# Generating Personalized Study Plan with Smaller Language Model

Christoph Winter
*Courant Institute*
*New York University*
NY, USA
jcw9376@nyu.edu

Ying Zeng
*Courant Institute*
*New York University*
NY, USA
yz11720@nyu.edu

## I. Project Overview

Personalized Learning has emerged as a critical field of educational study. It aims to tailor learning experiences to meet learners' goals, preferences, and prior knowledge. In contrast to the one-size-fits-all model of traditional education, personalized study plans enhance learning efficiency by adapting different learning paces.

Traditional approaches to generating personalized learning paths rely on rule-based cognitive diagnosis and require delicate annotations [1]. With the few-shot learning capability of Large Language models, personalized study plans can be generated more easily with minimal annotations. Furthermore, the logical reasoning ability of Large Language models enables them to provide adaptive explanations for each recommendation step in the study plan, making it more understandable for teenage students.

## II. Project Objectives and Contributions

Recent research has proved that Large Language Models (LLMs) have the ability to produce more explainable and understandable learning paths in an interactive way, enhancing learner performance and engagement [2]. However, few studies have highlighted the difficulty students with Internet of Things (IoT) edge devices face in accessing Large Language Models. To promote educational equity, our project explores to introduce mathematics-focused, smaller language models (SLMs) that can be deployed on smartphones while retaining most of the capacity to identify students' primary learning challenges and generate personalized study plans.

## III. Literature Review

Existing systems take a variety of approaches in an attempt to solve this problem. We have found a number of papers analyzing this problem but few that implement solutions similar to our proposal. One of the papers [3] breaks down the evolution of education in relation to LLMs and the issues that are becoming increasingly present in modern educational environments and how LLMs are able to solve them. This provides more background and context for the issue we are attempting to solve.

The current usage of LLMs in personalized learning path generation focuses on prompt engineering. For example, Chee et al. [4] designed prompts for initial assessment, clarification, and explanation to incorporate learner-specific information and generate personalized and coherent learning paths. Abu-Rasheed et al. [5] utilize a knowledge graph as a source of contextual information and involve domain experts in the loop to reduce hallucinations. Kuo et al. [6] harness the power of LLMs by introducing a cognitive model, an item model, and interactive procedures, which increase the accuracy of locating individuals' learning weaknesses by providing fewer but more precise questions. This approach offers a scaffolding method during remediation, enhancing the adaptive mechanism of the TALP platform.

In contributing to smaller language models, a few approaches have been introduced. One previous approach to a math-specific fine-tuned LLM [7], is most closely related to our approach in terms of implementation. The fine-tuning was done on a number of math-related datasets, in order to improve the model's arithmetic ability. The results showed that the model was able to specialize in this area, and outperformed other models with significantly larger parameter sizes. Another approach [8] to a similar problem makes use of much larger LLMs to generate learning paths for students with regard to a specific subject and subtopic. The prompts were carefully engineered for this task, and the models used were able to output coherent learning paths for a variety of students and topics.

An interesting method that we plan to include is model quantization [9]. This technique utilizes quantization and LoRA to significantly improve the efficiency of fine-tuning with regard to memory usage. This will decrease the amount of dedicated hardware needed for fine-tuning and inference, allowing us to train and deploy in more constrained environments.

## IV. Methodology

### A. Finetune SLMs with Math Dataset

To generate a personalized study plan, the model needs to demonstrate the ability in: a) problem-solving to assess whether students can perform correctly on certain topics; b) math reasoning to evaluate whether students accurately deduce and arrive at solutions; c) concept-relating to identify students' biggest difficulties based on their provided error sets.

To achieve these abilities, two datasets will be used to fine-tune smaller language models:

- For problem-solving and reasoning skills, the MATH Dataset [10], which contains 12,500 challenging competition mathematics problems, will be used. A joint task

focusing on producing the correct output and generating step-by-step solutions will be created.
- For concept-relating skill, the Chasat-Algebra-Sub02 Dataset [11], which contains the relation between questions and corresponding concepts, will be utilized.

### B. Prompt Engineering and Smartphone Deployment

After the best-performing SLM model is fine-tuned, we will create a prototype where students can obtain a personalized math study plan by inputting specific error sets. The existing prompt engineering [4] for personalized study plan generation will be reused to integrate the student input and the fine-tuned model. The output will identify the biggest difficulties the student is facing, recommend relevant concepts to learn with specific explanations and provide a sequenced study plan for the next steps.

### C. Idea Validation

The fine-tuned SLM model will be tested on the aforementioned datasets. For the Math Dataset, accuracy will be the primary metric. For the Chasat-Algebra-Sub02 Dataset, the ROUGE-N metric will be used. The baseline models will be GPT-4, the unfine-tuned LLaMA model and GPT-4 o1.

## V. Technical Considerations

We plan to implement our approach as a fine-tune atop Llama 3.1, a series of general-application LLMs developed by Meta. More specifically, we will start with the 8 billion parameter model to measure its ability to complete the tasks we have planned. If we determine that this model size is unable to capture the knowledge we require, we may choose to use a model with a greater set of parameters. The next model size in the Llama series contains 70 billion parameters, which may present difficulties during our fine-tuning process due to its size. If another model size is necessary, we will most likely need to evaluate the abilities of other available models with parameter sizes between 8 and 70 billion, of which we have the capability to fine-tune efficiently.

