Custom GPT - Democratizing LLM app access and fostering collaborative app development

Dr.B.Monica Jenefer[1]*, Manvith BV [2], Jeff Samuel S[3], Yugabharathi R[4]
Computer Science and Engineering, Meenakshi Sundararajan Engineering College, Chennai 600024, Tamilnadu, India

***Abstract*— In today's rapidly advancing technological landscape, the demand for accessible and powerful AI solutions is ever-growing. Large Language Models (LLMs) have emerged as fundamental tools across various applications, yet traditional LLMs present limitations such as a constrained context window and lack of additional features. To address these issues and meet the pressing demand for a comprehensive solution, this project introduces Custom GPT, an innovative platform designed to democratize access to LLM applications while incorporating advanced functionalities. Custom GPT is a user-friendly, free, and open-source ecosystem that fosters collaboration and innovation among a diverse user base, including students and professionals. Leveraging the capabilities of Mistral, an open-source language model, Custom GPT empowers users to effortlessly develop and share AI applications tailored to their specific use cases. Through a streamlined process of few-shot prompting, users can articulate their app ideas and create bespoke applications without the need for extensive coding skills. In addition to overcoming the limitations of traditional LLMs, Custom GPT offers a wide range of supplementary functionalities, including a code interpreter, to enhance the user experience. This comprehensive platform serves as a hub for users to discover, create, and share AI applications, thereby fostering a collaborative ecosystem where innovation thrives. By providing a platform that is both accessible and feature rich, Custom GPT aims to revolutionize the way users interact with and harness the capabilities of LLMs. Through its commitment to openness, accessibility, and innovation, Custom GPT stands as a potential benchmark in the realm of AI application development, empowering users to unlock the full potential of large language models for their diverse needs.**

***Index Terms:* Python, Fast API, Langchain, Mistral, PWA (Progressive Web Application), JavaScript, LLM (Large language model), Fine Tuning, Mongo Db.**

I. INTRODUCTION

In the era of personalized technology, Custom GPT emerges as a solution, allowing users to articulate their app ideas through prompts. This introduction sets the stage for a revolutionary approach to democratizing AI application development. Custom GPT is designed to democratize app creation, allowing users to bring their unique ideas to life without the need for extensive coding skills. The platform harnesses the power of Mistral, an open-source language model, to enable users to seamlessly craft and share their applications with the community. The automation of assistant creation will be facilitated through the utilization of few-shot prompting. The model will be prompted to generate a JSON output, enabling the seamless development of custom AI applications. Additionally, this web application will be designed as a progressive web app, offering users the convenience of accessing it as if it were their native app on either Android or iOS devices. In order to create their own custom-tailored app, all they need to do is just give a prompt on what they want to create. To craft their own bespoke applications, users will simply need to provide a prompt outlining their requirements. This streamlined process ensures accessibility and empowers users to effortlessly tailor applications to their specific needs.

II. LITERATURE SURVEY

A Survey on Large Language Models:
Applications, Challenges, Limitations, and Practical Usage. Muhammad Usman Hadi, Qasem Al-Tashi, Rizwan Qureshi.

A Survey on ChatGPT and Open-AI Models: A Preliminary Review Sheradha Jauhari, Chetan Aggarwal, Apoorv Gautam and Diksha Awal Virtual.

Mock Interview Assistant (Video Bot-based) Konstantinos I. Roumeliotis and Nikolaos D. Tselikas

A Survey on Algorithms, Techniques, and Applications: Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, S. Iyengar.
A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios: Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, Dietrich Klakow.
A Survey on ChatGPT: AI–Generated Contents, Challenges, and Solutions: Yuntao Wang, Yanghe Pan, Miao Yan, Zhou Su, Tom H. Luan.

III. PROPOSED METHODOLOGY

Custom GPT envisions a user-friendly web application where individuals can craft their own AI applications using a set of prompts. The proposed system not only streamlines the creation process but also encourages collaboration by allowing others to benefit from these custom applications. Custom GPT integrates the Mistral open-source language model, providing a robust foundation for natural language understanding and generation. Mistral enhances the app creation process by enabling the system to comprehend and respond to a wide range of user prompt .The proposed system offers a transformative solution to the existing challenges in AI application development, presenting a dynamic and inclusive web application platform Central to this innovative platform is the integration of Mem GPT, a cutting-edge language model that effectively addresses the context window problem, enabling users to generate more coherent and contextually relevant responses. Progressive Web Application: Our platform is designed as a progressive web application (PWA), ensuring seamless access and compatibility across various devices and operating systems.

