PROCEEDINGS OF SCIENCE

ChatQCD: Let Large Language Models Explore QCD

Antonin Sulc[a,*] **and Patrick L.S. Connor**[b]

[a] *Helmholtz-Zentrum Berlin fuer Materialien und Energie, Albert-Einstein-Str. 15, 12489 Berlin, Germany*

[b] *Center for Data and Computing in Natural Sciences, Universität Hamburg, Germany, Albert-Einstein-Ring 10, D-22761 Hamburg*

E-mail: antonin.sulc@helmholtz-berlin.de, patrick.connor@desy.de

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 their 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.

42nd International Conference on High Energy Physics
17-24 July 2024
Prague Congress Centre, Prague, Czech Republic

*Speaker

1. Introduction

This paper introduces ChatQCD, an approach that utilizes large language models (LLMs) to explore and consolidate QCD knowledge from a comprehensive corpus of research papers. The motivation for this work is multifaceted: the complexity of QCD literature necessitates advanced tools for efficient knowledge synthesis; modern artificial intelligence techniques offer promising avenues for accelerating scientific understanding, and LLMs have the potential to enhance the accessibility of QCD research for both experts and newcomers. Furthermore, this approach may serve as a model for knowledge consolidation in other scientific disciplines by leveraging existing research pipelines.

2. Method

Dataset. We compiled a dataset of 45,422 arXiv PDF documents, which were converted to an uniform text format using Nougat OCR [1]. This tool efficiently processes common scientific paper formats, exporting formulas and tables in LaTeX and structuring text into sections. The training dataset comprises two components: (1) an unsupervised dataset derived from raw documents, including tables and LaTeX formulas, and (2) a supervised dataset generated through a self-instruct [2]. It consists of question-answer pairs generated by `Mistral-7B-Instruct-v0.3` model by prompting paper paragraphs to 10 question-answer pairs in JSON format.

Model Architecture and Fine-tuning. We utilized the Unsloth Framework [3], a lightweight system optimized for efficient fine-tuning of LLMs. The base model is `llama-3-8b-Instruct-bnb-4bit`. The model was fine-tuned using Low-Rank Adaptation (LoRA) [4], a parameter-efficient technique that adapts the model via matrix multiplication. Key parameters for the LoRA configuration included a rank (r) of 32 and an α value of 8. Training arguments were configured as follows: per-device train batch size was set to 16, gradient accumulation steps to 32, with only 5 warmup steps and 1 training epoch. The learning rate (γ) was set to 2^{-4}, using an 8-bit AdamW optimizer regularized by a weight decay λ of 0.005, with linear scheduling of the learning rate.

3. Experiments and Evaluations

Question (1): What's the QCD vacuum's θ parameter value?

Claude 3.5: ✓≈ 0
Base: ✓0.
ChatQCD: ✓The QCD vacuum's θ parameter value is zero.

Question (2): What's the β function's leading coefficient in QCD?

Claude 3.5: ×$-11N_c/3$
Base: ×1.
ChatQCD: ✓The beta function's leading coefficient in QCD is $b_0 = \frac{1}{(4\pi)^2}\left(\frac{11}{3}N_c - \frac{2}{3}N_f\right)$

Question (3): How does the QCD coupling constant change with energy scale?

Claude 3.5: ✓It decreases at higher energies due to asymptotic freedom.
Base: (50/50)The QCD coupling constant, α_s, decreases with increasing energy scale (or decreasing momentum transfer) according to the renormalization group equation.
ChatQCD: ✓The QCD coupling constant decreases with energy scale.

Question (4): Why is QCD challenging to solve in the low-energy regime?

Claude 3.5: ✓Strong coupling makes perturbative methods ineffective at low energies.
Base: ✓QCD is challenging to solve in the low-energy regime because the theory is non-perturbative, meaning that the calculations become increasingly difficult as the energy scale decreases.
ChatQCD: (50/50)QCD is challenging to solve in low-energetic regime because it is a non-perturbative theory that involves strong interactions, which are difficult to handle using perturbation theory.

We analyze the logits, which serve as indicators of an LLM's confidence in predicting the next token in a sequence. We present two sequences where correct and incorrect statements are provided to the model, followed by a measurement of their respective logits (incorrect options are ~~striked~~): (1) `How many gluons exist? (8|~~9~~)` The logit for the correct answer `8` was ≈ 3.53, while for the incorrect `9`, it was ≈ 1.14. (2) `In QCD, quarks interact via the exchange of gluons, which carry (color|~~mass~~) charge.` The logit for the correct term `color` was ≈ 10.73, whereas for the incorrect `mass`, it was ≈ 8.20. These results demonstrate that in both cases, the logits (i.e., the model's confidence) are higher for the correct statements.

4. Conclusion

ChatQCD demonstrates a promising LLM alternative to commercial models. By utilizing a dataset of arXiv papers and employing advanced fine-tuning techniques, we have created a model capable of generating relevant responses to QCD-related queries, in some cases surpassing existing commercial models (Question 2), but qualify of the output depends on formulation of the question. Our experiments show encouraging results regarding the model's confidence in predicting correct information. However, challenges remain, particularly in handling complex scientific concepts and ensuring factual accuracy.

We propose the following improvements: (1) enhanced pre-selection of high-quality publications, (2) implementation of advanced feedback mechanisms [5, 6], (3) refinement of supervised instructions to eliminate misleading references, and (4) additional validation of source materials, as publication on arXiv does not guarantee content validity.

Acknowledgement: We would like to thank to Radek Zlebcik for review of correctness of answers from our model.

Resources Resources available at https://github.com/sulcantonin/CHATQCD_ICHEP24

References

[1] L. Blecher, G. Cucurull, T. Scialom and R. Stojnic, *Nougat: Neural optical understanding for academic documents*, 2023.

[2] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N.A. Smith, D. Khashabi et al., *Self-instruct: Aligning language models with self-generated instructions*, *arXiv preprint arXiv:2212.10560* (2022) .

[3] unslothai, "unsloth." https://github.com/unslothai/unsloth, 2023.

[4] E.J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang et al., *Lora: Low-rank adaptation of large language models*, *arXiv preprint arXiv:2106.09685* (2021) .

[5] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin et al., *Training language models to follow instructions with human feedback*, *Advances in neural information processing systems* **35** (2022) 27730.

[6] H. Lee, S. Phatale, H. Mansoor, T. Mesnard, J. Ferret, K.R. Lu et al., *Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback*, in *Forty-first International Conference on Machine Learning*, 2023.