www.pipelinepub.com
Volume 21, Issue 8

Preparing for GenAI Innovation

By: Mark Cummings, Ph.D.

With the explosive appearance of this generation of GenAI, it is hard to think about preparing for the next disruption. Hard, but necessary. The GenAI field is ripe for innovation in the two areas of software and hardware. It is easy to miss the true innovation bubbling up because of other AI news coming out. With new versions of existing LLMs (Large Language Models) coming out every few weeks, it is easy to fall into the trap of thinking that the incremental improvements we are getting in GenAI are the only genuine innovations. This is while we are going through a period of stumbling implementations turning into effective intelligent agents.

This ongoing stream of technical innovation creates a lot of anxiety about job losses from GenAI automation. Thus, innovation is critical to address this anxiety and the counterproductive forces it engenders.

Some may wonder if the financial situation around tariffs will slow AI innovation. It is not likely to. The battle for AI supremacy is expected to keep it going. However, it is possible to add to the anxiety.

Successful leaders will not fall behind in the ongoing AI-driven competition. This means that those developing GenAI products need to be careful not to get too far heads down in existing development streams and miss the next wave of disruption. For those implementing applications, there may be more time. Early signs of disruption may take longer to reach the stage where they are ready for application implementation. However, these application developers must now be careful to document and archive the detailed requirements they are using with this generation of GenAI. This documentation will allow them to speed up the deployment of the next disruptive generation. Leaders must recognize the anxiety around GenAI worker displacement to avoid productivity losses and take active steps to reassure employees that they will be cared for.

AI Technology Innovation

On the surface, current GenAI technology development is relatively incremental. Things like creating ever-larger models, new ways of doing tuning, making smaller LLMs by pruning larger ones, incremental improvements in current GPU and TPU architectures, etc., are happening. Under the surface, innovation is going on in new software and new hardware architectures.

In the software space, there are indications of new model architectures being developed in both big companies and early-stage start-ups. One example of this in large companies is Yann LeCun, VP and Chief AI Scientist at Meta and a professor at NYU. He is leading an effort to develop one view of the next generation of AI (Joint Embedding Predictive Architecture - JEPA). He suggests a new paradigm with a model that includes understanding the physical world, persistent memory, reasoning, and complex planning capabilities. He expects these models to be available for use in three to five years.

Another example is an early-stage start-up, Vital Statistics Inc., which has technology that makes today's static LLMs dynamic. Once trained, today's LLMs have a static structure: a set number of columns and rows. This technology makes the shape of the model dynamic. It promises dramatic improvements in speed of inference, accuracy of results, and lowering of power consumption.

In the hardware space, one of the key issues is increasing the efficiency of systems. The actual individual calculations in LLM are relatively simple. However, they are performed as part of extensive matrix calculations. One way of improving efficiency is speeding up data transfer to and from processors and between processors. Abacus Semiconductor is working on technology to lower latency and power consumption in this communications process. Those working in this space are trying to improve the movement of data to make more efficient processing rather than just faster processors. Some other early-stage start-ups are focusing on innovations in how data is moved. Another hardware approach being explored by an early-stage start-up is focused on extending the architecture developed for GPUs to more general-purpose processors. Others are working on analog approaches.

The amount of software focused on Nvidia's proprietary CUDA interface (combination of APIs and module libraries) makes it difficult for other semiconductor vendors. Groq and Cerebras are addressing this by using their proprietary chips to run data centers that offer inference as a service. Apple and Google Cloud are doing similar things. Apple, with its M Series processors, is being used in its data centers to support LLMs it runs to support its services. Google, using Google, developed TPUs it runs to support its services and Nvidia chips for running other people's LLMs in Google Cloud.

These are just examples. There are likely to be many more innovations on the threshold. Traditionally, software innovation has been faster than hardware because of the cost barriers to hardware innovation/proof of concept. With GenAI, however, the cost of proving a concept has become quite large. Building, training, and testing any LLM is quite expensive. Breakthroughs that lower costs are an advantage here. With that said, software innovations typically still come to deployment faster than semiconductor hardware innovations. Finally, these examples may not be successful. What appears to be good fundamental innovation can turn out not to be technically or business feasible. However, some of the streams of innovation that they exemplify will be successfully deployed.

Workforce Anxiety

Some predict that GenAI will automate 80% of jobs. CEOs are saying, some publicly, some privately, that in order to start a new hiring process it must be proven GenAI can't do the job. All of this is leading to high anxiety in the workforce. The anxiety is broad-based. A leading industry consultant told me privately that the word that best described the year's MWC (Mobile World Congress - one of the largest trade shows held every Spring) was anxiety.

Workforce anxiety, anger, and resentment lead to poor individual performance, which can lead to organizational chaos. The challenge is for leaders to develop ways of supporting their employees, reducing anxiety, and avoiding poor performance and organizational chaos.

Training, outplacement, and termination packages have been used in the past. These may still be helpful. However, they are based on the assumption that automation will replace low-skill jobs and that

there will be up-skill opportunities. GenAI has brought that into question. Some predict that only people who can build the next generation of AI systems will have jobs. Others maintain that AI itself will create the next generation of AI.

This is a challenging area ripe for innovation. However, it appears that those with the expertise, experience, and incentives to create this innovation are distracted by other things. Thus, successful business leaders will make it clear to their organizations that this is a problem area that needs to be solved, and good work will be rewarded. Leaders who take the path of trying to ignore the problem. Pretend it doesn't exist. Continue business as usual. Will confront growing employee morale problems, such as declining performance, etc.

Preparing for the Next Generation

There is so much pressure to release the next incremental update that it is easy to miss the next disruptive innovation in GenAI software and hardware. Current naming conventions around software incremental changes are confusing, and disruptive software products may suffer from the same confusion. It appears that APIs and associated tools are moving in a direction that may be workable. However, ways of describing behaviors of particular GenAI systems can be problematic.

Application developers can be too heads down as well. Focused on solving today's problems. They also need to think about the path forward to next-generation systems. One way to do that is to document and archive the detailed requirements they are using carefully. Try to understand what they are seeking to achieve clearly. What portions of that had to be sacrificed to fit in this generation? Where they were successful. Where they failed. What they expect the path to the next generation to look like. And how they want to get there. How they can make the transitions between generations easy for their end users, system maintainers, etc. Such documentation will allow them to speed up developing and deploying next-generation solutions.

Understanding the behaviors of particular GenAI systems is particularly important in intelligent agents responsible for controlling physical things. Examples include autonomous vehicles, drones, factory robots, farming robots, controlling electrical grids, etc. Replacing an older generation GenAI system in such applications with a new generation one is not a simple plug-and-play operation. Even if the APIs seem close enough to work, the differences in fundamental behavior may produce unexpected and unacceptable responses. Application developers must be cautious until better ways of describing particular GenAI systems' behaviors are common.

Each project plan should have a component on addressing employee anxiety around GenAI in general and how displacements generated by the deployment of this application will be handled.

Conclusion

Success leaders in the GenAI innovation environment will understand and prepare for the coming generations of innovation. The competitive drivers behind AI will not stop. These leaders will drive current activities to make efficient use of current generations of AI while preparing for future generations of disruption. At the same time, they will find innovative ways to take care of their employees, ways that lower employee anxiety, creating high levels of performance and organizational success.