRF) cavities does European XFEL have? Ours: European XFEL has a total of 800 superconducting radio frequency

(SRF) cavities. partially correct (808)







How many superconducting radio frequency (SRF) cavities does European XFEL have?

<u>Ours</u>: European XFEL has a total of 800 superconducting radio frequency (SRF) cavities. partially correct (808)

<u>falcon-180b</u>: The European XFEL has a total of 16 superconducting radio-frequency (SRF) cavity stations that contain approximately 170 individual SRF cavities... Wrong







How many superconducting radio frequency (SRF) cavities does European XFEL have?

<u>Ours</u>: European XFEL has a total of 800 superconducting radio frequency (SRF) cavities. partially correct (808)

<u>falcon-180b</u>: The European XFEL has a total of 16 superconducting radio-frequency (SRF) cavity stations that contain approximately 170 individual SRF cavities... wrong

<u>ChatGPT</u>: As of my last knowledge update in September 2021, the European XFEL (X-ray Free-Electron Laser) facility had a total of 100 superconducting radiofrequency (SRF) cavities. wrong



LLM - Future - Multimodality



Human How should I make something like this?





LLM - Future - Multimodality



Human How should I make something like this?

MiniGPT-4 To make this dish, you will need the following ingredients: No Q-Former

- * 4 large live lobsters
- * 2 tablespoons olive oil
- * Salt and pepper
- * 2 cloves garlic, minced
- * 1 lemon, juiced
- * 1 tablespoon chopped fresh parsley

Here's how to make it: 1. Preheat your oven to 425°F (220°C).





> Improve our logbook (root-cause analysis, automated problem solving etc.).





- > Improve our logbook (root-cause analysis, automated problem solving etc.).
- > Automate some procedures (writing logbook entries, writing documents, documentations)





- Improve our logbook (root-cause analysis, automated problem solving etc.).
- > Automate some procedures (writing logbook entries, writing documents, documentations)
- > Assessment (of documents).

DESY.¹ Findables Accessible Materoperable Rei Sable ichler, Tim Wilksen | Hamburg,





- Improve our **logbook** (root-cause analysis, automated problem solving etc.).
- Automate some procedures (writing logbook entries, writing documents, documentations)
- > Assessment (of documents).
- Can improve the FAIR-ness.¹

DESY¹ Findable, Accessible, Interoperable, Reusable ichler Tim Wilksen I Hamburg



- Improve our logbook (root-cause analysis, automated problem solving etc.).
- Automate some procedures (writing logbook entries, writing documents, documentations)
- > Assessment (of documents).
- Can improve the FAIR-ness.¹
- And who knows what future holds?

¹Findable, Accessible, Interoperable, Reusable, tim Wilkson L Hamburg



Observation: Sometimes it works to just wait.





 Log anomaly detection using word embeddings and Hidden Markov Models (HMMs have a very few parameters!).

LOG ANOMALY DETECTION ON EUXFEL NODES

A. Sulc*, A. Eichler, T. Wilksen, DESY, Hamburg, Germany

Abstract

This article introduces a method to detect anomalies in the tog data generated by control system nodes at the European XFEL accelerator. The primary aim of this proposed method is provide operators a comprehensive understanding of the availability, status, and problems specific to each node. This information is vitor a comprehensive smooth operation. The sequential nature of logs and the absence of a rich text corput hat is specific to our node, posses significant limitations for traditional and learning-based approaches for anomaly detection. To overcome this limitation, we propose a method that uses word embedding and models individual nodes as a sequence of these vectors that commonly co-occur, using a Hidden Markov Model (HMM). We score individual log curitis by computing a probability ratio between the proba-

to mitigate potential problems from arising. Monitoring the logs of the watchdog nodes by textual analysis of their logs not only provides an automated means of comprehending the European XFEL accelerator system conditions but also enables early detection and resolution of issues that would otherwise only gain significance in the event of a specific node failure.

The structure of the paper is the following: First, we summarize the related work in log anomaly detection. In the next section, we show four main steps of our approach with important justifications and examples. Lastly, we show several examples and sketch a potential future work in this field.



- Log anomaly detection using word embeddings and Hidden Markov Models (HMMs have a very few parameters!).
- Represents logs as vectors (Word2Vec), and models their representations as HMMs.

LOG ANOMALY DETECTION ON EUXFEL NODES

A. Sulc*, A. Eichler, T. Wilksen, DESY, Hamburg, Germany

Abstract

This article introduces a method to detect anomalies in the top data generately ocotrol system modes at the European XFEL accelerator. The primary aim of this propoed method is provide operators a comprehensive understanding of the availability, status, and problems specific to each node. This information is vitor as comprehensive smooth operation. The sequential nature of logs and the absence of a rich text corpus that is specific to our nodes posses significant limitations for traditional and learning-based approaches for anomaly detection. To overcome this limitation, we propose a method that uses word embedding and models individual nodes as a sequence of these vectors that commonly co-occcur, using a Hidden Markov Model (HMM). We score individual log curities by computing a probability ruito between the probe

to mitigate potential problems from arising. Monitoring the logs of the watchdog nodes by textual analysis of their logs not only provides an automated means of comprehending the European XFEL accelerator system conditions but also enables early detection and resolution of issues that would otherwise only gain significance in the event of a specific node failure.

