Methodology:
The methodology begins with initializing the system and verifying the use of the CUDA device to ensure GPU acceleration for enhanced computational efficiency. The LLM model and its associated configuration files, such as config.json, model.safetensors, and tokenizer.json, are downloaded and loaded into memory, with checkpoint shards successfully integrated for compatibility. Performance evaluation is conducted by administering a predefined set of questions to the model, calculating key metrics like average score, accuracy, and speed. The data is further analyzed to identify strong topics and areas requiring improvement, with detailed insights provided for incorrect answers, including the question, the student's response, and the correct answer. Performance metrics and analysis results are visualized and saved as a PNG file for reference.

Subsequently, AI-driven recommendations are generated based on the performance summary and mistake analysis, offering targeted suggestions such as identifying key misconceptions, topics for review, study strategies, and practice methods. An interactive learning assistant is also incorporated, enabling dynamic interaction to address student queries and guide study plans based on performance data. The process concludes with a comprehensive report, summarizing performance, mistake analysis, recommendations, and suggested study topics for improvement, ensuring a structured and effective approach to learning enhancement.

How to use the models and code:
To use the provided code for the personalized recommendation system, follow these steps:

1. The code is available on Kaggle and can be accessed through the following link: https://www.kaggle.com/code/sanjay20052/personalized-recommendation-system.

2. First, visit the Hugging Face website (https://huggingface.co/) and create a free account. Once your account is set up, request access to the LLaMA 3.2 3B model through Hugging Face. Approval typically takes some time.

3. After your access request is approved, go to your Hugging Face account settings. Navigate to the "Access Tokens" section and create a read token. Copy the generated token.

4. Open the Kaggle code notebook. Locate Code Block 2, Line 38, and paste the token into the appropriate variable.

5. Ensure the runtime accelerator is set to P100 GPU on Kaggle for optimal performance. Execute the notebook cells sequentially to run the entire process.

6. The chatbot functionality is integrated into the system to provide real-time guidance and recommendations. Engage with the chatbot for a more interactive learning experience.

For reference, the code has already been executed, and the results are provided as an IPYNB file. The notebook includes detailed outputs, performance summaries, and visualizations such as graphs to help you understand the results.

By following these steps, you can successfully run the code, explore the chatbot, and review the detailed analysis and recommendations provided by the system.