Ben Carpenter and James Stevens

Professor Hao Loi

Intro to Software Engineering CSC 212-90

21 February 2025

Generative AI and Learning: Using Retrieval-Augmented Generation (RAG) for C++ Tutoring

The objective of our project is the design and implementation of an AI-based tutor app that addresses introductory-level C++ programming questions. By using a Retrieval-Augmented Generation (RAG) approach to providing targeted feedback, our goal is to develop a basic, but effective AI framework that facilitates learning. Instead of relying on a pre-stored database of answers, the system will dynamically generate context-specific responses by interfacing with a local Llama3 AI model through Ollama’s local REST API.

The backend, implemented in C++, will serve as the engine of the application. It will process incoming questions from users, communicate with the Llama3 model via the REST API, and ensure that the answers generated are both concise and relevant. This design minimizes the risk of oversaturated or generic responses, providing students with clear and direct explanations tailored to their queries.

We chose to use Retrieval-Augmented Generation because it solves many of the outstanding issues that exist with generative AI. One of generative AI’s largest and most well know flaws is that of halucinazation, where the AI just makes ups stuff it thinks sounds good, but is not at all factual. By implement RAG we intend to make it so that this app has a stricter data pool that it can draw from making it so that the AI doesn’t to generate stuff that has no basis in reality, thus solving a key problem with using AI as a tutor.