An Integrated Framework for Applying LLMs with Custom Metrics: a Use Case for Personalized Study Plan Generation

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

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

***Abstract*—Personalized learning with Large Language Models (LLMs) has become a focal point of recent educational research, as the logical reasoning ability of LLMs offer the potential to enhance learning efficiency by adapting to various learning needs. However, previous works rely largely on closed-source metrics and internal datasets, which are neither broadly accessible nor reusable for future studies. To address this problem, this paper proposes an integrated framework for building education-oriented LLM assistants through a pipeline that incorporates finetuning, prompt engineering, and evaluation customization. We validate the effectiveness of our framework through a comprehensive use-case for personalized study plan generation. By leveraging the broad abilities of LLMs, this framework can be extended to produce effective responses for tasks in domains that lack public evaluation metrics.**

I. Introduction

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 styles and needs.

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 (LLMs), personalized study plans can be generated more easily with minimal annotations. Furthermore, the logical reasoning ability of LLMs enables them to provide adaptive explanations for each recommendation step in the study plan, making it more understandable for teenage students.

However, the application of using LLMs to generate personalized study plans faces challenges due to limited datasets and close-source evaluation metrics. Existed studies [1]–[3] validate their approaches either using internal dataset obtained from educational platforms or relying on platform-specific metrics, such as retention rates and student satisfaction rates collected within campuses. Furthermore, these resources are neither broadly accessible nor reusable as common benchmarks. This dependency limits the reproducibility of research, forcing researchers to collect new data and define new evaluation in subsequent studies.

To solve this problem, we propose an integrated framework to assist researchers in building their domain-specific finetuned models and customizing evaluation metrics for tasks lacking common benchmarks, such as assessing the quality of personalized study plan. The introduced metrics customization pipeline leverages large language models as domain experts, enabling researchers to define widely applicable and reusable rubrics for domains without public evaluation metric in a cost-effective manner.

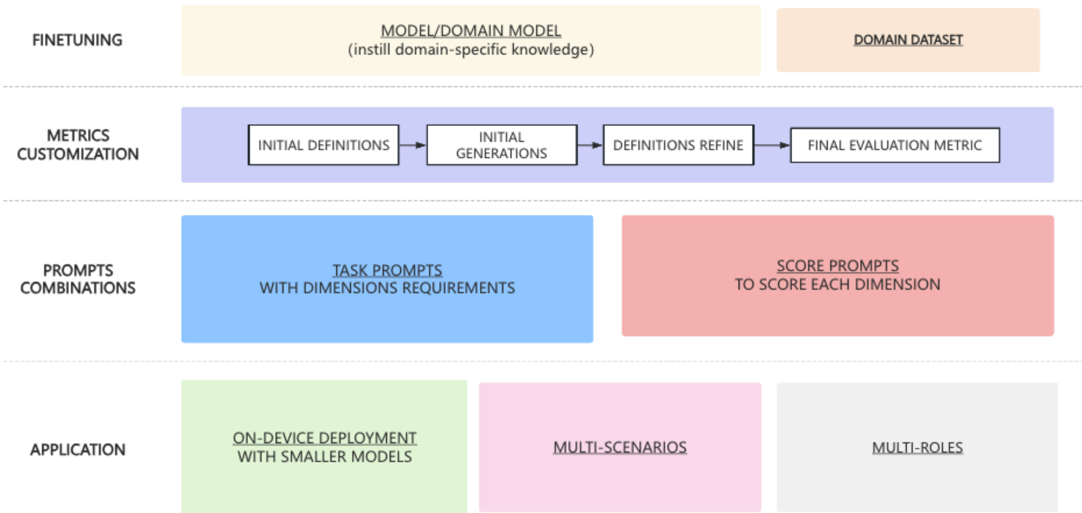
We validate the effectiveness of our framework through a comprehensive use-case focused on personalized study plan generation, demonstrating the successful development of evaluation metrics that can be broadly applied to assess personalized learning plans.

