**Podcast_Episode_28- Demystifying Large Language Models for Future of Work - Part 2_Transcript**

You were listening to the HCLTech Digital Workplace Podcast, the place where industry experts, analysts, and veterans help us identify, understand, and prepare for the upcoming digital workplace technologies and trends. If you haven't subscribed to the channel already, do it now for regular updates. This episode starts in 3, 2, 1.

**TJ** - **Welcome. This is the second part of a two-part series. We interacted with Varun Singh, the President and Founder at Moveworks. We discussed the upcoming implications and opportunities related to large language models and generative AI in the enterprise world.**

**So just interacting with and using models like these will essentially help them get better every day or is training them in a different activity altogether?**

**Varun** - Language models can improve over time, and their usage in businesses depends on their application. For instance, if better datasets and user feedback improve the prompt and information provided, the model doesn't need to improve. However, these models also undergo retraining steps, such as ChatGPT, which can be incrementally retrained. The broader training of a new model, like ChatGPT 4, costs over $100 million, so it's unlikely that this will happen regularly. Incremental training is a common practice for these models, but it's not a guarantee of regular retraining. In the absence of sufficient ships, incremental training can still be used to improve the performance of these models.

User feedback at the application level leads to improved language models, not necessarily because the language models improve with it. Fine-tuning these models is not expensive and can be done incrementally, several times a day, or weekly. Additionally, providing more integration can improve performance. Language model limitations may arise from not having the right enterprise system or content quality. Different ways in which applications improve with user feedback can be instant or require more model retaining. Overall, user feedback plays a crucial role in enhancing the performance of applications.

**TJ** - **So, whatever you said right now, is it correct to summarize it in a way that the amount of effort and the frequency of retraining will eventually reduce over time because the more work you put into helping, the model understands you better over time?**

**Varun** - The marginal return on incrementally similar data decreases over time. Retraining can improve when new datasets are introduced. For instance, when COVID first hit, it created a new phrase for our language model. Regular fine-tuning and retraining can improve these models, including automatic entity extraction. Our models were able to recognize the emerging concept of COVID in the enterprise, which is related to remote work, new applications, and potential outages. These retraining cycles help detect and address these new data-related issues.

**TJ** - **So there seems to be some misconception about what's happening under the hood with ChatGPT. A lot of business leaders believe they just take a strategy, ChatGPT plugin and have it take action across their business. But can you talk to us about why that's not the case?**

**Varun** - A ChatGPT plugin is a concept where a ChatGPT is not directly connected to any system or data source on the backend. However, it can be augmented and connected to various things, such as search

engines or restaurant booking websites. When deploying this model in the enterprise, it is important to consider that it doesn't fundamentally solve the problem of factuality. Factual accuracy is crucial for enterprise digital assistants and virtual assistants, and plugins don't solve this problem per second.

Security is another important aspect of deploying ChatGPT. It needs to be done in a secure manner, and it's not guaranteed to solve this problem unless the deployment method has security built into it. Additionally, ChatGPT is a consumer-grade model, and if it doesn't understand the language of work, it may misfire and get things wrong. This could lead to people losing trust in the system and a loss of trust in the system.

Business leaders should consider this when deploying applications, as they don't want a system that hallucinates and gets things wrong. To solve this problem, business leaders should lead analytics on the system to understand its performance, improvements, control, iteration, and access. Overall, deploying ChatGPT in the enterprise is not super straightforward.

**TJ** - **For businesses who have decided that they need a very robust and effective generative AI-based product in their organizations to enable their workforces, is it recommended to go for the biggest and the best LLMs or is it better to have multiple models working together?**

**Varun** - The bigger LLMs are powerful in many ways but are not practical for many purposes due to their high costs, latency, and difficulty in control. Instead, it is important to consider the enterprise in a new movement of smaller language models that are good at specific tasks. Many applications deploy these smaller language models in combination with larger language models when creating AI applications.

An example of a future architecture for AI applications is ChatGPT 4. GPT 3.5, which takes a request combined with conversation context and other language models. This system can then use tools like ChatGPT four to find opportunities, perform specific operations, and summarize information. Each tool has a description of what it does, and ChatGPT four can read that description and assemble the response for the user.

This architecture is extensible, as it can perform calculations on its own without the need for specific logic built into the application. This architecture represents the future where large language models are surrounded by smaller, task-specific language models that are faster to run, more manageable, more controllable, and very performant.

In summary, the future of AI applications will involve a combination of larger language models for reasoning and task-specific language models for specific tasks. This architecture represents the future, as many of these models are coming together to outperform a single model.

**TJ** - **So Vision 2030 ideal case scenario, what do you envision move works to be and what other aspects of business do you think Moveworks will get involved in the daily working of each and every person in the workforce?**

The company's vision focuses on eliminating the challenges faced by knowledge workers and frontline workers, aiming to improve productivity and eliminate the drudgery of low productivity. we don't have a vision carved out for product strategy that takes us very far. We tend to look focus on the next 18 months because we believe we can make significant advancements with the latest technology and align our roadmap accordingly. However, the journey of building a product depends on the successful execution of

the strategy. If something doesn't work, the company must decide whether to continue, suspend it, or reduce investment.