The structure of the paper is the following: First, we summarize the related work in log anomaly detection. In the next section, we show four main steps of our approach with important justifications and examples. Lastly, we show several examples and sketch a potential future work in this field.



- Log anomaly detection using word embeddings and Hidden Markov Models (HMMs have a very few parameters!).
- Represents logs as vectors (Word2Vec), and models their representations as HMMs.

DESY

Scores entries by probability ratio to detect anomalies (how well is the message fitting to the sequence).

LOG ANOMALY DETECTION ON EUXFEL NODES

A. Sulc*, A. Eichler, T. Wilksen, DESY, Hamburg, Germany

Abstract

This stride introduces a method to detect anomalies in the tog data generately by control system nodes at the European XFEL accelerator. The primary aim of this proposed method is provide operators a comprehensive understanding of the availability, status, and problems specific to each node. This information is visit for ensuring the smooth operation. The sequential nature of logs and the absence of a rich text corput hat is specific to our nodes poses significant limitations for traditional and learning-based approaches for anomaly detection. To overcome this limitation, we propose a method that uses word embedding and models individual nodes as a sequence of these vectors that corrumoly co-occur, using a Hidden Markov Model (HMM). We score individual to curitis by computing a probability ratio between the proba-

to mitigate potential problems from arising. Monitoring the logs of the watchdog nodes by textual analysis of their logs not only provides an automated means of comprehending the European XFEL accelerator system conditions but also enables early detection and resolution of issues that would otherwise only gain significance in the event of a specific node failure.

The structure of the paper is the following: First, we summarize the related work in log anomaly detection. In the next section, we show four main steps of our approach with important justifications and examples. Lastly, we show several examples and sketch a potential future work in this field.





- Log anomaly detection using word embeddings and Hidden Markov Models (HMMs have a very few parameters!).
- Represents logs as vectors (Word2Vec), and models their representations as HMMs.
- Scores entries by probability ratio to detect anomalies (how well is the message fitting to the sequence).
- Tested on EuXFEL logs, identifies score spikes corresponding to errors.

LOG ANOMALY DETECTION ON EUXFEL NODES

A. Sulc*, A. Eichler, T. Wilksen, DESY, Hamburg, Germany

Abstract

This stricle introduces a method to detect anomalies in the tog data generated by control system nodes at the European XFEL accelerator. The primary aim of this proposed method is provide operators a comprehensive understanding of the availability, status, and problems specific to each node. This information is visit for ensuring the smooth operation. The sequential nature of logs and the absence of a rich text corpus that is specific to our nodes poses significant limitations for traditional and learning-based approaches for anomaly detection. To overcome this limitation, we propose a method that uses word embedding and models individual nodes as a sequence of the vectors that commonly co-occur, using a Hidden Markov Model (HMM). We score individual to curities by computing a probability ruito between the probacurities by computing a probability ruito between the proba-

to mitigate potential problems from arising. Monitoring the logs of the watchdog nodes by textual analysis of their logs not only provides an automated means of comprehending the European XFEL accelerator system conditions but also enables early detection and resolution of issues that would otherwise only gain significance in the event of a specific node failure.

The structure of the paper is the following: First, we summarize the related work in log anomaly detection. In the next section, we show four main steps of our approach with important justifications and examples. Lastly, we show several examples and sketch a potential future work in this field.



(TEST,OK,





(TEST,OK,TEST,OK,



(TEST,OK,TEST,OK,TEST,OK,





(TEST,OK,TEST,OK,TEST,OK,TEST,OK)











(TEST,OK,





(TEST,OK,TEST,OK,





(TEST,OK,TEST,OK,TEST,OK,



)



(TEST,OK,TEST,OK,TEST,OK,TEST,**TEST**)









(TEST,OK,





(TEST,OK,TEST,OK,





(TEST,OK,TEST,OK,TEST,OK,







(TEST,OK,TEST,OK,TEST,OK,TEST,**ERROR**)









Log Anomaly Detection - Real Example







```
from hmmlearn import hmm
import numpy as np
x = np.stack([[0,1],[1,0],[0,1],[1,0],[0,1],[1,0],[0,1],[1,0]])
model = hmm.GaussianHMM(n_components=2, covariance_type="diag")
model.fit(x[:-1,:])
loqp = []
for i in range(1, x.shape[0]+1):
    logp.append(model.score(x[:i]))
logp = np.array(logp)
score = logp[:-1] - logp[1:]
```





Texts Expose Accelerator Secrets

> Analyze conference proceedings to reveal research trends, topics, and collaborations.





Texts Expose Accelerator Secrets

- Analyze conference proceedings to reveal research trends, topics, and collaborations.
- Semantic (text) search, topic modeling, and graph analysis methods





Texts Expose Accelerator Secrets

- Analyze conference proceedings to reveal research trends, topics, and collaborations.
- Semantic (text) search, topic modeling, and graph analysis methods
- > Uncovers latent topical structures.





Rule 2: Follow formatting rules and notation if you want to get your work recognized.







Thank you!

Contact

Deutsches Elektronen-Synchrotron DESY

www.desy.de

Antonin Sulc, Annika Eichler, Tim Wilksen 0000-0001-7767-778X MCS antonin.sulc@desy.de