II. Related Work

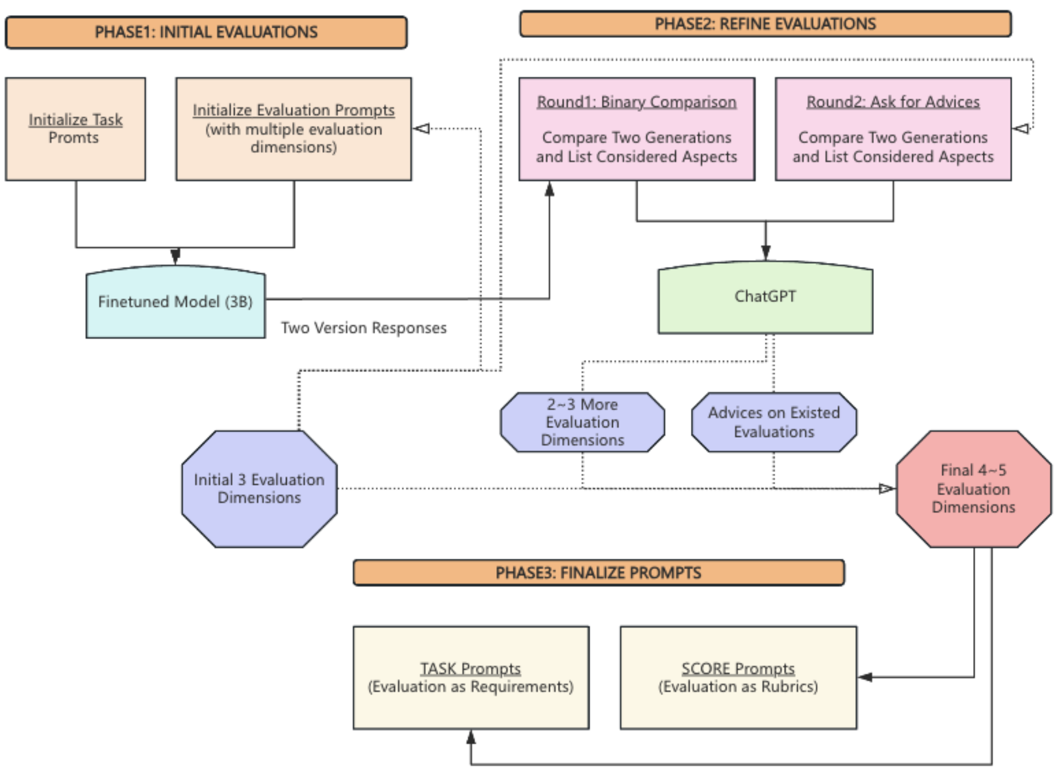
In this section, we review related work in the domain of Learning Path Planning. Xu et al. [4] breaks down the evolution of education in relation to LLMs and the issues that are becoming increasingly present in modern educational environments and how LLMs are able to solve them. This provides more background and context for the issue we are attempting to solve.

The current usage of LLMs in the generation of personalized learning pathways focuses on prompt engineering. For example, Chee et al. [5] designed prompts for initial assessment, clarification, and explanation to incorporate learner-specific information and generate personalized and coherent learning paths. Abu-Rasheed et al. [6] utilize a knowledge graph as a source of contextual information and involve domain experts in the loop to reduce hallucinations. Kuo et al. [1] 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, some approaches have been introduced. Park et al. [6] makes use of much larger LLMs to generate learning paths for students with respect to a specific subject and subtopic. The prompts were carefully engineered for this task and the models used were able to produce coherent learning paths for a variety of students and topics. Another approach [7], defines a system for tutoring that dynamically generates a pre-prompt based on



(a) Framework Architecture.



(b) Metrics Customization Pipeline.

Figure 1: An Integrated Framework for Applying LLMs with Custom Metrics.

the answers a student provides to a set of questions to more effectively personalize the tutoring conversation that follows.

III. Methodology

A. Framework Architecture

Our proposed approach for customizing evaluation metrics for domain-specific tasks is an integrated framework that combines finetuning, metrics customization, prompts combination, scoring and application, as illustrated in Figure1(a). Each component of the framework serves the following functions:

1) Finetuning Section: This stage aims to assist users in efficiently building their fine-tuned models. By providing domain-specific datasets, a smaller large language model (Llama 3.2 3B) is trained using the Unsloth framework [8].

2) Metrics Customization: In this stage, the evaluation generation pipeline guides users through the process of creating customized evaluation metrics for their domain-specific tasks, step by step. The pipeline integrates both human-designed evaluation rubrics and the insights provided by a large language model, such as ChatGPT. After selectively merging information from ChatGPT and human feedback, a well-customized set of evaluation rubrics is finalized by the end of this section.

3) Prompt Combination: With the finalized rubric from the previous stage, users can now complete their task prompts using this rubric, leading the fine-tuned model to generate more accurate responses for the current task. Additionally, users can define a scoring task for the generated responses based on this rubric, enabling the LLM to function as a scorer and labeler for their domain-specific dataset. This stage provides a feasible and cost-effective approach for constructing domain-specific datasets, especially for less popular domains.