## VI. Timeline and Milestones

- Offline Training Framework for Fine-tuning: Develop and establish the offline training framework to facilitate the fine-tuning process of the smaller language models. Deadline: 11.12.
- Refinement and Hyperparameter Tuning: Refine the model by tuning hyperparameters and addressing issues related to the expressive power of the smaller models. Deadline: 11.26.
- Prototype Development and Smartphone Deployment: Create the prototype of the application and deploy it on smartphones. Deadline: 12.06.

## VII. Expected Outcomes and Evaluation

The product of our work will be a prototype based on the fine-tuned smaller language model. We plan to evaluate this model on a few metrics in comparison with other available models.

We will compare the accuracy and ROUGE-N metrics based on two math-related datasets. We will evaluate our model on its ability to answer the questions correctly, by comparing its final answer to the correct value from the dataset.

We also plan to evaluate the model qualitatively by generating a set of outputs from validation prompts. These outputs, along with outputs from other available models on the same prompts, will be shuffled into a dataset. We will evaluate these outputs blindly ourselves, by creating a system that presents us with two outputs for a specific prompt, where we will choose the one which more effectively satisfies the query. We also may replicate this system in a way where we replace ourselves with GPT-4 to determine which output is more effective, which would allow us to evaluate a much larger dataset.

We plan to compare our model's results to generalized models of varying sizes. We will use Llama 3.1 8B, GPT-4 and GPT-4 o1. We will compare the results of the math exam dataset based on the amount of correctly answered questions and the results from our output selection strategy based on the proportion of selection results.

We find that this approach will provide a comprehensive view of the performance of our model in relation to other available LLMs.

## VIII. Potential Challenges and Mitigation Strategies

There are a multitude of challenges that we may face during the implementation of our ideas. The first of which is related to the availability of computing power to fine-tune our model. The NYU HPC cluster seems to be quite busy running computing jobs from many students and researchers and our concern is that available computing on the cluster will not be abundant enough to complete our research in the time period necessary. We have limited time to complete our research and some amount of necessary testing to fulfill the requirements, so there is some proportion of availability that we require from the cluster to meet the deadline and complete the project successfully. If the cluster is too busy for us to use to fine-tune our model effectively, we may need to utilize external computing solutions such as AWS EC2.

Another challenge we may face is the possibility that 8 billion parameter base model we plan to use is insufficiently large to complete the tasks we propose. We are attempting fine-tuning with a relatively large dataset and we are not sure of the model "capacity" to learn this information. A related issue is that the model can learn all of the new information we use for fine-tuning, but some of its necessary, more general abilities are lost in the process. Both of these concerns would lead to a model that does not meet the requirements we have set. In this scenario, we may need to evaluate which other generalized pre-trained models are available, with larger parameter sizes, that we can use in its place. This substitute model would need to be openly available, larger than our original 8 billion parameter model such that it can learn a larger amount of information, yet also small enough that we are still able to fine-tune it effectively, and in a timely manner. If we cannot find

a model that is suitable, we may need to utilize commercial LLM APIs to fulfill our needs.

## IX. Conclusion

Educating effectively is a difficult task, which becomes nearly impossible to achieve without the necessary resources. In the age of LLMs, there exists the ability to begin solving this problem on a global scale. We believe that our research can improve the accessibility of educational materials in a personalized and interactive way for students who do not have access to these resources elsewhere, so that any student with the motivation to learn can be given the opportunity to pursue their goals.

## References


[1] B.-C. Kuo, F. T. Y. Chang, and Z.-E. Bai, "Leveraging LLMs for Adaptive Testing and Learning in Taiwan Adaptive Learning Platform (TALP)," 2023.

[2] B. Jiang, X. Li, S. Yang, Y. Kong, W. Cheng, C. Hao, and Q. Lin, "Data-driven personalized learning path planning based on cognitive diagnostic assessments in MOOCs," *Applied Sciences*, vol. 12, no. 8, p. 3982, 2022.

[3] H. Xu, W. Gan, Z. Qi, J. Wu, and P. S. Yu, "Large Language Models for Education: A Survey", arXiv [cs.CL]. 2024.

[4] C. Ng and Y. Fung, 'Educational Personalized Learning Path Planning with Large Language Models', arXiv [cs.CL]. 2024.

[5] H. Abu-Rasheed, C. Weber, and M. Fathi, "Knowledge Graphs as Context Sources for LLM-Based Explanations of Learning Recommendations," arXiv preprint arXiv:2403.03008, 2024.

[6] B.-C. Kuo, F. T. Y. Chang, and Z.-E. Bai, "Leveraging LLMs for Adaptive Testing and Learning in Taiwan Adaptive Learning Platform (TALP)," 2023.

[7] T. Liu and B. K. H. Low, "Goat: Fine-tuned LLaMA Outperforms GPT-4 on Arithmetic Tasks", arXiv preprint arXiv:2305. 14201, 2023.

[8] C. Ng and Y. Fung, "Educational Personalized Learning Path Planning with Large Language Models", arXiv [cs.CL]. 2024.

[9] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, 'QLoRA: Efficient Finetuning of Quantized LLMs', arXiv preprint arXiv:2305. 14314, 2023.

[10] D. Hendrycks et al., 'Measuring Mathematical Problem Solving With the MATH Dataset', NeurIPS, 2021.

[11] Manas Bansal, "Chasat-Algebra-Sub02 Dataset," Hugging Face, 2022. [Online]. Available: https://huggingface.co/datasets/themanas021/chasat-algebra-sub02. [Accessed: 25-Oct-2024].