The company's vision for Vision 2030 is to touch thousands of organizations across various help and service use cases, including recruiting, finance, customer support, and IT support. The aim is to be the best company in leveraging language models in the work context and building smart applications that employees enjoy using. It is difficult to predict the future of technology, as it has taken a surprising rate of improvement in recent years. The founder of OpenAI, for example, emphasized the importance of leveraging language models in the work context and building smart applications that employees enjoy using.

In conclusion, the company's vision is to eliminate the challenges faced by knowledge workers and frontline workers, focusing on improving productivity and leveraging language models in the workplace.

**TJ** - **That's so 15 years down the line you might even have quantum computing. In enterprises, do you feel that would amplify or exponentially increase the effectiveness of something like ChatGPT or will it be more linear?**

**Varun** - In the industry, people are exploring different architectures to train models that might be better than transformer-based models. Google's Pathways model is a powerful model that trains their palm model. Yann LeCun and Meta have defined a model that can understand the real world, make predictions, plan tasks, and accomplish goals in the real world. Critics argue that large language models don't directly understand the real world by observing it. However, if architectures could allow these models to understand the real world, they might be better.

The next big change could be a non-linear moment, with reinforcement learning continuing to provide models that behave in surprising ways. ChatGPT 4 has outperformed ChatGPT three in rigorous tests, making it a non-linear move. The application landscape will change dramatically in the next five years, as people start to use language models in powerful ways without any progress on the language model front. Even if language models don't improve from where they are today, they will be transformative in the next 10 years in a non-linear manner.

**TJ** - **Can you talk about the role of LMS fine-tuning and grounding? Why those two things are significant for businesses looking to derive maximum value from LMS?**

**Varun** - Large language models like ChatGPT are trained on internet data and aim to be truthful to that data. However, businesses have proprietary datasets that these models are not trained on. To make predictions within a business, it is crucial to fine-tune these models to be more truthful to the business's data set. This involves fine-tuning the weights of the models to be more truthful to the proprietary dataset, which these models may not have access to.

Grounding is another step in fine-tuning large language models. By providing better descriptions from enterprise data sets through engineering applications or interfaces, these models can perform better in certain scenarios. This is the idea of grounding, which is to be truthful to the dataset.

Both fine-tuning and grounding are independent of the quality of large language models. The concept of grounding is crucial in understanding the performance of large language models in enterprise data. Enterprise data is proprietary and unique for each business, and fine-tuning is essential for achieving

meaningful results. Moveworks, for example, has a vast data set of millions of interactions with chatbots, which is combined with their machine learning operations platform. This data can be fine-tuned, improving the performance of large language models by 50% in certain areas. This not only improves the model's understanding of HR, benefits, and finance queries but also improves the overall performance of the application.

**TJ - What are the core capabilities of Moveworks that help enable enterprise transformation? And how does Moveworks help employees navigate the entire workday and remain productive?**

**Varun** - Moveworks aims to ensure employees don't have to wait for help and remain productive while businesses are not inefficient at delivering services to employees. To achieve this, Moveworks has several core applications that customers use, including an IT-specific application with language models that understand technology's language, integrated with various enterprise systems, identity systems, ticketing management, case management systems, and workflow systems. Corolla, an HR space, integrates with various systems to understand HR language and solve issues for employees.

Customers have seen a 60-70% reduction in ticket volume due to these applications. Additionally, Moveworks has an enterprise search application that connects Moveworks to knowledge bases, files, documents, and PowerPoints, allowing for conversational questions and response. This application has been built incrementally to improve workforce productivity by proactively informing employees about their environment.

Moveworks also focuses on analyzing unstructured data, such as case management systems and ticketing systems, to understand the voice of customers and prioritize interventions, projects, and service delivery. Employee Experience Insights is a new application that applies language models to unstructured data, allowing businesses, CIOs, and technology leaders to understand the state of service delivery in a completely unvarnished way.

The latest release, Creator Studio, empowers service owners and developers within an enterprise to use enterprise-specific large language models to build new conversational use cases within the business. This enables developers to build hundreds of use cases using these language models, enhancing the value that businesses receive from the Moveworks platform.

**TJ** - Is this a no-code platform?

**Varun** - It is a no-code, low-code service ownerlatorm, enabling customers to deploy a dozen use cases within a day. We have participated in hackathons and have seen customers deploying multiple use cases within a day, demonstrating the platform's efficiency and adaptability.

**TJ - So, before we let you go, tell us more about Movework's partnership with HCLTech.**

**Varun** - HCLTech is a global leader in the digital workplace, serving prominent businesses for digital transformation, automation, and revenue goals. As partners, our vision is to transform these businesses to become more efficient, grow faster, and serve their end customers better. By partnering with HCLTech, we gain a deeper understanding of other technologies, such as Moveworks, which helps us better serve our joint customers. As a leader in generative conversational AI, we appreciate the opportunity to partner with leaders like HCLTech in the digital workplace and serve our joint customers.

**TJ** - Thank you very much, Varun, this was a very enlightening conversation and I think when it comes to discussing generative AI, we are still scraping the tip of the iceberg so we look forward to hosting you again where we can continue this discussion.

And we wish you a good stay, a happy stay here in India. See you soon.

**Varun:** Fantastic. Thank you for having me.

This episode of the HCLTech Digital Workplace Podcast has ended, but be sure to subscribe for more insights on how to identify, understand, and prepare for a world of possibilities around the new and upcoming digital workplace technologies and trends. Don't forget to rate and review this episode so that we can keep bringing you the most relevant content.

Thank you for listening.