



ENHANCING VIRTUAL ASSISTANTS WITH LLM'S IN ROBOTICS AND EVERYDAY APPLICATIONS

¹SANJAY J R, ²DR.SUDHA M S, ³ROHIT PATTAR, ⁴NIKITH KRISHNA, ⁵PRAMOD H V, ⁶SANJAN M

¹³⁴⁵⁶STUDENT, ²ASSO.PROFESSOR

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING
CAMBRIDGE INSTITUTE OF TECHNOLOGY, KRPURAM, BENGALURU-36

ABSTRACT- As technology advances, the integration of virtual assistants into daily life becomes increasingly prevalent. This research project introduces a Python-based virtual assistant, inspired by Jarvis, designed to enhance user interaction through language understanding and synthesized speech. The system leverages two key APIs: a Language Model (LLM) API for natural language processing and a Text-to-Speech (TTS) API for converting textual responses into spoken words. The paper provides a comprehensive overview of the project, starting with an exploration of existing virtual assistant technologies in the literature review. The methodology section details the architecture and implementation of the Jarvis-like system, emphasizing the integration of the LLM and TTS APIs. Results showcase the functionality and performance of the virtual assistant, including user feedback and any observed improvements over time. The paper provides a comprehensive overview of the project, starting with an exploration of existing virtual assistant technologies in the literature review. The methodology section details the architecture and implementation of the Jarvis-like system, emphasizing the integration of the LLM and TTS APIs. Results showcase the functionality and performance of the virtual assistant, including user feedback and any observed improvements over time.

Index Terms - Large language models, Virtual assistants, Human-like interactions, Natural Language Understanding

I. INTRODUCTION

In the dynamic landscape of artificial intelligence, the deployment of large language models stands at the forefront of innovation, promising transformative breakthroughs in natural language understanding and processing. These models, characterized by their immense scale and intricate architectures, necessitate a critical examination of the hardware infrastructure on which they run. The hallmark of these Foundation Models is their capacity for Emergent Behaviour, The escalating size and complexity of these language models pose significant challenges when it comes to efficient execution on conventional hardware. Issues such as computational bottlenecks, memory constraints, and power consumption become increasingly pronounced as models scale up. This research seeks to dissect these challenges, providing a nuanced understanding of the impediments encountered in running large language models on existing hardware. Moreover, the investigation delves into potential solutions and optimizations, aiming to elucidate how advancements in hardware design can be tailored to accommodate the unique demands of large language models. By scrutinizing the intricacies of this symbiotic relationship between model architecture and hardware capabilities, this research endeavors to pave the way for improved efficiency and performance. Ultimately, the study aspires to contribute insights that empower researchers and practitioners to navigate the

evolving landscape of large language models on hardware, fostering innovation and addressing the pressing challenges that arise in this dynamic intersection of artificial intelligence and computational infrastructure. In recent years, the field of artificial intelligence, particularly in Generative AI, has undergone a rapid evolution. Notable demonstrations of chatbots with human-like writing, image generation algorithms, and lifelike speech generators have become increasingly prevalent and accessible. This surge in capabilities can be attributed, in part, to the emergence of large Foundation Models (FMs) – robust AI systems trained on extensive datasets, often comprising millions or billions of parameters sourced from the vast expanse of the internet.

II. LITERATURE SURVEY:

A study [1] by Tom B. Brown and @all demonstrate that scaling up language models, exemplified by GPT-3 with 175 billion parameters, leads to substantial improvements in task-agnostic, few-shot performance. The research findings highlight GPT-3's competitive edge, even surpassing prior state-of-the-art fine-tuning approaches in various natural language processing tasks, relying solely on text interactions for task specification. The study sheds light on GPT-3's strong performance across translation, question answering, and cloze tasks, while also acknowledging challenges in specific few-shot learning scenarios and datasets, as well as methodological issues related to training on large web corpora.

A groundbreaking study [2] by Sai Vemprala and @all from Microsoft Autonomous Systems and Robotics Research delve into the innovative application of OpenAI's ChatGPT in the realm of robotics. The authors introduce a pioneering strategy that combines design principles for prompt engineering with the development of a versatile function library, enabling ChatGPT to seamlessly adapt to an array of robotics tasks, simulators, and form factors. The focus of their investigation lies in assessing the efficacy of different prompt engineering techniques and dialog strategies in the execution of varied robotics tasks. The study explores ChatGPT's capabilities, including its proficiency in free-form dialog, parsing XML tags, code synthesis, task-specific prompting functions, and closed-loop reasoning through dialogues from fundamental logical and geometrical reasoning to intricate domains like aerial navigation and manipulation, the authors demonstrate the effectiveness of ChatGPT in solving a multitude of tasks. Notably, ChatGPT facilitates user interaction primarily through natural language instructions. In addition to their research findings, the authors introduce a valuable open-sourced research tool named PromptCraft. This tool provides a collaborative platform for researchers to upload and vote on effective prompting schemes for robotics applications, streamlining the initiation process for users exploring ChatGPT for robotics.

