SciGalactica: Integrating Nougat and Llama 2 for Comprehensive Scientific Data Synthesis

Younes Brahimi

November 17, 2023

Abstract

SciGalactica represents a groundbreaking endeavor in the realm of scientific research, harnessing the capabilities of advanced AI models Nougat and Llama 2 to synthesize and apply vast scientific knowledge. Inspired by the monumental Galactica model, this project focuses on streamlining the analysis and interpretation of scientific data across multiple disciplines, with an example on cryptography. SciGalactica aims to motivate the science community to overcome the challenges posed by the exponential growth of scientific data, offering a novel approach to navigating and leveraging this wealth of information for accelerated scientific discovery.

1 Introduction

The exponential growth of scientific information in the digital era poses significant challenges in data management and utilization. Projects like Galactica have pioneered the use of large language models to address these challenges, but there remains a gap in specialized, discipline-specific applications. SciGalactica is conceived as an extension of this pioneering work, leveraging the data processing capabilities of Nougat and the advanced modeling prowess of Llama 2. This project is not only an exploration in AI-driven data synthesis but also a testament to the potential of AI in revolutionizing scientific inquiry.

2 Methodology

2.1 Data Collection

Using a combination of automated scripts, APIs, and web scraping techniques, SciGalactica aggregates a vast corpus of scientific documents. The collection spans an extensive range of disciplines, with a particular focus on the evolving field of cryptography. This process involves not just the acquisition of data but also its categorization and preliminary analysis for relevance and quality. The dataset could be acquired similarly to Galactica's and filtered to our needs.

2.2 Data Processing with Nougat

The raw data undergoes transformation through Nougat's advanced OCR capabilities. This step is crucial for converting diverse document formats into a structured, machine-readable format. Special attention is given to maintaining the integrity of scientific notations, formulas, and diagrams. Although only 77% accurate on tables and graphs, there can be other solutions to checking them at the cost of performance. Or, we can wait and see Galactica using images to truly capture science. (Geometry in chemistry, graphs in computer science, etc.)

2.3 Model Training

Utilizing the processed data, Llama 2 is fine-tuned to not just understand and regurgitate information but to synthesize and apply it across contexts. This training emphasizes the development of a model that can reason, deduce, and generate novel insights in scientific domains.

2.4 Retreieval Augmented Generation

It is true that it has been shown how large language models can absorb large bodies of scientific knowledge, however, detaield context through retrieval has a place for fine-grained types of knowledge, and this is believed to heavily complement the flexible weight memory of the Transformer. Examples include LlamaIndex, ChatGPT with RAG, etc.

3 Implementation

The implementation of SciGalactica revolves around three core components: data integration, annotation, and training emphasis. By combining varied scientific data and employing targeted annotations, particularly in the realm of cryptography, the model is trained to effectively synthesize information from multiple sources. This includes linking theoretical concepts with practical applications and bridging gaps between seemingly disparate scientific domains.

4 Evaluation

SciGalactica's performance is evaluated based on its ability to accurately generate and reason about scientific content. The evaluation process includes:

- Generating responses to complex scientific queries.
- Demonstrating understanding and application of scientific principles.
- Comparing generated content with established scientific literature for accuracy and novelty.

- Combining and using multiple scientific phenomena to solve a problem.

Example questions for model testing include:

- Chemistry: "Sulfuric acid reacts with sodium chloride, giving ___ and ___."
- Physics: "Describe the principle behind the operation of a cyclotron."
- Mathematics: "Explain the concept of the Riemann hypothesis."
- Astrophysics: "What are the main characteristics of a neutron star?"
- Cryptography: "Outline the fundamental principles of the RSA algorithm."
- Aerospace: "Explain the aerodynamic scaling of ice accretions on airfoil"

5 Conclusion

SciGalactica marks a significant advancement in the field of AI-driven scientific research. By effectively synthesizing and applying knowledge across multiple scientific disciplines, this project not only demonstrates the potential of AI in managing the wealth of scientific data but also paves the way for future developments in AI-assisted scientific discovery and exploration.