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

DESY. HELMHOLTZ

NEURAL INFORMATION PROCESSING SYSTEMS

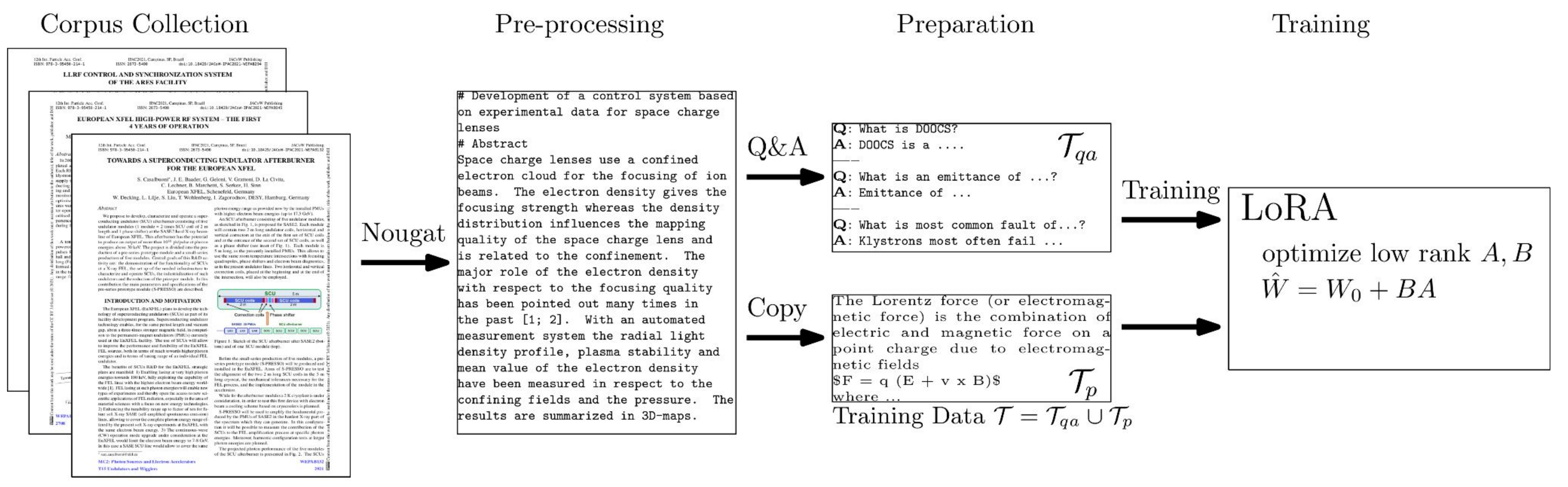
Antonin Sulc, Raimund Kammering, Annika Eichler, Tim Wilksen

## 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.

## Motivation

- Modern AI assistants like ChatGPT and Claude show impressive creativity but rely on questionable internet sources and can hallucinate incorrect responses when lacking knowledge.
- The complexity of large commercial language models makes it difficult to easily fine-tune them with the latest advancements.
- Fine-tuning smaller, powerful models on reputable sources like scientific papers creates more reliable assistants for specific domains.
- Scientific papers provide a trustworthy training dataset to create assistants with domain-specific language.
- These can improve FAIR principles and help in diverse applications like particle accelerator logbooks.
- Parameter-efficient fine-tuning methods adapt smaller models into domain-specific assistants able to handle complex conceptual questions.
- The challenge is automating preparation of high-quality, domain-specific training data.
- We present an approach that automates data preparation for fine-tuning publicly available models to train customized AI assistants.
- We introduce the first accelerator assistant, trained without human involvement in dataset preparation, that can answer questions commercial chatbots cannot.
- We show an automated pipeline to enable this for arbitrary domains without manually creating a supervised dataset



```
# Development of a control system based
on experimental data for space charge
lenses
# Abstract
Space charge lenses use a confined
electron cloud for the focusing of ion
beams.  The electron density gives the
focusing strength whereas the density
distribution influences the mapping
quality of the space charge lens and
is related to the confinement.  The
major role of the electron density
with respect to the focusing quality
has been pointed out many times in
the past [1; 2].  With an automated
measurement system the radial light
density profile, plasma stability and
mean value of the electron density
have been measured in respect to the
confining fields and the pressure.  The
results are summarized in 3D-maps.
```