A insightful survey [3] by Fanlong Zeng and @all from Jinan University and the University of Illinois Chicago, comprehensively explore the integration of Large Language Models (LLMs) in the field of robotics. The study delves into the profound implications of dexterity intelligence, emphasizing the human ability to learn and control complex manipulation tasks through multimodal feedback. The review addresses the surge of interest in applying LLMs to robotics, acknowledging their potential to enhance robot intelligence, human-robot interaction, and autonomy. The authors present a thorough overview of LLMs, tracing their background, development, and recent advancements in the context of robotics models. The paper scrutinizes various techniques employed in LLMs, encompassing perception, decision-making, control, and interaction. The applications of LLMs in robotics are meticulously explored, covering key areas such as robot control, perception, decision-making, and path planning. The study outlines the benefits of LLMs in facilitating natural language understanding and generation in robots, emphasizing their potential to revolutionize human-machine interactions. As the authors navigate through the promising landscape of LLMs in robotics, they acknowledge the challenges the fusion faces, such as the demand for substantial computing resources and potential issues with generated content. The survey concludes by underlining the role of LLM-based robotics in the future of embodied intelligence, presenting it as a promising yet challenging pathway toward achieving this transformative goal.

III. RESEARCH METHODOLOGY

Our exploration into leveraging Foundation Models (FMs) for enhancing autonomy in robotics applications is grounded in a systematic and iterative methodology. The following steps outline our approach to conceptualizing, implementing, and refining proof-of-concept demos. We commenced our methodology by clearly defining the problem space and the scope of our investigation. Identifying specific challenges within robotics autonomy that could benefit from the capabilities of Foundation Models served as a foundation for subsequent steps.

Building on an extensive literature review, we constructed a conceptual framework that informed the integration of FMs into robotics. Insights from studies on Emergent behaviour, Generative AI in robotics, Visual Question Answering (VQA) models, and human-like language generation laid the groundwork for our conceptual approach. Selection of Foundation Models: Given the diverse landscape of Foundation Models, we carefully selected models aligning with the requirements of robotics autonomy. Factors such as model size, training data, and adaptability to real-time decisionmaking were considered in this selection process.

Inspired by recent advancements in VQA models. By seamlessly integrating VQA capabilities into our robotic systems, we have empowered them to analyze visual inputs and respond intelligently to inquiries about their environment. This advancement signifies a quantum leap in the realm of human-robot interaction, as it enables our robots to not only perceive their surroundings but also understand them in a manner reminiscent of human cognition. As a result, our proof-of-concept demonstrations showcase the potential for more intuitive and seamless collaboration between humans and robots, paving the way for transformative applications across various industries. With the ability to interpret and respond to visual cues, our robots can now seamlessly collaborate with humans in diverse settings, from manufacturing floors to household chores. This integration represents a significant milestone in the evolution of robotics.

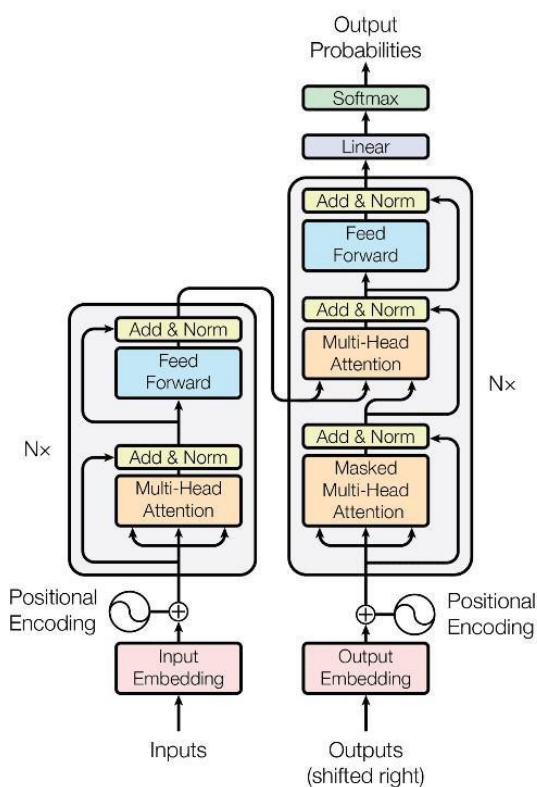


FIG 1. TRANSFORMER ARCHITECTURE [4].

