PROCEEDINGS OF SCIENCE

Approach to Physics Education Using Local AI with RAG for Open Educational Materials Generation

Dmitriy Beznosko [a,*], **Tatiana Krivosheev** [a] **and Alexander Iakovlev** [b]

a College of STEM, School of Sciences, Clayton State University, 2000 Clayton State Blvd, Morrow, GA 30260 USA,

b Upper School Science Department, Woodward Academy, 1662 Rugby Avenue, College Park, GA 30337

E-mail: dmitriybeznosko@clayton.edu

The new tool that starts to enter all parts of our life is AI – and it enters education as well. There are two large concerns with using AI for education - the safety of students' data and the AI missing specific knowledge about the given class. The approach of using the Retrieval Augmented Generation provides the user data to the locally run LLM model (using Ollama framework, a free and open-source tool that allows you to run large language models locally on your system) as a context for the generation of the OER materials. As this AI model of user's choice is run fully locally, no data is transmitted and it can be used to do tasks with students' data, such as summarization of evaluations and simple grading and image recognition tasks. The background, installation highlights and use examples are presented.

*39th International Cosmic Ray Conference (*ICRC2025*)*
15–24 July 2025
Geneva, Switzerland

ICRC 2025
The Astroparticle Physics Conference
Geneva July 15-24, 2025

*Speaker

ISSN 1824-8039 . Published by SISSA Medialab.

1. Introduction

The new tool that starts to enter all parts of our life is AI – and it enters education as well. There are two large concerns with using AI for education: the safety of students' data and AI missing specific knowledge about the given class thus giving hallucinated answers.

The approach of using the Retrieval Augmented Generation (RAG) provides the user data to the locally run LLM model (using Ollama [1] framework, a free and open-source tool that allows you to run large language models locally on your system) as a context for the generation of the OER materials.

As this AI model of user's choice is run fully locally, no data is transmitted and it can be used to do tasks with students' data, such as summarization of evaluations, simple grading and creation of quiz materials using RAG.

2. Running AI Locally

The easiest approach is to install Ollama framework (https://ollama.com/download) or a similar framework onto user's computer. Assuming Ollama, the next step is to pick a model that suits the needs of the user and the existing hardware. This can be done at https://ollama.com/library?sort=newest. Do not install into the virtualized environment!

Decide the model size to get based on video card memory available (typically around 8Gb on the modern machines) or RAM on a device like a laptop. A larger number of parameters in the model (i.e. 3b, 8b) will have better performance but will require also better hardware as well, mostly amount of memory as shown in Figure 1. If the chosen model doesn't fit into video memory completely with some space to space for the operating system itself – it will run slow!

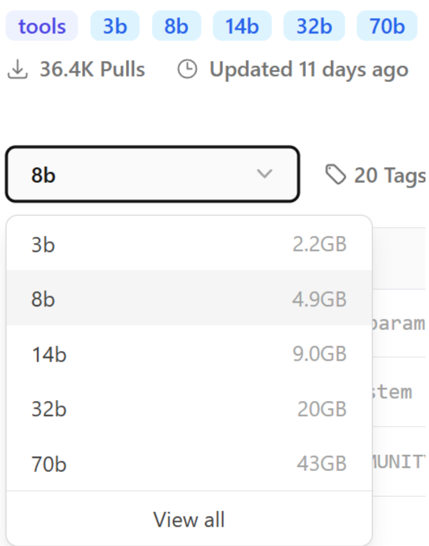


Figure 1. Number of parameters in a model and the needed amount of memory.

If running on a laptop, additional steps may need to be taken to allow Ollama to run on the discrete video card. The framework will not run on the built-in cards and will use CPU instead.

Download the chosen model using the following command line: ollama pull model_name (for example: ollama pull llama3.2:3b). After the download is complete (Figure 2), use the command: ollama run model_name. A command 'ollama ls' will list all local models that were downloaded, command 'rm' can be used to remove the models you don't need any longer as they take a lot of space.

```
C:\Users\[redacted]>ollama pull phi4-mini
pulling manifest
pulling 3c168af1dea0... 100%                                        2.5 GB
pulling 813f53fdc6e5... 100%                                        655 B
pulling fa8235e5b48f... 100%                                        1.1 KB
pulling 8c2539a423c4... 100%                                        411 B
verifying sha256 digest
writing manifest
success
```

Figure 2. A completed download of the model.

