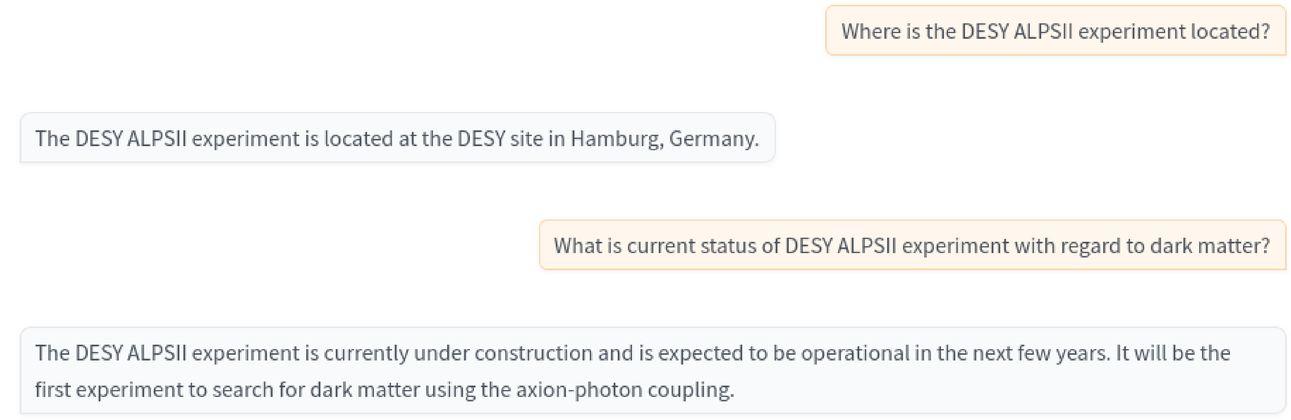


https://github.com/sulcantonin/CHATQCD_ICHEP24

ChatQCD: Let Large Language Models Explore QCD

Antonin Sulc *(HZB), Patrick L.S. Connor (UHH)*

Quantum chromodynamics (QCD) has yielded a vast literature spanning distinct phenomena. We construct a corpus of papers and build a generative model. This model holds promise for accelerating the capability of scientists to consolidate our knowledge of QCD by the ability to generate and validate scientific works in the landscape of works related to QCD and similar problems in HEP. Furthermore, we discuss challenges and future directions of using large language models to integrate our scientific knowledge about QCD through the automated generation of explanatory scientific texts.

MOTIVATION

- Broad QCD literature
- Modern tools can accelerate knowledge acquisition in QCD
- LLMs could enhance QCD research accessibility.
- Potential model for knowledge consolidation in other fields.

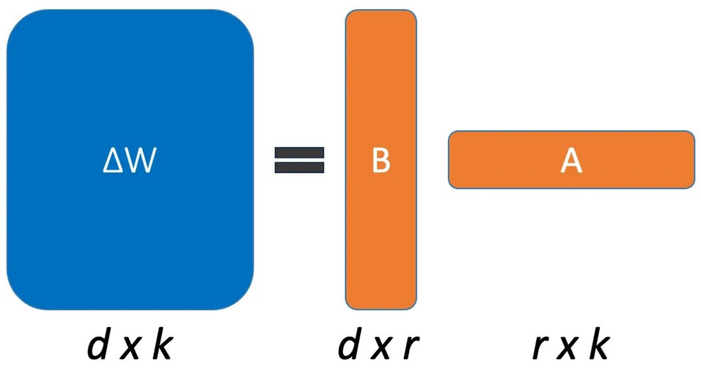
DATASET

- The training dataset consist of
- arxiv 45422 PDF documents
- PDF to text with Nougat OCR [1]
- Two datasets
  - **Unsupervised** from raw documents (incl. tables and latex formulas)
  - **Supervised** generated from
    - *Paragraph* of orig. paper
    - *Summary* from whole paper (or up to max.len)
    - *Prompt* that asks for JSON output

FINETUNING

Unsloth Framework [2] : A lightweight framework optimized for efficient fine-tuning of LLMs.

LoRA [3] (Low-Rank Adaptation): Parameter efficient fine-tuning via matrix multiplication



APPLICATIONS

- **Querying** current QCD knowledge
- **Validation** from measuring likelihood of input
- **Data mining** on the fine-tuned model

PARAMETERS

```
LORA(
 r = 32,
 target_modules = ["q_proj",
   "k_proj","v_proj", "o_proj",
   "gate_proj", "up_proj", "down_proj",],
 lora_alpha = 8,
 lora_dropout = 0,
 bias = "none")
```

```
TrainingArguments(
 per_device_train_batch_size = 16,
 gradient_accumulation_steps = 32,
 warmup_steps = 5,
 num_train_epochs = 1,
 learning_rate = 2e-4,
 optim = "adamw_8bit",
 weight_decay = 0.005,
 lr_scheduler_type = "linear",
 seed = 3407)
```

CONCLUSION

- Generative model shows promise for QCD knowledge consolidation.
- Potential for accelerating progress in QCD and related fields like knowledge consolidation
- Challenges remain but future refinements could enhance impact.
- AI-assisted research tools may become integral to scientific discovery

REFERENCES

[1] Blecher, L., Cucurull, G., Scialom, T. and Stojnic, R., 2023. Nougat: Neural Optical Understanding for Academic Documents. arXiv preprint arXiv:2308.13418.
[2] unslothai, 2023. unsloth. [online] GitHub. Available at: https://github.com/unslothai/unsloth [Accessed 14 July 2024].
[3] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S. and Chen, W., 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv preprint arXiv:2106.09685.