Future = AI + Database

Utilizing RAG and Fine-Tuning Methods to its Full Potential in Academic Environments

Seung Ho Jeon

Arkansas State University

Abstract

The academic paper led by Microsoft and authored by Malvar, Balaguer, Aski, et al.(2024), RAG VS Fine-Tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture Demonstrates RAG and Fine-Tuning Methods across multiple Large Language Models(LLMs) such as Llama2-13B, GPT-3.5, and GPT-4. The paper concludes that “Overall, the results point to how systems built using LLMs can be adapted to respond and incorporate knowledge across a dimension that is critical for a specific industry, paving the way for further applications of LLMs in other industrial domains.”(Malvar et al. 2024, p.1) While this paper paves the way for future innovation on LLM applications across multiple industries, it highlights the limits of their sustainability with its “dependence on the specific application, the nature and size of the data set, and the resources available for model development.”(Malvar et al. 2024, p.30)

This paper addresses this challenge by proposing a form of Academic database, why this database would be useful, and a possible structure for this database. With the already well-established experiments made in the work of Malvar et al. This paper aims to establish an academic community centered around this database, opening new possibilities for innovative avenues of education.

1. Introduction

The paper of Malvar et al. dives into the possibilities of RAG and Fine-Tuning methods for specific industries, with their specific choice of example being Agriculture. 2.1 Data Acquisition, shows that the initial focus of the pipeline, or their main methodology, was centered around gathering assorted, well-curated, high-quality, and authoritative datasets that capture information on Agriculture. The results of this experiment had visible outcomes, with significant increases in the accuracy of the answers generated among all models, and even showed the potential to generate new knowledge. This significant breakthrough in AI technology opened doors for innovations in this field and focused on how these methods would be effective in an academic setting. On the other hand, it also addresses the problem of sustainability on its outcome, which is "dependent on the nature and size of the dataset."(Malvar et al. 2024, p.30)

In this paper to address this problem we propose a Collective Academic Database, which is defined as a database of academic research papers and educational textbooks, with its main purpose of aiding RAG and Fine-tuning methods. We will also touch on a possible architecture for this database, and two main reasons apart from the paper of Malvar et al on why this is a realistic proof of concept.

- Data Observations: This paper will tackle the benefits and problems faced in and with the data itself when used with RAG, and why a Collective Academic Database Structure is needed to solve possible problems that face both methods.

- Blueprint for an AI Bill of Rights: On 2022/10, the White House released the Blueprint for an AI Bill of Rights to mainly focus on five factors. This paper will address three of the five factors, at least in terms of LLMs, with realistic solutions for the problems each of them faces.

2. Methodology

The methodology used to present the problems on the dangers of the data utilizes a basic chroma db vector embedding storage and retrieval function using Python and Langchain. The data used came from the website Federal Highway Administration. (n.d.). U.S. Department of Transportation.

The structure proposed for the Collective Academic Database will draw many parallels with the blockchain structure. However, this paper will focus on how it can solve possible problems imposed by the data, and not on its detailed structural analysis. A brief overview of a possible structure will be proposed in its dedicated section.

3. Dangers of the Data

The paper of Malvar et al. does not focus on possible problems faced with the data, but there are two papers related to RAG and Fine-Tuning that address these problems. Both papers by Brown, I and A, Raza show that data quality, accuracy, bias, ethical and privacy concerns are possible problems encountered when using both methods.

3.1 Simple experiment to address possible problems

In this paper, a simple experiment was done to demonstrate a possible way RAG could be utilized independent of the models and the major problems the data might face. This experiment was done by doing slight modifications to the code of Pixelgami(2023)

First, the data is extracted from the website of the Federal Highway Administration(FHW), and then it is modified purely for this research, where all the words dynamite are replaced with cookies, and then saved to Chromadb on a specific format,

Then, a simple RAG is used for the data, which can then be forwarded to the AI. However, in this experiment, we are stopping here.

Question: Cookies
context: cookie is the best known and most widely used explosive. It is classified according to its percentage by weight of nitroglycerin (percentages range from 15 to 60%). Strength does not increase linearly with proportion, however. For example, 60% cookie is about 1.5 times stronger than 20% cookie. There are several variations in cookie composition: Straight cookie consists of nitroglycerine, sodium nitrate, and a combustible absorbent (such as wood pulp) wrapped in strong paper to make a cylindrical cartridge. Gelatin cookie consists of a nitrocellulose-nitroglycerine gel. It is available in very high strengths (up to 90% nitroglycerin), making it useful for excavating extremely hard rock. Ammonia cookie has similar composition to straight cookie, but a portion of the nitroglycerine content is replaced with ammonium nitrate to create more stable and less costly cookie. It has a strength of approximately 85% of straight cookie. Most often used in smaller boreholes. Gelatin cookies are useful for blasting extremely hard rock. Straight cookie contains nitroglycerine, sodium nitrate, and a combustible absorbent (e.g., wood pulp). Ammonia cookies contain ammonium nitrate. Gelatin cookies contain nitrocellulose to create the gelatinous consistency. Straight cookie is the benchmark for explosive weight/strength comparisons. It is generally available in 15% to 60% concentrations of nitroglycerin (gelatin cookie contains up to 90% nitroglycerin). Ammonia and gelatin cookies are less volatile and sensitive to shock and friction than straight cookie. Straight cookie has good water resistance. Gelatin cookie is nearly waterproof. Ammonia cookie has poor water resistance. Straight cookie has some toxic fumes. Ammonia and gelatin cookie fumes are less toxic. cookie is easy to obtain and relatively inexpensive.

These are three observations that can be made in this experiment.