4) Application: With the finalized rubric and a smaller fine-tuned model, users can seamlessly implement on-device deployment. Moreover, the rubric can be reused for similar tasks, enabling the framework to support multiple scenarios.

B. Pipeline for Generating Evaluation Metrics

As shown in Figure1(b), at this pipeline, ChatGPT [9] serves as an expert, offering additional evaluation aspects and advices on refining the initial rubrics. The whole procedure would be:

1) Initialize Evaluation Rubrics: Users begin by defining the task prompt and initial evaluation prompts based on several key aspects for evaluating the quality of LLM responses to the task. Responses are then generated using either the task prompt alone or a combination of task and evaluation prompts.

2) Refine Evaluation Rubrics: Using the two responses from the previous stage, ChatGPT performs a binary comparison and justifies its choice based on its own evaluation criteria. This process enables users to obtain supplementary evaluations for their tasks. Additionally, ChatGPT assesses the initial evaluation dimensions to identify any element that may require modification.

3) Finalize Evaluation Rubrics: By combining the additional aspects and suggestions on the initial evaluations, users can finalize their evaluation rubrics with 4 to 5 dimensions.

IV. Experiments

To validate the effectiveness of the framework described earlier, we conducted experiments focused on generating personalized learning paths. The experiments consist of two sections. The first section involves finetuning a smaller large language model on a math dataset to instill the model with mathematical knowledge, thereby improving its performance when generating study plans for math. The second section entails customizing evaluation rubrics to evaluate the quality of the generated study plans.

A. Model Finetuning

1) Environment Setup:

- System: NVidia RTX 4090 GPU (24 GB VRAM)
- Base Model: Llama-3.2-3B-Instruct
- Finetune Dataset: openai/Gsm8k
- Evalution Metrics: Accuracy
- Training Framework: Unsloth and Huggingface

2) Implementation: We fine-tuned the Llama-3.2-3B-Instruct model using the Unsloth library. The model was trained on GSM8K dataset for 60 steps, with a batch size of 8. To optimize memory usage and accelerating the training process, we incorporated the Low-Rank Adaptation (LoRA) technique [10], which facilitates efficient fine-tuning by reducing the number of trainable parameters.

TABLE 1: Performance Comparison of Different Models

Model	Accuracy (%)
Finetuned Model	63.5
Llama-3B-Instruct	73.5
Llama-1B-Instruct	29.5

Due to limitations in computational resources, the performance of our fine-tuned model does not outperform the base model, as presented in TABLE1. However, the finetuning process is fully supported, allowing users to easily substitute their own domain-specific datasets and initiate their training.

B. Prompt Engineering

1) Environment Setup:

- System: NVidia RTX 4090 GPU (24 GB VRAM)
- Base Model: Llama-3.2-3B-Instruct
- OpenAI API: GPT-4o (gpt-4o-2024-08-06)

2) Intialize Evaluation Prompt: In this step, we defined the task prompt to generate a personalized study plan and initialize the evaluation prompt. For the initial evaluation prompt, being detailed and hierarchical were considered, as shown in Figure 2(a).

In Figure 2(b), the generated responses using either the task prompt alone or a combination of task and evaluation prompts are presented. The left half shows the study plan generated with the task prompt alone, while the right half displays the one generated with the combined prompts. As can be seen, the response generated with the initial evaluation prompts is more well-organized and detailed, aligning with the requirements outlined in the rubrics, whereas the response

```python
task_system_prompt = (
    "You are an expert in math education. "
    + "You should provide clear, descriptive, and helpful answers and explanations to the student's questions."
    + "The student may also ask or provide additional specific information that you should take into account when assisting them."
    "\n\n"
    + "You are helping a student with questions about the direction of their studies. "
    + "You will help the student generate a study plan for their course. "
)

system_prompt = (
    "\n\n"
    + "If there are specific requirements or constraints, you should satisfy the requests in your response"
)


evaluation_system_prompt = (
    "\n\n"
    + "Consider the following dimensions when generating the study plan: "
    + "1. Be Detailed: Provide a detailed plan that includes the topics and concepts they should focus on, in a clear and organized manner. "
    + "2. Be heirarchical: Create a heirarchical study plan, grouping similar topics and concepts together. "
)

raw_system_prompt = task_system_prompt + system_prompt + evaluation_system_prompt
partial_system_prompt = task_system_prompt + system_prompt
```

