Development of Microcontroller-Based AI Robot Tour Guide Utilizing Custom Language Models

Ian S. Jackson
West Virginia University
Morgantown, United States
isj0001@mix.wvu.edu

Adien Ballard
West Virginia University
Morgantown, United States
agb00033@mix.wvu.edu

***Abstract*—**

I. Introduction

Advancements in artificial intelligence (AI) and natural language processing (NLP) have enabled new forms of human-computer interaction, particularly in the realm of educational engagement. Universities increasingly seek innovative ways to connect with prospective students, providing them with immersive experiences that showcase academic programs, research opportunities, and campus life. One such approach is the integration of AI-powered robotic tour guides that allow visitors to interact dynamically with departmental resources.

This research focuses on the development of an AI-driven robotic tour guide designed to enhance the experience of prospective students visiting the Lane Department of Computer Science and Electrical Engineering (LCSEE). The goal is to create an interactive medium where students can ask questions about the department, and the robot, powered by language models, generates informative responses in real-time. Unlike static presentations or pre-scripted responses, this approach enables natural, context-aware interactions, providing a more engaging and personalized tour experience.

To achieve this, two distinct methodologies were explored: (1) DirectLLM, which fine-tunes a large language model (LLM) specifically for LCSEE-related queries, and (2) a Classify-Retrieve-Generate (CRG) pipeline, which first classifies the user's question, retrieves relevant information, and then generates a response. A custom Q&A dataset was curated using LCSEE-specific data, ensuring that the AI model delivers accurate and contextually relevant information.

Due to the constraints of running an AI model on embedded hardware, the system was deployed on a Raspberry Pi 4, paired with a MangDang Mini Pupper [1] robot as the physical embodiment of the tour guide. The robot is equipped with a simple microphone and speaker interface, facilitating natural language communication with users. The choice of hardware necessitated the use of a lightweight AI model optimized for fast inference, ensuring real-time conversational interactions without significant latency.

This paper details the design, development, and deployment of the AI-powered robotic tour guide, evaluating the performance of both DirectLLM and the CRG-based approach. By comparing these methods, we aim to determine the most effective strategy for delivering responsive and informative AI-driven interactions in an embedded robotics context.

II. Related Work

A. Fine-Tuning Large Language Models for Domain-Specific Tasks

Fine-tuning pre-trained LLMs has become a standard approach to adapting general-purpose models for domain-specific applications. Studies show that fine-tuned LLMs outperform generic models in specialized domains, such as healthcare and legal services [2], by leveraging tailored datasets. The DirectLLM approach in this work applies the same principle, fine-tuning a model specifically on Lane Department of CSEE (LCSEE) information. However, fine-tuning requires significant computational resources, which poses challenges for deployment on edge devices like microcontrollers and single-board computers [3].

B. Deploying AI Models on Resource-Constrained Devices

Deploying AI models on low-power devices such as Raspberry Pi and microcontrollers presents unique challenges, primarily due to hardware limitations in memory and computation. Prior studies have explored optimization techniques such as quantization, pruning, and distillation to reduce model size while preserving accuracy [4]. Recent advancements in small language models (SLMs) offer promising alternatives for efficient, real-time inference [5]. This work integrates such optimizations to deploy a lightweight yet effective AI system on a Raspberry Pi 4.

III. Background

IV. Our Approach

A. DirectLLM

B. Classify-Retrieve-Generate

V. Methodology

VI. Results

VII. Conclusion

Future Work:

Computational resources were provided by the WVU Research Computing Thorny Flat HPC cluster, partly funded by NSF OAC-1726534.

References

[1] Mangdang, ”Mangdang Store,” Available: https://mangdang.store

[2] M. Rajkomar et al., ”Scalable and accurate deep learning with electronic health records,” npj Digital Medicine, vol. 1, no. 18, 2018.

[3] T. Dettmers et al., ”Sparse fine-tuning for efficient deployment of large-scale language models,” NeurIPS Conference Proceedings, 2022.

[4] Z. Jiang et al., ”Efficient deep learning inference on edge devices: Challenges and techniques,” ACM Computing Surveys, vol. 55, no. 6, 2023.

[5] H. Touvron et al., ”Llama 2: Open foundation and fine-tuned chat models,” arXiv:2307.09288, 2023.

APPENDIX