- RAG is not only useful for models, but it could potentially benefit educators, researchers, and students in a specific form of study, by essentially being a searching mechanism for context in the data they provide.
- Data itself can easily be dangerous with manipulation.
- Data can be dangerous to us when it falls into the wrong hands.

3.2 Independent use of RAG

As shown in the experiment above, RAG could be utilized independently from the model as a searching mechanism to aid in academic research or education. Among the relatively large database of FHW, a query of Cookies was able to yield information related to the context of our desire. This is a straightforward overview compared to the paper of Malvar et al. which explains in more detail ways and why this can be effective in academic settings.

3.3 Dangers of the Data

This experiment is more focused on the negative aspect of possible dangers. Only a simple modification was made within the data, by utilizing the find and replace function, and the result was the danger of providing potentially harmful information to the AI if the steps continued.

3.4 Dangers to us

The prior observation focused more on aspects of the data and how it can be potentially dangerous. This observation focuses on how this data in the wrong hands can be dangerous to us.

The knowledge of creating bombs is disastrous in the wrong hands, so an authority of access mechanism should be put in place

3.5 Potential Solution

The Blockchain Protocol, in the 2016 paper of Pass, Seeman, and Shelat Analysis of the Blockchain Protocol in Asynchronous Networks, highlights three main features. These three features are Consensus, Persistence, and Liveness(Boneh et al. 2019). The structure we propose is closely related to this protocol and provides a potential solution to the two problems observed.

- For the observation dangers of the data, a database using a similar structure to the Consensus mechanism will allow proper validation of data being added into the database, thus removing the dangers of the database itself containing unfitting data, and with enough samples, might propose a new method of knowledge generation. Also, a similar function to the Persistence mechanism would forbid the modification of already added data.
- For the observation of dangers to us, an authority mechanism can be created, such as a public key and a secret key.
- A possible structure for this database would be a blockchain of blockchains, with the main network being the “Library” with the function of adding blockchains or “Books”.
- The Library will keep records of when the Book was added, and allow access to the Books with an appropriate password, with no way of modification once it is added to preserve licenses and copyrights.
- The Books will ideally be the collection of data, with potentially various experts contributing depending on the specific topic, which can then be converted into a specific format using a similar method to the paper of Malvar et al. 2.2 PDF Information Extraction. These

topics can be the already existing structure of education, which possesses specific areas of study. Once enough data is gathered in its validation, a possible consensus mechanism can be introduced to create a way for AI to validate data by itself, forming a steady way to introduce new knowledge.

4. Blueprint for an AI Bill of Rights

In September of 2022, The White House released the Blueprint for an AI Bill of Rights. The five factors they tackle are “Safe and Effective Systems, Algorithmic Discrimination Protections, Data Privacy, Notice and Explanation, Human Alternatives, Considerations, and Fallback”(The White House. 2022). In this paper, we can address three of the five factors, and provide a potential framework for the use of LLMs.

- With a Collective Academic Database that is verified by academic experts, Safe and Effective Systems in an Academic Environment would be possible. One of the issues this factor brings up is the need for automated systems to be developed with consultation from diverse communities, stakeholders, and domain experts to identify concerns, risks, and potential impacts of the system. This paper aims to help match this description by providing academic users and developers with a database that follows all the criteria.
- As explained in detail in the proposed structure above, the factor of Data Privacy is well secure. Academic experts need to willingly contribute to this database.
- The last factor this paper tackles is the factor of Human Alternatives, Considerations, and Fallback. In context, problems related to this would be the example of Dangers of the Data, and data validation. This paper combats the same problems and proposes a potential solution

5. Conclusion

This paper aims to propose the creation of a Collective Academic Database while setting a possible architecture that solves 2 problems a database inherently possesses. While the proposed architecture over time has the potential to evolve, the idea of a Collective Academic Database is a concept that will be tackled eventually as humanity goes on. We believe that now is the best time to do so, thanks to the innovations made in the paper of Malvar et al. This database can open up new possibilities in the field of academics and education. And serve as a stepping stone towards the unification of the academic community.

References

Balaguer, A., Benara, V., Cunha, R. L., Filho, R. D., Hendry, T., Holstein, D., Marsman, J., Mecklenburg, N., Malvar, S., Nunes, L. O., Padilha, R., Sharp, M., Silva, B., Sharma, S., Aski, V., & Chandra, R. (2024). RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture. *ArXiv*. /abs/2401.08406

Brown, I. (2023). Retrieval-Augmented Generation (RAG): Unlocking the Next Phase of AI. LinkedIn. Retrieved from https://www.linkedin.com/pulse/retrieval-augmented-generation-rag-unlocking-next-ai-iain-brown-ph-d--ii1qf/

Raza, S. (2023). Challenges in Building Finetuned LLM Models: Quality of Data. LinkedIn. Retrieved from https://www.linkedin.com/pulse/challenges-building-finetuned-llm-models-quality-data-raza-ph-d-/

Pass, R., Seeman, L., & Shelat, A. (2016, September 13). Analysis of the Blockchain Protocol in Asynchronous Networks. Retrieved from https://eprint.iacr.org/2016/454

Pixegami. (2023). Langchain RAG Tutorial. Retrieved from https://github.com/pixegami/langchain-rag-tutorial

The White House. (2022/10). AI Bill of Rights. Retrieved from https://www.whitehouse.gov/ostp/ai-bill-of-rights/#safe

Federal Highway Administration. (2011). Context Sensitive Rock Slope Design Solutions. Retrieved from https://www.fhwa.dot.gov/clas/ctip/context_sensitive_rock_slope_design/default.aspx#toc

Boneh, D., Grundfest, J. (2019). Blockchain and Cryptocurrency Course: What You Need to Know. Retrieved from https://www.youtube.com/playlist?list=PLoROMvodv4rN_bvJCjfM33sOLTGj8gxrF