(a) Task prompt and initial evaluation prompt.

```
I'd be happy to help you generate a study plan for Calculus 1!

To create a personalized study plan, I'll need to know a bit more about your learning style, schedule
, and goals. Please answer the following questions:

1. What is your current level of math knowledge (e.g., algebra, geometry, etc.)?
2. How many hours per week can you dedicate to studying Calculus 1?
3. What are your goals for the course (e.g., do you want to earn a good grade, prepare for a future m
ath course, etc.)?
4. Do you have any specific areas of Calculus 1 that you're struggling with or would like to focus on
 (e.g., limits, derivatives, integrals)?
5. Do you prefer a structured, schedule-based approach or a more flexible, self-directed approach?

Once I have this information, I can help you generate a tailored study plan that suits your needs and
 learning style.

Also, I'll provide you with a general outline of a study plan for Calculus 1, which you can modify as
 needed.

**General Study Plan Outline:**

**Weeks 1-2:**

* Review and refresh your math foundation (algebra, geometry, etc.)
* Introduce yourself to the basics of Calculus 1 (limits, functions, etc.)
* Complete any assigned readings or homework

**Weeks 3-6:**

* Focus on limits, derivatives, and basic applications
* Practice problems and exercises to build confidence and fluency
* Review and reinforce key concepts

**Weeks 7-10:**

* Introduce yourself to integrals, sequences, and series
* Practice problems and exercises to build confidence and fluency
* Review and reinforce key concepts
```

```
I'd be happy to help you generate a study plan for Calculus 1.

**Calculus 1 Study Plan**

To create an effective study plan, we'll break down the course into manageable chunks. Here's
 a suggested outline, grouping similar topics and concepts together:

**I. Introduction to Calculus (Weeks 1-2)**

* Topics:
        + Limits
        + Basic concepts of calculus (derivative, integral, etc.)
        + Graphing and visualizing functions
* Study Tips:
        + Review basic algebra and trigonometry
        + Practice graphing and visualizing functions
        + Focus on understanding the concept of limits and its significance in calculus

**II. Differentiation (Weeks 3-6)**

* Topics:
        + Rules of differentiation (power rule, product rule, quotient rule, etc.)
        + Geometric interpretation of derivatives
        + Differentiation of trigonometric and exponential functions
* Study Tips:
        + Practice applying the rules of differentiation
        + Focus on understanding the geometric interpretation of derivatives
        + Review and practice differentiating various types of functions

**III. Applications of Derivatives (Weeks 7-8)**

* Topics:
        + Optimization problems (maxima and minima)
        + Related rates and related functions
        + Applications of derivatives in physics and engineering
* Study Tips:
        + Practice solving optimization problems
        + Focus on understanding related rates and related functions
        + Review and practice applying derivatives in physics and engineering contexts
```

(b) Task prompt alone response versus Combined prompts response

Figure 2: Initial evaluation prompt and initial study plans

generated with the task prompt alone tends to be more general.

3) Refine Evaluation Rubrics: To compare the generated study plans from the previous step, we designed a prompt, as shown in Figure 3(a), to enable ChatGPT to assess these plans and provide 2 to 3 evaluation aspects. As presented in Figure 3(b), the response indicates that the first two metrics provided by ChatGPT align with our predefined evaluation dimensions, while also introducing the additional evaluation aspect of ”Being Customizable” to our criteria.

Meanwhile, we developed a prompt (Figure 4(a)) to solicit feedback on the initial evaluation prompt, aiming to identify any elements that may need modification. The feedback, shown in Figure 4(b) and 4(c), confirmed that the original metrics ”Being detailed” and ”Being hierarchical” should be retained, while ”Being actionable” and ”Being adaptable” were recommended to be included.

4) Finalized Evaluation Rubrics: By selectively merging suggestions from ChatGPT, we finalized the evaluation prompts, as shown in Figure 5(a), with the key considerations: ”Being Detailed and Actionable”, ”Being Hierarchical” and ”Being Adaptable”. The final generated study plan, presented in Figure 5(b), is more structured and actionable compared to the initial output in Figure 2(b), as it clearly divides the plan into short-term, mid-term, and long-term goals, enhancing its adaptability.

