AI for Equal, On Demand Education

Angelo Botticelli
Elizabethtown College
Scranton, USA
ajbotcs@gmail.com

Abstract

A large language model answers based on what it has been taught so if that knowledge bank could be fed directly from a professor seeking to teach their notes to students, a model ought to be able to serve as a online tutor available whenever neither a professor nor human tutor would be. Additionally, such a model should be able to learn from what questions are most asked to defer that information to the professor such that they can provide clarification in lessons when students are too shy to outwardly ask for help.

CCS Concepts

• **Applied computing** → **E-learning**; *Interactive learning environments*; **Computer-assisted instruction**; **Computer-managed instruction**; *Document searching*; • **Computing methodologies** → *Information extraction.*

Keywords

LLM, MathGPT, Tutoring

ACM Reference Format:
Angelo Botticelli. 2018. AI for Equal, On Demand Education. In *Proceedings of (Written).* ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXX X.XXXXXXX

1 Introduction

After two years of tutoring for lower level CS courses and seeing the advent of LLM capturing the attention of students I was left uncertain. Models like ChatGPT and its ilk were fully capable of answering much of the coding questions that students would come into my office hours asking for; but despite that, many still did attend. The element of intractability was irreplaceable.

I should not have been too surprised though, as many questions before could easily be found online from sources like Chegg before LLM became commonplace. Beyond LLM services generally being free compared to paid online homework Q-A websites, they are also closer in human interactivity as the LLM mimics a person.

After some ruminating and a suggestion from a friend in a Machine Intelligence club, the idea for an AI tutor began. LLM are already capable of this on their own but where I seek to innovate is in the content that is taught. Rather than just using info scrapped from the web, this tutor would be fed directly from the notes of the professor. It would be a software directly for the class being tutored and as such, a professor could have complete control on what is being taught.

As a tutor, students seeking out online options like homework Q-A websites for answers is considered the worst-case as they subvert the learning component of completing homework. The goal of this project is to design a better worst-case such that students are still easily able to seek out their answers but through an AI tutor which is trained to mimic a real tutor in leading students to their answers and encouraging the learning process . This would not seek to replace real tutors nor to I expect it would, as a real tutor provides more interactivity than any existing options for off-hours help and that would remain unchanged even with this work. Rather, this would just be a professor-controlled supplement to help learning that students could use rather than reaching out for third-party help which may not teach the same contents nor even teach, instead just giving information with no learning process.

For goals beyond the minimum viable product, I also seek to allow a professor to see which topics students most ask about such that they can tune their lessons to address those items. Often students will not speak up in class and instead let their questions ruminate out of embarrassment so this can help bridge that information to a professor in a more anonymous way.

2 Background

To supplement my project, I have searched for related studies in this field as well as into the foundations of the topics I will be analyzing as LLM are, to many, a black box of unknown function. First, I read a paper by Woo-Hyun Kim and Jong-Hwan Kim which leverages neural networks to design individualized AI tutors. They leverage a trained neural network model, the ART network to make a multi-pronged approach in tracking the status of the individual learner as well as their focus areas such that the model can add new nodes into specified areas to further hone the learner's skills. Compared to my own goals, this work seeks a much deeper depth where my own seeks more breadth, only individualized as much as the model has learnt in the current conversation with the user with its learning being more general to the notes the professor has elected to teach the model.

Second, I analyzed a paper by Tommaso Calo and Christopher Maclellan which leverages generative AI to design tutoring interfaces. They integrate generative AI and drag-and-drop no-code solutions to allow educators to create tutoring interfaces for their students. Compared to my own work, this is much closer in usage set-up than the prior as it focuses on an educator and their actions alongside AI to benefit students; however, this paper focuses on the interface only. The tutor in this case is from ATB (apprentice tutor builder) from an earlier work by one of the authors. In tandem, these papers propose a no code method to design AI based tutor

Written, Sept 9, 2024, Elizabethtown, PA

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

plans. However, in my project I seek to leverage LLM as they provide an element of interaction closer to a real human tutor which just AI based lessons cannot provide.

