
ChatQCD: Let Large Language Models Explore QCD

Antonin Sulc*

Helmholtz-Zentrum Berlin fuer Materialien und Energie,
Albert-Einstein-Str. 15, 12489 Berlin, Germany
antonin.sulc@helmholtz-berlin.de

Patrick Connor

Center for Data and Computing in Natural Sciences,
Universität Hamburg, Germany,
Albert-Einstein-Ring 10, D-22761 Hamburg
patrick.connor@desy.de

Abstract

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.

1 Introduction

Quantum Chromodynamics (QCD), the theory of strong interactions, stands as one of the fundamental pillars of modern particle physics. Since its inception in the early 1970s, QCD has spawned a vast and intricate body of literature, encompassing a wide array of phenomena from the subatomic to the cosmic scale. The complexity and breadth of QCD research present significant challenges for scientists seeking to navigate and synthesize this extensive knowledge base efficiently.

In recent years, the rapid advancement of artificial intelligence, particularly in the domain of natural language processing, has opened new avenues for knowledge consolidation and discovery. Large Language Models (LLMs) have demonstrated remarkable capabilities in understanding and generating human-like text across diverse domains. This paper introduces ChatQCD, an innovative approach that harnesses the power of LLMs to explore and consolidate QCD knowledge from a comprehensive corpus of research papers.

The motivation for this work is multifaceted. Firstly, the sheer volume and complexity of QCD literature necessitate advanced tools for efficient knowledge synthesis. Traditional methods of literature review and knowledge aggregation are time-consuming and often limited by human cognitive constraints. Secondly, modern artificial intelligence techniques offer promising avenues for accelerating scientific understanding by identifying patterns, connections, and insights that might elude human researchers. Lastly, LLMs have the potential to enhance the accessibility of QCD research for both experts and newcomers to the field, facilitating knowledge transfer and interdisciplinary collaboration.

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

ChatQCD aims to serve as a model for knowledge consolidation in scientific disciplines by leveraging existing research pipelines and state-of-the-art AI technologies. By creating a specialized LLM trained on a curated corpus of QCD literature, we seek to develop a tool that can assist researchers in navigating complex QCD concepts, generating hypotheses, and identifying promising research directions.

In this paper, we detail the methodology behind ChatQCD, including the construction of our dataset, the architecture of our language model, and our fine-tuning approach. We present a series of experiments and evaluations designed to assess the model’s performance in understanding and generating QCD-related content. Furthermore, we discuss the challenges encountered in applying LLMs to highly specialized scientific domains and propose future directions for improving the integration of AI technologies in theoretical physics research.

2 Method

The development of ChatQCD involved several key steps, from dataset creation to model fine-tuning. This section outlines our methodology in detail, providing insights into the technical aspects of our approach.

2.1 Dataset

The foundation of ChatQCD is a carefully curated dataset of QCD-related research papers. We compiled a corpus of 45,422 arXiv PDF documents, encompassing a wide range of topics within the field of QCD. To ensure uniformity and machine-readability of the dataset, we employed the Nougat OCR tool [1], which is specifically designed to process scientific documents efficiently.

Nougat OCR offers several advantages for our purposes. It can handle common scientific paper formats, exporting formulas and tables in LaTeX format. This capability is crucial for preserving the mathematical content integral to QCD research. Additionally, Nougat structures the extracted text into sections, maintaining the logical flow of the original documents.

Our training dataset comprises two main components:

Unsupervised Dataset : This consists of raw text extracted from the documents, including tables and LaTeX formulas. This component preserves the original structure and content of the papers, allowing our model to learn from the natural flow of scientific discourse in QCD literature.

Supervised Dataset : To enhance the model’s ability to engage in question-answering tasks, we generated a supervised dataset through a self-instruct method [6]. Using the Mistral-7B-Instruct-v0.3 model, we processed paper paragraphs to create 10 question-answer pairs in JSON format for each document. This approach allows us to train the model on targeted QCD-related queries, improving its performance in interactive scenarios.

2.2 Model Architecture and Fine-tuning

For the core of ChatQCD, we utilized the Unsloth Framework [5], a lightweight system optimized for efficient fine-tuning of LLMs. Our base model is llama-3-8b-Instruct-bnb-4bit, chosen for its balance of performance and computational efficiency.

To adapt this general-purpose model to the specific domain of QCD, we employed Low-Rank Adaptation (LoRA) [2]. LoRA is a parameter-efficient fine-tuning technique that adapts the model via matrix multiplication, allowing for significant performance improvements without the need to retrain the entire model.

Key parameters for our LoRA configuration included:

- Rank (r): 32
- Alpha (α): 8

Our training configuration was as follows:

- Per-device train batch size: 16
- Gradient accumulation steps: 32
- Warmup steps: 5
- Training epochs: 1
- Learning rate (γ): 2^{-4}
- Optimizer: 8-bit AdamW
- Weight decay (λ): 0.005
- Learning rate scheduling: Linear

This configuration was designed to balance training efficiency with model performance, allowing us to fine-tune ChatQCD effectively on our QCD-specific dataset.

In the following sections, we will discuss the experiments conducted to evaluate ChatQCD’s performance and analyze its capabilities in understanding and generating QCD-related content.

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 quality 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 [4, 3], (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.

References

- [1] Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. Nougat: Neural optical understanding for academic documents, 2023.
- [2] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- [3] Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, et al. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. In *Forty-first International Conference on Machine Learning*, 2023.
- [4] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

- [5] unslothai. unsloth. <https://github.com/unslothai/unsloth>, 2023. Accessed: 14 July 2024.
- [6] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.