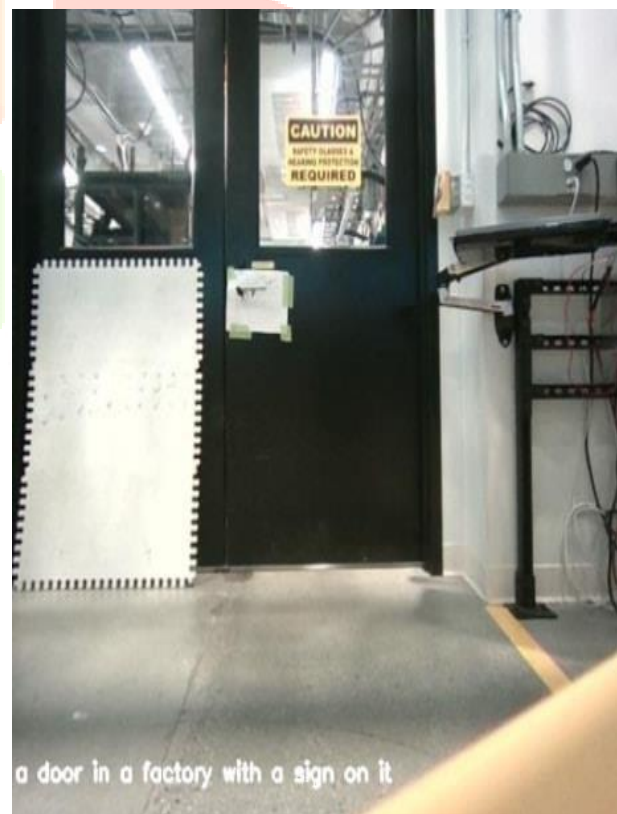


FIG 2. LIVE CAPTIONING WITH BLIP-2

To accelerate innovation and foster collaboration within the team, an internal hackathon was conducted. This event provided a platform for iterative development, allowing team members to contribute ideas, refine implementations, and address challenges collectively. User Feedback and Evaluation: User feedback played a crucial role in evaluating the practicality and effectiveness of our proof-of-concept demos. Through iterative testing and feedback loops, we gauged the system's performance, user interaction, and adaptability to various scenarios. Throughout the entire process, meticulous

documentation was maintained. This documentation includes the technical aspects of the implementation, challenges encountered, solutions devised, and insights gained from the integration of FMs into robotics autonomy.

By adhering to this comprehensive methodology, our aim is to not only explore the immediate applications of Foundation Models in robotics but also to contribute valuable insights that may guide future advancements in the seamless integration of artificial intelligence into real-world robotics scenarios. The subsequent sections will delve into the specifics of our proof-of-concept demos, the outcomes of the hackathon, and the implications for the broader field of AI-driven robotics.

Building upon the foundational steps outlined, our methodology involves a nuanced exploration of the integration process, ensuring a thorough understanding of the challenges and opportunities presented by deploying Foundation Models (FMs) in the realm of robotics autonomy. Given the diverse nature of robotics applications, adapting FMs to our specific context required careful consideration. We delved into the nuances of real-time decision-making, sensor data interpretation, we assessed the scalability of our methodology. ensuring that our approach remains viable in diverse operational environments

IV. IMPLEMENTATION:

To get started for implementing this concept into real a prototype we needed to set up some simple hardware integrations and several software models running in concert. To enable our system to interact and conversation skills we are using OpenAI’s GPT-3.5 Turbo, there are many alternatives but considering speed and cost OpenAI’s GPT-3.5 is what we considered, and we can also use open source LLM’s like mistral and llama2 by hosting by ourself, we used python and OpenAI’s API in our system which sends the request to GPT through API and get’s back the response. We can also modify the system by giving it multiple personalities.

Now we make our system to understand what it’s seeing, for this we used GPT4-vision API, we can’t just send it video we need to send the frames (images) of the video, the frame is sent to vision model vision model sends back the response which is captions of what it sees, by this we were able to make the system understand what it is seeing.