Third, I sought to look into the foundations of MathGPT as I forsaw anything involving mathematics to be a big issue as most LLM are poor with such processes. To this, I do not have a scholarly article, but a collection of pages to cite as MathGPT is not a single model but a name that many companies and individuals have independently made each with their own quirks as the name MathGPT is seemingly not reserved to one model. In searching for a MathGPT, I have found several different options: MathGPT by Photostudy, an online tutoring service which claims its model can teach a pre-algebra textbook with high accuracy. There is also MathGPT by Mathpresso-Team Qanda which ranked first in benchmarks. Then there is MathGPT Mathos which is a free available online webpage similar to Wolfram Alpha. Lastly, I also found MathGPT, a pre-trained GPT model hosted directly on ChatGPT's webpage by Fei Hou. Which of these, if any, I will leverage will come after which I am able to best pull the API for as well as performances compared to one another.

3 Designs

The first step of this procedure is choosing a model to operate with. There are many publicly available options for free or very cheap to integrate into my program.

	Average	Multi-choice Qs	Reasoning	Python coding	Future Capabilities	Grade school math	Math Problems
Claude 3 Opus	84.83%	86.80%	95.40%	84.90%	86.80%	95.00%	60.10%
Gemini 1.5 Pro	80.08%	81.90%	92.50%	71.90%	84%	91.70%	58.50%
Gemini Ultra	79.52%	83.70%	87.80%	74.40%	83.60%	94.40%	53.20%
GPT-4	79.45%	86.40%	95.30%	67%	83.10%	92%	52.90%
Claude 3 Sonnet	76.55%	79.00%	89.00%	73.00%	82.90%	92.30%	43.10%
Claude 3 Haiku	73.08%	75.20%	85.90%	75.90%	73.70%	88.90%	38.90%
Gemini Pro	68.28%	71.80%	84.70%	67.70%	75%	77.90%	32.60%
Palm 2-L	65.82%	78.40%	86.80%	37.60%	77.70%	80%	34.40%
GPT-3.5	65.46%	70%	85.50%	48.10%	66.60%	57.10%	34.1%
Mixtral 8×7B	59.79%	70.60%	84.40%	40.20%	60.76%	74.40%	28.40%

Figure 1: Sept 8 2024 LLM Benchmarks. Table by Anita Kirkovska. via Google Images Commons. (https://cdn.prod.website-files.com/63f416b32254e8679cd8af88/65e8ac9e390f8072d8696352_llm-leaderboard.png).