1.1 Running a Model

In Figure 2, a very small model was downloaded, it takes only 2.5Gb in video memory. As such, it is very basic and can answer most questions in a most minimalistic fashion. After running the model (ollama run phi4-mini), here's the answer to the question "What are Newton's laws?".

`Newton's Laws of Motion consist of three fundamental principles that describe the relationship between a body and the forces acting upon it, as well as its motion in response to those forces. These can be summarized into:`

`1. ** First Law (Law of Inertia) **: An object at rest will remain at rest unless acted on by an external force; similarly, an object moving uniformly straight line continues doing so indefinitely if no net external force acts upon it.`

`2. ** Second Law **: The acceleration of an object is directly proportional to the net force acting upon it and inversely proportional to its mass (F = ma). This means that a heavier object's motion changes less when you apply more force compared with lighter ones.`

`3. ** Third Law (Action-Reaction) **: For every action, there's an equal and opposite reaction. If object A exerts a force on object B, then simultaneously object B is exerting the same amount of force back onto object A but in the opposite direction.`

`Newton's Laws are foundational to classical mechanics; they describe how objects move under various conditions (like friction or air resistance) when forces act upon them, assuming that all involved masses have non-relativistic velocities.`

The response is very basic from a small model. The question is whether it can be expanded.

3. Using RAG

1.2 RAG Explained

RAG stands for Retrieval Augmented Generation. It is a way to make your basic model more efficient in a specific area by providing it with additional materials without retraining the actual model by providing it with information from user's materials on the topic.

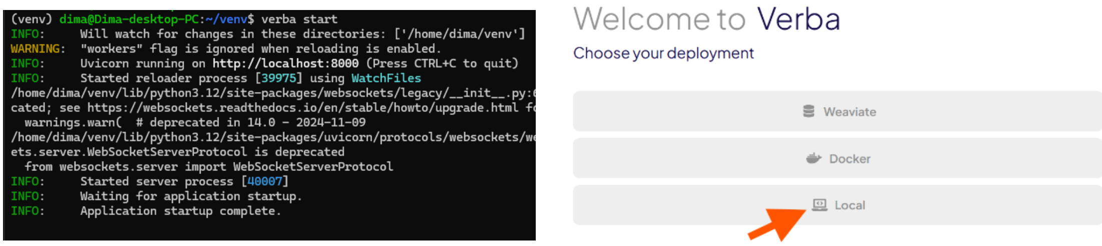
Provided materials are vectorized into chunks and saved locally, with each inquiry the relevant chunks are provided to a model as context from which answers are drawn. This way, you effectively search for the answer within provided documents but using an AI model.

Two programs are widely used for RAG and are reviewed here briefly: 'the Golden RAGtreiver' Verba[1] (available from https://github.com/weaviate/) and the AnythingLLM [2] framework (https://anythingllm.com/desktop).

1.3 Using Verba

Verba is a Python-based RAG environment with an interface via web browser. From the experience of this study, Verba performs more efficiently over other options but is rather difficult to install and has a steep learning curve. Due to limitations of the vectorization library *weaviate* (https://github.com/weaviate/weaviate), Verba can only be installed on a Linux-based OS, or using the WLS2 (Windows Linux subsystem), or in a Docker container, but not directly on Windows. This introduces a possible issue as all these steps require certain topic familiarity from the user. To use WLS2, go to Microsoft© Store and install Ubuntu or another Linux distribution.

After installation, running Verba is simple: verba start. The application will report the starting process as in Figure 3a. The web interface will typically be available at http://localhost:8000/ and the first screen is shown in Figure 3b.



(a) (b)

Figure 3. (a) Verba starting up; (b) Loading web interface.

Quick note on settings – in the folder, from which a user starts Verba, there should be a file named '.env' that contains the basic settings for Verba. For example, the minimal set will be:

OLLAMA_URL=http://172.28.224.1:11434

OLLAMA_MODEL=llama3.2:3b

The first line tells Verba where the Ollama server is running (use command: ollama serve), and the second specifies which model should be used. Note that the WLS2 may see the IP address of user's PC different from its localhost or external address. To check for that, a traceroute command should be used, as shown in Figure 4.