C. Scoring Using Evaluation Rubrics

In addition to enhancing the quality of generated study plans, the finalized evaluation metrics can also be leveraged to construct a scoring task for domain-specific tasks. We explored this application by defining a scoring task based on the finalized evaluation dimensions. In this case, a zero-shot prompt was employed, simplifying the scoring task into three levels: -1, 0, and 1, which correspond to poor, neutral, and good, respectively, as shown in Figure 6(a). As shown in Figure 6(b), the scoring prompt effectively guided ChatGPT to assess scores for each evaluation metric as well as assess the overall quality of the study plan.

```
compare_prompt = (
    "The following texts are two different sample study plans to be used by students to further their studies."
    + "Compare the following study plans and provide feedback on which one is more effective. "
    + "Provide an explanation of your decision and define the 2-3 most important metrics you used to arrive at your conclusion."
)
```

(a) Prompt for comparison

```
### **Conclusion: Study Plan 2 is More Effective**

**Explanation:**
Study Plan 2 is more effective because it scores higher on **clarity and structure** and **specificity and detail**. It provides a clear, sequential breakdown of topics with explicit time allocations, actionable study tips, and smaller, manageable chunks of content. This level of detail helps students understand what to study, how to approach the material, and how it connects to real-world applications.

While Study Plan 1 focuses on customizability, its lack of depth and detailed structure makes it less effective for most students who need guidance to navigate complex subjects like Calculus 1.

---

### **Key Metrics Used**

1. **Clarity and Structure**: A well-organized plan ensures logical flow, manageable sections, and ease of understanding.
2. **Specificity and Detail**: Explicit breakdown of topics, actionable study tips, and examples of applications make the plan actionable and practical.
3. **Customizability**: Flexibility to adapt the plan to individual learning needs or constraints.

Study Plan 2 excels in the first two metrics while maintaining sufficient flexibility, making it the better option overall.
```

(b) Comparison output

Figure 3: Compare initial study plans

```
advice_prompt = (
    "For this task we considered the following dimensions during the generation process: "
    + evaluation_system_prompt
    + "\n\n"
    + "Please evaluate the dimensions we use generation criteria, and suggest if any dimensions should be added or removed or modified."
)
```

(a) Prompt for seeking advice

```
### **Current Dimensions Evaluation**

1. **Be Detailed** (Keep with Modification):
   - **Strength**: Emphasizing detail ensures students receive a clear and actionable plan with sufficient depth.

   - **Suggested Modification**: Add the requirement of **actionable guidance**, such as study tips, examples, or practical applications. Detail alone is helpful, but actionable elements make the plan more usable.
   - **Revised Dimension**: **Be Detailed and Actionable**: The study plan should include clear, specific topics, with actionable study tips, techniques, or examples to help the student approach the material effectively.

2. **Be Hierarchical** (Keep):
   - **Strength**: Grouping similar topics into sections provides logical flow and clarity, making the plan easier to follow and reducing cognitive overload.
   - **No Change Needed**: This dimension is well-suited for organizing concepts in an effective and hierarchical manner.
```

(b) Assessment on initial evaluation metrics

```
### **Revised Dimensions**
1. **Be Detailed and Actionable**: Provide a detailed study plan with specific topics, study tips, and actionable strategies.
2. **Be Hierarchical**: Organize topics logically, grouping similar concepts into clear and structured sections.
3. **Be Adaptable**: Allow flexibility for students to personalize the plan based on their progress, time constraints, or focus areas.

---

These three dimensions collectively ensure the study plan is **clear**, **organized**, **practical**, and **flexible**, making it more effective for a wide range of students.
```

(c) Revised evaluation metrics

Figure 4: Refine initial evaluation metrics

V. Conclusion

Through our research, we have determined an effective process for refining prompts to satisfy user-defined tasks. The results highlight our framework's ability to self-improve and its capability to generate structured, detailed and coherent outputs. We hope that this framework will assist users and researchers across a wide range of educational contexts, enabling the use of novel user-defined tasks.