IV.ARCHITECTURE DIAGRAM

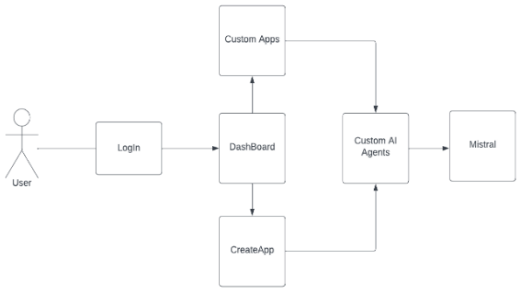

Figure.1 Block diagram for Custom GPT

Based on the provided headings, here's a breakdown of how you might explain the architecture diagram of Custom GPT:

1. User
 - This section represents the users interacting with the Custom GPT platform.
 - Users can include individuals, teams, or organizations seeking to utilize AI applications or contribute to the platform.
2. Login
 - Describes the authentication and authorization process for users to access the Custom GPT platform.
 - Includes components related to user account management, such as registration, login, and profile settings.
3. Custom Apps
 - Explains the functionality for users to create, explore, and utilize custom AI applications within the Custom GPT ecosystem.
 - This section includes components related to

browsing existing applications, creating new ones, and managing application settings.

4. Dataset
 - Discusses the management and integration of datasets within the Custom GPT platform.
 - Components may include tools for importing, preprocessing, and utilizing datasets in AI application development.
5. Create App
 - Details the process and components involved in creating a new AI application within Custom GPT.
 - This section encompasses features such as application templates, development environments, and deployment options.
6. Custom AI Agents
 - Describes the concept of custom AI agents within Custom GPT, which are specialized instances of AI models tailored to specific tasks or domains.
 - Components may include tools for training, fine-tuning, and deploying custom AI agents.
7. Mistral
 - Introduces Mistral, the open-source language model integrated into the Custom GPT platform
 - Discusses how Mistral powers various AI functionalities within Custom GPT, such as natural language processing and response generation.

This breakdown provides an organized explanation of the architecture diagram, highlighting key components

and their functionalities within the Custom GPT platform

V.RESULT AND IMPLEMENTATION

After fine tuning mistral we were able to make the model to give good results in a Json format without any errors, the Training loss when fine tuning was very low. This has been fine-tuned using a method called as Qlora, It was made possible through the

process of quantization. With this we got a very nice adapter that we can use to make inference on the model

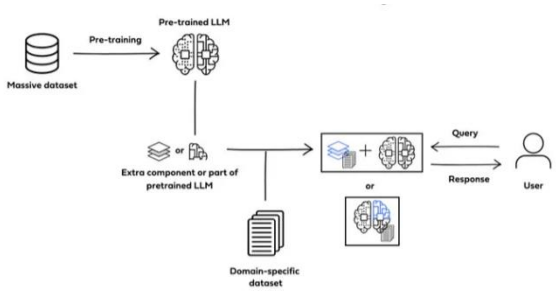

Figure 2 Parameter-Efficient Fine-Tuning

With the help of fast Api and MongoDB, a nice backend was made for users to create their apps. The configuration for user created apps has been stored in the MongoDB for future use.

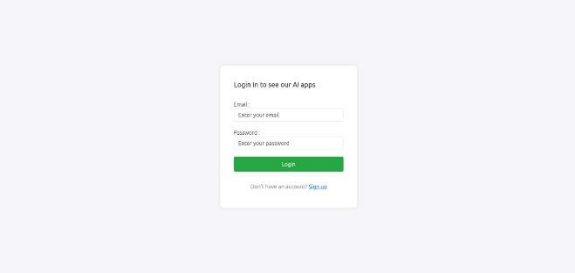

Figure 3 Login page

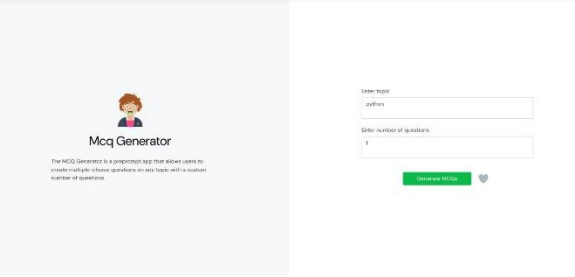

Figure 4 Customized AI generator

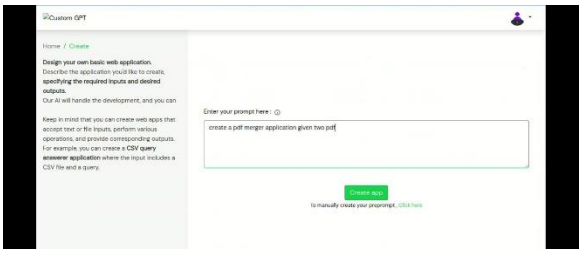

Figure 5 Customized App Creation

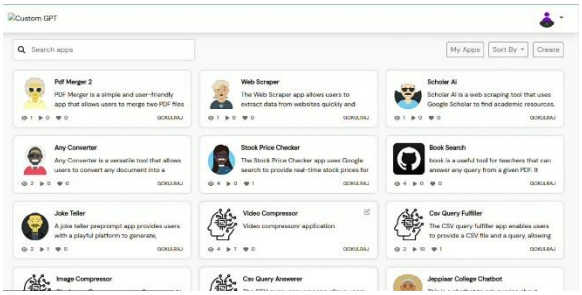

Figure 6 Home page

VI.CONCLUSION

In conclusion custom GPT can be adapted to solve day to day problems without any constraints and in more efficient and user-friendly manner. It would be able to accept pictures as well as videos and also provide information about real time data which could be done by many AI's out there in the internet and moreover this software is a progressive web application allowing the users to download the apps created in this platform both in mobiles as well as windows devices, also it is responsive. This project encourages all the users to create an app solving their specific problem and moreover allowing other users to install the previously created applications making it very socially active application.

