Breaking Barriers: Enabling User-Focused Applications through LLMS (Extended Abstract)

Macphail Magwira Junior
AIDAS (Artificial Intelligence Data Analytic Systems, an early-stage startup, www.aidas.co.za)
Malawi University of Business & Applied Sciences and the Inspire Africa PEACH project, a pan-African Covid-19 Data Initiative (www.inspiredata.network)

Abstract:

This paper introduces the Hive platform, a user-friendly and accessible solution powered by large language models (LLMs) available at https://www.hiveapp.dev, that aims to enable user-focused applications. LLMs have emerged as a powerful tool in Artificial Intelligence, but their adoption has been hindered by barriers such as technical knowledge and resources. The Hive platform addresses these challenges by providing an intuitive graphical flow-builder and pre-designed templates, allowing users to build tailored AI assistants quickly. By leveraging techniques such as in-context learning and prompt construction while incorporating moderation and evaluation features, the platform not only enables the development of incredible LLM apps but also promotes responsible and ethical AI usage. The potential impact of the Hive platform extends across various domains, empowering businesses, educators, and content creators to enhance their processes and engage with users more effectively. Looking ahead, the Hive platform will continue to evolve, incorporating advancements in LLM technology and exploring the integration of LLMs with other AI technologies.

1 Introduction:

Large language models (LLMs), including OpenAI's GPT-3, Google's BERT, and META's LLaMa, have emerged as powerful tools in Artificial Intelligence, although not a relatively new technology, researchers and technologists are actively exploring the vast potential of LLMs and seeking effective ways to leverage their capabilities. LLMs represent advanced artificial intelligence systems designed to comprehend and generate human-like text. These models undergo extensive training on massive datasets, enabling them to excel at understanding complex patterns, grasping language nuances, and generating coherent responses. They exhibit versatility and are capable of performing diverse language-related tasks, such as translation, text completion, summarization, and engaging in

conversational interaction and fall under the umbrella of Generative AI. Throughout history, the convergence of technology creation and user adoption has been pivotal for widespread benefit, for example, Steve Jobs effectively communicated the potential of computers and the graphical user interface by presenting them in a user-friendly form most commonly known as "the home computer." Following a similar path, OpenAI has made LLM technology easy to use and accessible, resulting in one of the fastest-growing tech adoptions worldwide. This highlights the importance of not only creating new technology but also innovating around its adoption. By recognizing the significance of user-friendly adoption, we can unlock the true potential of LLMs and accelerate their impact across various domains. In this paper, I introduce our ongoing work, the Hive platform, it serves as a user-friendly and accessible solution, akin to WordPress for AI assistants, powered by LLMs. We have developed an easy-to-use graphical flow-builder, along with pre-designed templates, allowing anyone to quickly build assistants tailored to specific needs. As an example, one of our initial templates focuses on a course tutor AI companion, the motivation behind this template stems from our own personal problem as online distance learners, being outside the classroom, it had always been a bit challenging to learn new challenging concepts on our own, even with our instructor available via email, a personal tutor available to us 24/7 that could teach and help us learn in a way that we understood and preferred felt invaluable, and thus was born our first template, enabling users/students to input their course material such as lecture notes, audio, and videos onto our platform and utilize an AI companion to enhance their learning experience. With the Hive platform, we aim to facilitate the adoption of LLM technology and empower users to harness the capabilities of LLMs in a user-friendly manner. This paper provides an overview of our ongoing work, its objectives, and the potential impact it can have in various domains. In the words of Professor Adam Habib, former vice chancellor of Wits University, the trick is to utilize the world's knowledge and apply it to our circumstances, that way we innovate and contribute to the world.

2 Emerging architectures for LLM applications

Pre-trained AI models represent the most important architectural change in software since the internet. They make it possible for individual developers to build incredible AI apps in a matter of days/minutes that surpass supervised machine learning projects that took big teams months to build.

2.1 In-context learning

In-context learning leverages large language models (LLMs) without fine-tuning by controlling their behavior through clever prompting and conditioning on private contextual data. Rather than sending all the documents to the LLM, in-context learning selects the most relevant documents determined by the LLMs themselves. The workflow involves data preprocessing and embedding, prompt construction and retrieval, and prompt execution and inference. In-context learning reduces the need for fine-tuning and specialized ML engineers, making it a data engineering problem that most companies can solve. It outperforms fine-tuning for small datasets and allows for near real-time incorporation of new data. While expanding the context window is possible, it comes with trade-offs on cost and inference time.