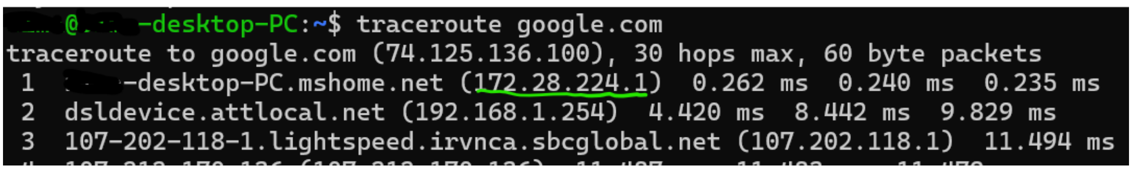


Figure 4. Running a traceroute command in WLS2 command line.

[1] Installation guide is at https://medium.com/@ria.banerjee005/build-your-rag-with-verba-ollama-and-talk-to-your-own-data-187f61b72408

Once Verba is up and running and can communicate with Ollama server and the downloaded model, it's time to add user's materials. At this point, it's simple: step 1 - go to 'Import data' and choose files in pdf or docx formats that are currently supported as step 2. Choose chunker in step 3 – for different choices, different parameters will appear that need to be set, or defaults can be used as a starting point. After step 4 - importing data is complete, the user can go to Chat and ask questions. This process is outlined in Figure 5.

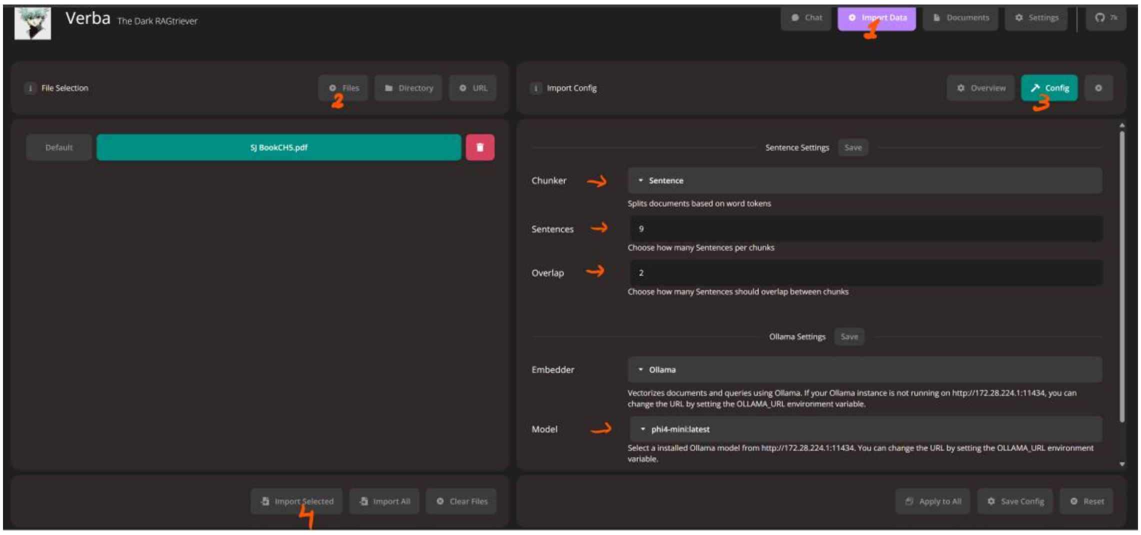


Figure 4. Importing data. Follow steps 1, 2, 3 and 4. Note that different chunkers have different options.

1.4 RAG and Questions

To show the difference between the response of the mini model used and the effect of the RAG, the same question was asked - "What are Newton's laws?". There is only one file imported – it is an old physics textbook chapter on Newton's laws.