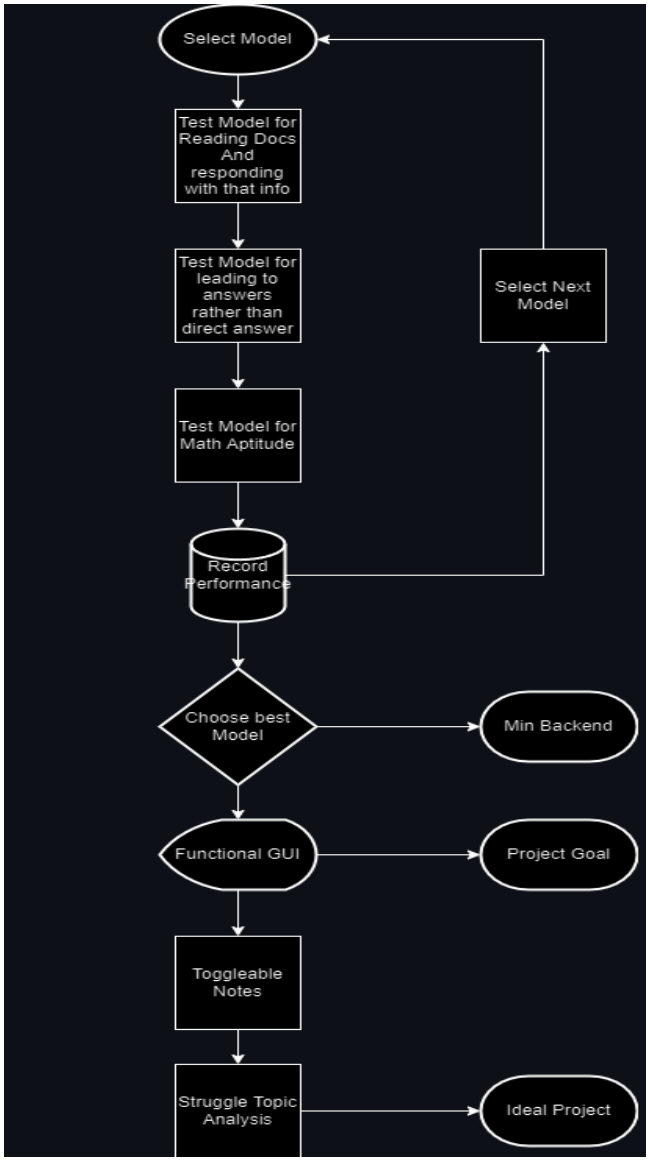

Figure 2: Sept 9 2024. Flowchart by Angelo Botticelli. via Draw.io Commons. (https://github.com/ajbot22/DS400AITutor/blob/main/DS400Workflow.drawio.png).

Amongst these are many viable models; however, one important area discussed prior that I want to ensure is compatible is mathematics. To that end, I want to choose a model which has work completed in integrating CAS or other mathematical solving engines to at least be able to help solve up to algebra. Because of this desire, I will be defaulting first to a GPT model as they tend to have the most community support; however, I will also be testing with others.

My workflow shown above is the general approach I plan to take. The first half of my work will focus on the testing of models for the bare functionality required for this project which will then be followed by general usability in creating a user interface and integrating as many features as possible in the time period allotted for this work. Metrics for what the best model is will be discussed shortly in the Experiments section.

Once more work on the back end has been finished, I will append here its progress as well as the front end when the GUI has been drafted out further than the concept phase.

4 Experiments

As of now, my plan for an experimentation metric is human oriented. There are several areas that can be bench-marked for each

model. Their general performance with a typical benchmark dataset. Their specific performance in a specified field with a benchmark dataset such as focusing on a mathematics dataset to test for math proficiency. Testing through use by trusted people to record its anecdotal performance.

To me, it makes little sense for the batch approach for this as this will not be creating its own Large Language Model but pulling existing ones and modifying them; and with this, such models are certainly bench-marked else wise on general performance. Subject specific may be valuable, especially in the case of mathematics where I expect models to struggle.

However, anecdotal performance is also important as the mission of this project is for the models to be taught by a professor and to focus on a narrow subject range so even a more specific benchmark is likely to go beyond the scope of notes provided for a single course; though, the model's ability to recognize topics outside of its own scope is important. So too, is the model's ability to infer what topics outside of its scope it should answer from an assumed knowledge by the questioner and which it should reserve as topics it expects to be covered at a later date by the professor.

The end result is likely to be a combination of all of these options. Broad testing to see how well the model understands its intended answering range. Subject specific testing to judge performance and more acutely how well it can differentiate topics it ought to not discuss. And hands on testing with specific notes to verify at least a moderate accuracy and usability.

5 Timeline and Contributions

This is a solo project which I have a semester to accomplish to the fullest extent possible; however, I give credits to my advisor and professor Dr. Li as well as my friend who first proposed this idea to me as a project Steven Klinefelter. For our midway milestone, I seek to have accomplished some model comparison testing as well as the first draft of the GUI For our second milestone, I seek to have a functional GUI with a usable model but potentially with some missing or not fully implemented features For the final milestone, I seek to have all projected features integrated into a user-friendly GUI; however, at minimum, I want a functional GUI with the base tutoring functionality.

sample-base

Received 9 September 2024