**Q**: What is DOOCS?
**A**: DOOCS is a ....

---

**Q**: What is an emittance of ...?
**A**: Emittance of ...

---

**Q**: What is most common fault of...?
**A**: Klystrons most often fail ...

The Lorentz force (or electromagnetic force) is the combination of electric and magnetic force on a point charge due to electromagnetic fields
$F = q (E + v x B)$
where ...

Training Data $\mathcal{T} = \mathcal{T}_{qa} \cup \mathcal{T}_p$

## Pipeline

1. **Collect**
2. **Pre-process**
3. **Auto-generate**
4. **Fine-tune (LoRA, eventually LongLoRA)**

### 1. Collect

To create a robust dataset, this work uses three complementary sources:

- **books** for foundational knowledge,
- **conference proceedings** containing recent facility developments,
- **arXiv pre-prints** in accelerator physics since 2015.

### 2. Pre-process

Nougat OCR digitizes documents into machine-readable text, equations, and tables. These outputs are transformed into a common format of **Markdown text**, plain text tables, and LaTeX equations. arXiv sources in LaTeX are also converted to this format using Pandoc.

### 3. Pre-process

To generate supervised **question-answer pairs**, `vicuna-7B-16k-v1.5` is prompted with

**`"Generate ten questions for a paper:"$TEXT""`**

for sections of books and papers, discarding improperly formatted responses.

## References

- Blecher, L., Cucurull, G., Scialom, T. and Stojnic, R., 2023. Nougat: Neural optical understanding for academic documents.
- Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I. and Xing, E.P., 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. https://lmsys.org/blog/2023-03-30-vicuna/
- Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. and Chen, W., 2021. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.

## Examples

**Question** about DESY ARES : *Does ARES use normal or superconducting cavities?*

- PACuna: *ARES uses normal cavities for its accelerator*, Correct
- falcon-180b: *I do not have information about a specific ares system. However, ares is a general acronym that can refer to various systems....* Wrong, the answer is missing.
- ChatGPT: *As of my last knowledge update in September 2021, ARES (Accelerator for Research in Subatomic Physics) had plans to use superconducting radiofrequency (SRF) cavities. ...*, Wrong, they are not superconducting.

**Question about DESY European XFEL**: *How many superconducting radio frequency (SRF) cavities does European XFEL have?*

- PACuna: *European XFEL has a total of 800 superconducting radio frequency (SRF) cavities.*, Approximately correct, it is 808 .
- falcon-180b: *The European XFEL has a total of 16 superconducting radio-frequency (SRF) cavity stations that contain approximately 170 individual SRF cavities...*, Wrong.
- ChatGPT: *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.

**Question about beam emittance at REAGE**: *What is the significance of measuring the beam emittance in the REGAE linear accelerator?*

- PACuna:*Measuring the beam emittance in the REGAE linear accelerator is significant because it enables the calculation of the beam's emittance, which is a measure of the beam's transverse size and intensity. This information is crucial for optimizing the performance of the REGAE linear accelerator and the other devices it serves*, Partially correct, word intensity is not correct.
- falcon-180b: *Measuring the beam emittance in the REGAE linear accelerator is significant because it allows for the optimization of the accelerator's performance. Emittance is a measure of how tightly packed the particles in the beam are, and a smaller emittance means a more focused and uniform beam...*, Correct.
- ChatGPT: *Measuring the beam emittance in a linear accelerator like REGAE (Relativistic Electron Gun for Atomic Exploration) is of significant importance because emittance is a fundamental parameter that characterizes the quality of the charged particle beam. The emittance measurement provides crucial information about the beam's properties, and it has several significant implications for accelerator operation and scientific experiments...*, Correct.