Here's the chat output, also indicating the name of the file from which the chunks were used:

- SJ BookCH5.pdf 6 chunks
- Newton's three fundamental principles for understanding forces between two bodies at a time (law 1), as well as their consequences—acceleration due to force causing an increase or decrease in velocity per law 2; the third principle involves conservation of momentum, explaining why objects remain still unless acted upon by another object according to Newton's first and second laws. These three principles govern everything from how a ball flies through air resistance on its way toward you (a specific example) to what happens when galaxies collide billions of years ago in space.
- Note: This explanation covers the essence, but not all possible nuances related to newtonian mechanics, such as rotational motion or complex systems which might require additional context for complete understanding. For a deeper dive into Newton's laws and their applications across different scenarios including those mentioned like block acceleration on an incline (as shown with blocks in your provided texts), it would be beneficial to consult comprehensive resources that

can elaborate further based on the specific questions you have about these principles or any related calculations.

- For more detailed explanations, especially regarding applied examples such as ones involving forces between objects at a distance including gravitational interactions like those mentioned for mobiles and mobile systems (e.g., Figure P5.74), additional reading from textbooks specifically aimed towards physics education can offer deeper insights with step-by-step problem-solving methods which align closely to the principles laid out by Sir Isaac Newton in his works, such as "Philosophiæ Naturalis Principia Mathematica" (Mathematical Principles of Natural Philosophy).

The model response is not much more detailed and to the point. Of course, be careful what you ask for, as now the model makes a distinction between Newton's laws themselves and a description about them. The laws were listed following this prompt: "Describe all three Newton's laws", with the lengthy output that is skipped here.

1.5 RAG and Making a Quiz

After a user feeds RAG with existing materials on the topic, a prompt can be used, like the following: "Write 5 problems with solutions on the topic of Newton's laws, for the college-level quiz with biological content. Format output as LaTeX." Caveats:

- Various models understand the term college level' differently so you may need to be more specific, such as calculus-based or algebra-based questions
- Different models have varying success with LaTeX so a bit of clean-up may be needed
- Larger models do better
- Carefully read the resulting output as there could be mistakes, big and small

More details on the generation of the quizzes and other class materials can be found in [3].

1.6 AnythingLLM Framework

A few words need to be said about another option for using LLM on the personal computer – AnythingLLM framework. It is a much simpler program to use overall - one installer for everything needed, will also install Ollama during the setup process. In the program window, click on AnythingLLM logo to select and install models automatically, some suggested models with descriptions are listed in the window that opens, and update often.

The overall advantage of using this framework is the ease of use, no need for the user to install Linux subsystems, to search for models, etc. If needed, the framework can use pre-existing Ollama instance or external generative LLM. It can work with images using model llama3.2-vision:11b or similar, note that this takes rather long to process.

Downsides also exist. In chat settings, user can also add documents or do that in the chat via prompt – this also adds the attached file to pool of documents, but can cause confusion and double additions, thus adding a need to remove duplicates manually. RAG option exists as the 'agent' in the chat window, and it is not fully reliable during use, needs to be selected with every prompt.

PoS(ICRC2025)1238

4. Conclusion

Large Language Models (so-called AI) can comfortably run on most modern PCs; the main requirements are the amount of RAM and the presence of an advanced video card with large amount of video memory. These models are still small and don't hold much information, thus RAG provides enhancements in AI responses by tuning to topics as needed by providing specific materials on the topic in question.

AI can help educators with quizzes and study guides, in summarizing students' responses; it can recognize images and handwriting and provide invaluable assistance with grading and is useful in the labs as well [3]. All this is possible because the smaller and distilled LLM models can secure and privately run on a local PC, eliminating the worry about the privacy of the data and materials.

Acknowledgement

Work is sponsored by ALG (Affordable Learning Georgia) round 26 grant RG017.

References

[1] Ollama Inc, 435 Hamilton Ave, Palo Alto, CA 94301, US

[1] Mintplex Labs Inc, 49 Stouts Ln Irvine, CA, 92606, USA

[2] Dmitriy Beznosko, Tatiana Krivosheev, Alexander Iakovlev. *Using Local AI with RAG for Open Educational Materials Generation for Physics Education.* In proceedings of 2025 European Physical Society Conference on High Energy Physics, `PoS(EPS-HEP2025)618`, 2025

[3] Dmitriy Beznosko, Tatiana Krivosheev, Alexander Iakovlev. *Ultimate makeover for an introductory physics lab*. In proceedings of 38th International Cosmic Ray Conference **444** 1597, 2023/7/25, `PoS(ICRC2023)1597`, DOI: https://doi.org/10.22323/1.444.1597