RamBot: An Early-Stage Chatbot for Academic Advising Tailored for West Chester University

Jarrett Trapani
West Chester University
JT958079@wcupa.edu

ABSTRACT

RamBot is an innovative chatbot currently in early development, utilizing Hugging Face libraries to incorporate text-to-text large language models. Designed to provide West Chester University students with 24/7 instant response academic advising, it aims to include a retrieval augmented generation system to provide relevant information regarding West Chester University. Currently in its infancy, RamBot is undergoing testing on various graphics cards and models via CloudLabs (12 Hardware, n.d). The vision of RamBot is to create a web application similar to the functionality of ChatGPT, which is a space to have a 1 on 1 conversation with RamBot. RamBot is a type of generative AI (Chambers et al., n.d.) that specifically uses Large Language Models (LLMs) such as OpenHermes 2.5 Mistral 7B to create new, never seen data. The OpenHermes model was fine-tuned using 1,000,000 examples of GPT-4 generated data, allowing it to be efficient at text generation and question answering, which is the functionality needed to achieve RamBots goal. At first these models have to go through pre-training; this is where the model is exposed to a mass amount of datasets to learn the patterns and nuances of language and contextual meaning. This is a fundamental stage for all Large Language Model's where they learn to excel in various applications, including academic advising and providing guidance to students about class selection and course planning. After the pre-training stage Large Language Models go through fine tuning. During this stage developers utilize fine-tuning which is a process aimed at optimizing the performance of the model. Specifically, Parameter-Efficient Fine-Tuning (PEFT) (Chambers et al., n.d.) is a discovered method of Fine-Tuning that allowed developers to enhance the model's computational efficiency without compromising the model's performance quality. Another method used to enhance the performance of the model is Reinforcement Learning with Human Feedback (RLHF) this is the process where humans can refine the responses of the Large Language Model by personally interacting with the model and providing feedback which could be corrections or ratings on the output quality. Future work includes incorporating a Retrieval-Augmented-Generation (RAG) (Chambers et al., n.d.) system which will be a major enhancement for RamBot. This new system will allow RamBot to fetch for valuable information from external sources. This is pivotal because this will allow RamBot to have knowledge on specific information regarding West Chester University. Providing students with reliable and relevant responses that are seeking help. The end goal would be to provide students with a fully functional web application aiming to mimic the success of ChatGPT's ability to provide personal 1 on 1 interactions. The poster will highlight the potential to revolutionize academic advising for students at West Chester University, it will also include information about the development of Large Language Models and specifically the development of RamBot. The poster will also provide insight on personal research regarding the performance variances observed based on the different hardware provided by cloudlabs and different choice of Large Language Models provided by Hugging Face. In general, RamBot was developed using personal and public research providing me with the knowledge to plan out and begin development of a chatbot specifically crafted for students at West Chester University with the functionality to provide academic advising. Generative AI is a tool that can be used in unlimited ways, and this is just one example of how beneficial these tools can be. Shining light on Artificial Intelligence will provide individuals to be better prepared for the AI driven future.

Bibliography

12 Hardware. (n.d.). Docs.cloudlab.us. https://docs.cloudlab.us/hardware.html

Chambers, M., Eigenbrode, S., Barth, A., & Fregly, C. (n.d.). Generative AI with Large Language Models. Coursera. https://www.coursera.org/learn/generative-ai-with-llms/home/info