2.3 Prompt construction and retrieval

Developers adopt increasingly complex strategies for prompting large language models (LLMs) and incorporating contextual data to differentiate their products. While simple prompts with direct instructions or example outputs are commonly used, they often fall short of production-level accuracy. Advanced prompting strategies, such as chain-of-thought, self-consistency, and generated knowledge, are being explored to ground model responses in truth and provide external context. Orchestration frameworks like LangChain and LlamaIndex simplify prompt chaining, API interfacing, contextual data retrieval, and memory maintenance. They offer templates for common applications and are popular among hobbyists and startups. ChatGPT, although primarily an app, can also serve as an API and perform similar functions as orchestration frameworks, potentially becoming an alternative to prompt construction in the future.

2.4 Prompt Execution & Inference

OpenAI is currently the leader in the field of large language models (LLMs), and most developers start their LLM applications using the OpenAI API with models like gpt-4. These models offer good performance and are easy to use without the need for fine-tuning or self-hosting. However, as projects scale, developers have a wider range of options. Some common strategies include switching to the more cost-effective and faster gpt-3.5-turbo model, exploring proprietary vendors like Anthropic's Claude models for customization options and larger context windows, and triaging requests to open-source models, especially for high-volume B2C use cases. Open-source models are also improving, with the LLaMa models from Meta are setting a new standard for accuracy and prompting the development of alternative base models by other providers. As open-source LLMs approach the accuracy levels of proprietary models, we anticipate a surge in experimentation, and sharing of fine-tuned models, facilitated by hosting companies and operational tooling. Various tools are being used for caching, logging, tracking, evaluation, validation, and prompt injection detection. Hosting solutions range from standard options like Vercel and cloud providers to specialized startups providing end-to-end hosting or platforms combining model hosting and code deployment.

3 LLM Application Building Blocks

The Hive platform incorporates various building blocks to facilitate the development and deployment of LLM applications:

3.1 Graphical Flow-Builder: The graphical flow-builder offers a visual interface that enables users to design their AI assistants. Users can create interactive dialogues and define the prompts and responses to guide the LLM's behavior. This intuitive interface eliminates the need for extensive programming knowledge, making it accessible to a wide range of users.

3.2 Pre-designed Templates: A collection of pre-designed templates that serve as starting points for building LLM applications. These templates are designed for specific domains or use cases, such as course tutoring, and customer support. Users can customize these templates to suit their requirements, saving time and effort in the development process.

3.3 In-context Learning Integration: Leveraging the power of in-context learning, our platform incorporates techniques to control the behavior of LLMs without the need for fine-tuning. By intelligently selecting and presenting relevant documents to the LLM, the platform ensures that the model's responses are grounded in accurate and contextual information. This integration enables the real-time incorporation of new data and enhances the performance of LLM applications.

3.4 Moderation and Ethics: Recognizing the importance of responsible and ethical AI, our platform incorporates moderation features to ensure appropriate content generation. It enables content classification, output validation, and chain of thought reasoning techniques to promote safe and responsible AI usage. By providing tools for testing, evaluation, and monitoring, the platform allows users to assess the performance and behavior of their AI assistants.

4 Potential Impacts and Future Directions

The Hive platform has the potential to unlock the true power of LLMs by enabling user-focused applications in various domains. By democratizing LLM technology, it reduces barriers to entry and facilitates the adoption of AI by businesses and individuals. The accessibility and user-friendly nature of the platform make it possible for users with diverse backgrounds to leverage LLMs for their specific needs, ranging from education and customer service to content generation and personal assistants. Looking ahead our platform will continue to evolve and incorporate advancements in LLM technology.

In conclusion, our platform represents a significant step towards enabling user-focused applications through LLMs. By providing a user-friendly interface, pre-designed templates, and integration with in-context learning, the platform empowers users to harness the capabilities of LLMs without the need for extensive technical expertise. The potential impact of the Hive platform spans industries and domains, driving innovation and transforming the way we communicate, learn, and create. As LLM technology continues to advance and the Hive platform evolves, we can expect to see a wide range of innovative applications that leverage the power of LLMs to break barriers and enable user-focused solutions in the AI landscape.

References:

[1] Adiwardana, Daniel, Minh-Thang Luong, David R. So, and 8 more. "Towards a Human-like Open-Domain Chatbot." 2020: https://ui.adsabs.harvard.edu/#abs/2020arXiv200109977A/abstract

[2] Askell, Amanda, Yuntao Bai, Anna Chen, and 19 more. "A General Language Assistant as a Laboratory for Alignment." 2021:https://ui.adsabs.harvard.edu/#abs/2021arXiv211200861A/abstract

[3] Brown, Tom B., Benjamin Mann, Nick Ryder, and 28 more. "Language Models are Few-Shot Learners." 2020. :https://ui.adsabs.harvard.edu/#abs/2020arXiv200514165B/abstract