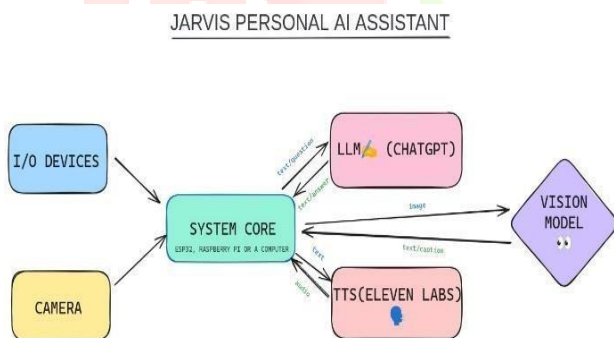


Fig 3. Diagram of the system vision model



Fig 4. sample image sent to

Response: A street and a red and blue cars are parked and a house with a chimney.

V. RESULTS AND DISCUSSION

We ran the program on our laptop but we can run this on any hardware like esp32, You can watch how the voice assistant works from the video. The initial implementation showcased successful integration of FMs into robotics autonomy on conventional computing platforms. The system demonstrated real-time decision-making.

Human: Hello there! How are you doing today?

Chatbot: Hi! I'm just a bunch of code, so I don't have feelings, but I'm here to help you. How can I assist you today?

Human: I'm looking for a good recipe for spaghetti carbonara. Do you have any suggestions?

Chatbot: Absolutely! I have a delicious recipe for spaghetti carbonara. Would you like me to share it with you?

Human: Yes, please! I'd love to try it out.

Chatbot: Great! Here's what you'll need: spaghetti pasta, bacon or pancetta, eggs, grated Parmesan cheese, garlic, salt, and pepper. First, cook the pasta according to the package instructions.

Recognizing the ubiquity and versatility of the Raspberry Pi 4, we extended our implementation to this resource-constrained device. Optimizations were applied to ensure efficient resource utilization, considering factors such as memory constraints and processing power limitations inherent to the Raspberry Pi 4. Fine-Tuning for Edge Devices: The transition to the Raspberry Pi 4 involved fine-tuning the FM-integrated robotics system to better align with the specific capabilities and limitations of edge devices. This process aimed to maintain functionality while optimizing for resource efficiency [6]. Simulated Testing on Raspberry Pi 4: Simulated testing was conducted to validate the performance of the FM-integrated robotics autonomy on the Raspberry Pi 4. Scenarios were designed to mimic real-world situations, and the system's responses were assessed for responsiveness and accuracy.

V. FUTURE SCOPE:

In the coming years, the integration of chatbots with physical devices like Raspberry Pi, enhanced by voice interaction, promises a transformative user experience. Users will seamlessly control their surroundings through natural conversation, from managing household appliances to accessing personalized assistance and educational support.

VI. CONCLUSION:

In the wake of rapid advancements in artificial intelligence, our exploration into integrating Foundation Models (FMs) into robotics autonomy and virtual assistants marks a significant stride toward realizing intelligent. Our journey showcased the remarkable versatility of Foundation Models, particularly in the context of robotics autonomy. From real-time decision-making to Visual Question Answering, FMs exhibited emergent behavior, demonstrating their adaptability to diverse applications. Seamless Integration on Edge Devices: Extending our implementation to the Raspberry Pi highlighted the potential for deploying sophisticated AI-driven systems on resource-constrained edge devices. This not only enhances accessibility but also opens avenues for edge computing in robotics and beyond. Human-like The exploration of FMs in virtual assistants unveiled the potential for creating more dynamic, contextually aware, and human-like interactions. The integration of multimodal capabilities, personalization, and continuous learning sets the stage for a new era in conversational AI. Ethical As we venture into the future of AI applications, addressing ethical considerations and mitigating biases remain paramount. Responsible AI practices, including transparency, fairness, and user privacy, should be integral components of AI-driven systems.

REFERENCES

- [1]. Tom B. Brown, Benjamin Mann, Nick Ryder, "Language Models are Few-shot learners." 7, no. 4 (2020)
- [2]. Sai Vemprala, Rogerio Bonatti, Arthur Buckler, Ashish Kapoor. "ChatGPT for Robotics: Design Principles and Model Abilities." no 09 (2023)
- [3]. Fanlong Zeng, Wensheng Gana , Yongheng Wang, Ning Liu, Philip S. Yu "Large Language Models for Robotics: A Survey." (2023)
- [4]. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkorei, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin "Attention Is All You Need." (2017)
- [5]. Maria Balota, Sebastian Borgeaud, Adam Trischler. "Conversational AI: Advances, Challenges, and Future Directions." (2022)
- [6]. Alexander Amini, Ava Soleimany, Sandy Huang, Hongseok Namkoong, Aurick Zhou, Yi Ding, Eric Gao, Ye Yuan, Evan Krause, Guillaume Leclerc, Yuan Cao, Jonathan Frankle, Omer Levy, Luke Zettlemoyer. "Scaling Dialogue State Tracking Models." (2023)