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

```
evaluation_system_prompt = (
    "\n\n"
    + "Consider the following dimensions when generating the study plan: \n"
    + "1. Be Detailed and Actionable: Provide a detailed plan with specific topics and concepts they should focus on, and include actionable strategies.\n"
    + "2. Be Hierarchical: Create a hierarchical study plan, grouping similar topics and concepts together.\n"
    + "3. Be Adaptable: Allow flexibility for students to personalize the plan based on their progress, time constraints, or focus areas.\n"
)
```

(a) Final evaluation prompts

```
I'd be happy to help you create a study plan for Algebra 1!

**Short-term goals (next 2-3 weeks)**

1. **Understand the basics of linear equations**:
        * Review the definition of a linear equation and its graph.
        * Focus on the slope-intercept form (y = mx + b) and the standard form (Ax + By = C).
        * Practice identifying the slope and y-intercept from a graph.
2. **Practice identifying the equation of a line from a graph**:
        * Use online resources or worksheets to practice identifying the equation of a line from a graph.
        * Start with simple graphs and gradually move to more complex ones.
        * Focus on identifying the slope, y-intercept, and any other relevant information.
3. **Develop problem-solving strategies**:
        * Practice solving problems that involve finding the equation of a line from a graph.
        * Use a systematic approach to solve problems, such as:
                + Identify the slope and y-intercept from the graph.
                + Use the slope-intercept form to write the equation.
                + Check your work by plugging in values from the graph.

**Mid-term goals (next 4-6 weeks)**

1. **Expand your knowledge of linear equations**:
        * Review and expand your knowledge of linear equations, including:
                + Slope-intercept form (y = mx + b)
                + Standard form (Ax + By = C)
                + Point-slope form (y - y1 = m(x - x1))
        * Practice writing equations in different forms.
2. **Practice solving linear equations**:
        * Practice solving linear equations using different methods, such as:
                + Graphing
                + Slope-intercept form
                + Standard form
        * Use online resources or worksheets to practice solving linear equations.
3. **Develop problem-solving strategies for word problems**:
        * Practice solving word problems that involve linear equations.
        * Use a systematic approach to solve problems, such as:
                + Read and understand the problem.
                + Identify the relevant information.
                + Use the slope-intercept form to write the equation.
```

(b) Study plan generated with final evaluation rubrics

Figure 5: Final evaluation metrics and study plan generated with these metrics

```
score_prompt = (
    "You are a labeller and response evaluator. Please evaluate the following question and response pair. "
    + "Provide a set of scores based on the quality of the response and how well it fulfills each of the criteria, "
    + "with -1 indicating that the response does not meet the requirement, 0 being neutral, and 1 meaning the response effectively meets the required criterion. "
    + "Here are the criteria to consider when evaluating the response, please provide a score for each critereon based on how well the response meets the requirement, "
    + "and a brief explanation as to why each score was chosen: \n"
    + "1. Be Detailed and Actionable: Provide a detailed plan with specific topics and concepts they should focus on, and include actionable strategies.\n"
    + "2. Be Hierarchical: Create a hierarchical study plan, grouping similar topics and concepts together.\n"
    + "3. Be Adaptable: Allow flexibility for students to personalize the plan based on their progress, time constraints, or focus areas.\n"
)
```

(a) Scoring prompts

```
**Criteria Scores:**

1. **Be Detailed and Actionable**: 1
The response provides a clear and detailed plan with specific topics and concepts to focus on. It includes actionable strategie
s, such as practicing identifying the slope and y-intercept from a graph, and using a systematic approach to solve problems.

2. **Be Hierarchical**: 1
The response creates a hierarchical study plan, grouping similar topics and concepts together. It starts with short-term goals
and then moves on to mid-term goals, providing a clear structure for the student to follow.

3. **Be Adaptable**: 1
The response allows flexibility for students to personalize the plan based on their progress, time constraints, or focus areas.
 It provides a general framework that can be adapted to individual needs.

**Overall Score**: 1 (Excellent)

**Explanation**: The response effectively meets all the required criteria. It provides a detailed and actionable plan, creates
a hierarchical structure, and allows for adaptability. The student is provided with specific topics and concepts to focus on, a
nd actionable strategies to help them achieve their goals. The response is well-organized and easy to follow, making it an exce
llent study plan.
```

(b) Evaluation outputs

Figure 6: Scoring prompt and evaluation outputs for each evaluation dimension

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

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

[6] Park, M., Kim, S., Lee, S., Kwon, S. & Kim, K. Empowering Personalized Learning through a Conversation-based Tutoring System with Student Modeling. *Extended Abstracts Of The CHI Conference On Human Factors In Computing Systems*. pp. 1-10 (2024,5), http://dx.doi.org/10.1145/3613905.3651122

[7] 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.

[8] Daniel Han, Michael Han and Unsloth team, ”Unsloth”, http://github.com/unslothai/unsloth, 2023.

[9] OpenAI. ChatGPT. https://chat.openai.com/chat. 2023.

[10] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora:Low-rank adaptation of large language models. International Conference on Learning Representations. 2022.