REFERENCES

[1] T. Webb, K. J. Holyoak, H. Lu, Emergent analogical reasoning in large language models, Nature Human Behaviour 7 (9) (2023) 1526–1541. 2

[2] D. A. Boiko, R. MacKnight, G. Gomes, Emergent autonomous scientific research capabilities of large language models, arXiv preprint arXiv:2304.05332 (2023). 2

[3] G. Izacard, P. Lewis, M. Lomeli, L. Hosseini, F. Petroni, T. Schick, J. Dwivedi-Yu, A. Joulin, S. Riedel, E. Grave, Few-shot learning with retrieval augmented language models, arXiv preprint arXiv:2208.03299 (2022). 2, 17, 18, 33

[4] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter,

A. Wahid, J. Tompson, Q. Vuong, T. Yu, et al., Palm-e: An embodied multimodal language model, arXiv preprint arXiv:2303.03378 (2023). 2, 19, 21, 33
[5] A. Parisi, Y. Zhao, N. Fiedel, Talm: Tool augmented language models, arXiv preprint arXiv:2205.12255 (2022). 2, 18, 19
[6] B. Zhang, H. Soh, Large language models as zero-shot human models for human-robot interaction, arXiv preprint arXiv:2303.03548 (2023). 2, 33
[7] Q. Ye, H. Xu, G. Xu, J. Ye, M. Yan, Y. Zhou, J. Wang, A. Hu, P. Shi, Y. Shi, et al., mplug-owl: Modularization empowers large language models with multimodality, arXiv preprint arXiv:2304.14178 (2023). 2, 22
[8] W. Wang, Z. Chen, X. Chen, J. Wu, X. Zhu, G. Zeng, P. Luo, T. Lu, J. Zhou, Y. Qiao, et al., Visionllm: Large language model is also an open-ended decoder for vision-centric tasks, arXiv preprint arXiv:2305.11175 (2023). 2, 22
[9] R. Yang, L. Song, Y. Li, S. Zhao, Y. Ge, X. Li, Y. Shan, Gpt4tools: Teaching large language model to use tools via self-instruction, arXiv preprint arXiv:2305.18752 (2023). 2, 19, 22
[10] E. Saravia, Prompt Engineering Guide, https://github.com/dairai/Prompt-Engineering-Guide (12 2022). 2, 7, 17, 33
[11] A. Zeng, X. Liu, Z. Du, Z. Wang, H. Lai, M. Ding, Z. Yang, Y. Xu, W. Zheng, X. Xia, et al., Glm-130b: An open bilingual pre-trained model, arXiv preprint arXiv:2210.02414 (2022). 2, 10, 22, 23, 25
[12] Y. Wang, H. Le, A. D. Gotmare, N. D. Bui, J. Li, S. C. Hoi, Codet5+: Open code large language models for code understanding and generation, arXiv preprint arXiv:2305.07922 (2023). 2, 10, 24, 25
[13] S. Wang, Y. Sun, Y. Xiang, Z. Wu, S. Ding, W. Gong, S. Feng, J. Shang, Y. Zhao, C. Pang, et al., Ernie 3.0 titan: Exploring larger-scale knowledge enhanced pre-training for language understanding and generation, arXiv preprint arXiv:2112.12731 (2021). 2, 8, 23, 25
[14] J. Rasley, S. Rajbhandari, O. Ruwase, Y. He, Deepspeed: System optimizations enable training deep learning models with over 100 billionparameters, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 3505–3506. 2, 5
[15] S. Rajbhandari, J. Rasley, O. Ruwase, Y. He, Zero: Memory optimizations toward training trillion parameter models, in: SC20: International Conference for High Performance Computing, Networking, Storage and Analysis, IEEE, 2020, pp. 1–16. 2, 4, 23
[16] J. He, C. Zhou, X. Ma, T. Berg-Kirkpatrick, G. Neubig, Towards a unified view of parameter-efficient transfer learning, arXiv preprint arXiv:2110.04366 (2021). 2, 20, 21
[17] Z. Hu, Y. Lan, L. Wang, W. Xu, E.-P. Lim, R. K.-W. Lee, L. Bing, S. Poria, Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models, arXiv preprint arXiv:2304.01933 (2023). 2, 20
[18] B. Lester, R. Al-Rfou, N. Constant, The power of scale for parameter efficient prompt tuning, arXiv preprint arXiv:2104.08691 (2021). 2, 8, 20
[19] X. L. Li, P. Liang, Prefix-tuning: Optimizing continuous prompts for generation, arXiv preprint arXiv:2101.00190 (2021). 2, 20
[20] X. Ma, G. Fang, X. Wang, Llm-pruner: On the structural pruning of large language models, arXiv preprint arXiv:2305.11627 (2023). 2, 21