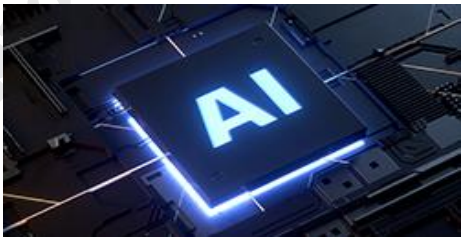
www.pipelinepub.com
Volume 20, Issue 11

Five Significant GenAI Innovations for Technology Leaders

By: Sanjay Basu, Ph.D, Akshai Parthasarathy



Generative AI (GenAI) is predicted to deliver value comparable to that of the Internet. While AI technology can be considered a solution to critical challenges faced by public and private organizations, the increasing complexity and scale of AI models are driving the need for high-performance, secure, and compliant solutions that meet the needs of large enterprises.

Organizations, including those in the public sector, must navigate regulatory requirements and applicable data privacy regulations, emphasizing the importance of sovereign clouds for data control and compliance. Meanwhile, for private enterprises, productivity can be hampered by time-consuming and error-prone tasks, and existing AI techniques may be limited in their ability to handle diverse data types and complex scenarios, highlighting the need for advancements that push the boundaries of AI capabilities.

Five key focus areas will shape the future of GenAI and tackle its challenges: AI infrastructure, sovereign clouds, agentic workflows, retrieval augmented fine-tuning (RAFT), and multi-modal AI. Solutions from these areas will help GenAI's computational power, data privacy, productivity, accuracy, and capability to handle diverse types of data.

AI Infrastructure: Running the most Demanding AI Workloads Faster

According to Goldman Sachs, AI is poised to drive 160 percent of all datacenter power demand. Behind models like Open AI GPT 4o and Meta Llama 3 are massive, scale-out infrastructures that consist of high-performance compute, storage, network, and software for AI. Datacenter power demand remained flat until 2019 and has steadily increased since then to meet the accelerated increase in workload demand.

AI is a notable contributor to datacenter power and workload demands. Cloud service providers, including OCI, are investing billions of dollars into massive computing clusters, such as the **OCI Supercluster**. These superclusters consist of GPU instances, cluster networking, and high-performance storage, including file systems that are designed to handle larger parallel AI training workloads.

While a significant amount of datacenter infrastructure is currently being used for AI training, with the advent of more capable, trained foundation models, we expect that AI inference may play an increasing role in datacenter power and workload consumption.

Sovereign AI: Achieving Digital Sovereignty and Control of AI Data

Sovereignty in the cloud is an important consideration for the public sector and for private enterprises. Cloud providers have already set up sovereign regions to help address these considerations. Sovereign clouds can control how the AI technologies are deployed and operated, including the hardware and software infrastructure used to build and operate the AI technologies, as well as the policies and personnel used to manage the AI technologies and protect the data. Sovereign clouds are powering key concepts like Sovereign AI and Sovereign LLMs.

The importance of Sovereign AI lies in its ability to address data residency. Sovereign AI also promotes innovation within local ecosystems by helping to enable countries to harness the potential of AI technologies while maintaining stronger operational control. This is particularly important for sensitive sectors such as the public sector and national security, where data integrity, security, and governance are paramount. As the world of AI evolves, sovereign clouds and Sovereign AI represent a new flexible choice for customers to maximize foundational and emerging technology where they need it.

AI-driven Agents: a Paradigm Shift in how we Compute

We are witnessing a paradigm shift in how we compute. Large language-based inference software now serves as both the interface and the computing platform, combining software and hardware accelerators. Users, programmers, and application developers interact with these new platforms using natural language. Applications developed on this new paradigm utilize foundational large language models for general interactions and route task-specific prompts to smaller language models trained on domain-specific data. This end-to-end interaction is known as an agentic workflow.

Agentic workflows powered by AI encompass a variety of tasks, such as helping enable automated financial management, where an AI application can be implemented to help scan bank statements, credit card expenses, and other financial data, extract and summarize relevant fields, and help generate an automated budget tracking spreadsheet without manual data entry. AI agents can also be implemented to help assist with targeted web searches, helping users find a new home in a neighborhood that suits their preferences or locate the best tutorials and materials for home projects. Some agentic workflows are also already in use, such as code generation and testing. AI can be implemented to help convert a developer's specifications into code, perform documentation, and handle basic testing, eliminating the need for a developer to manage these tasks, thus streamlining

the software development process by efficiently linking access to information with cognitive insights. The future of technological evolution will bring about a new generation of workflows that not only become more sophisticated but also more integrated across multiple domains. Agents may have the potential to handle more and more complex tasks. This can help improve productivity in our lives.

Retrieval-augmented Fine-tuning (RAFT): Richer Context for Cleaner Text

GenAI is transitioning from retrieval-augmented generation (RAG) to retrieval-augmented fine-tuning (RAFT) for domain-specific tasks which have not met our expectations, promising both better accuracy and minimal hallucination (errors).

RAG provides a way to optimize the output of an LLM with targeted information without modifying the underlying LLM. It allows the LLM to tap into new data without retraining and is used for GenAI applications to provide more context-specific answers to questions. RAFT is an improvement upon RAG because it enables models to learn from external, domain-specific knowledge during fine-tuning. RAFT goes beyond RAG by improving the accuracy of the LLM.

The RAFT process begins with initial fine-tuning of task-specific data. At each iteration, the retrieval system gathers relevant documents for each input example. These documents are then merged with the input data to form an augmented dataset. The model is further fine-tuned on this enhanced dataset, leveraging the additional context for improved performance.

There are positive advantages of RAFT. By incorporating useful external information, RAFT enhances the performance of language models on a variety of tasks. The diverse examples and contexts acquired through external data allow the model to generalize with improved accuracy. Additionally, RAFT can be implemented to efficiently use large external corpora during training without embedding the information into the model parameters.

RAFT can be implemented to be particularly beneficial for helping with knowledge-intensive tasks such as question-answering, summarization, and knowledge-based reasoning. It provides context for accurately answering queries in open-domain question-answering scenarios.

Multi-modal AI: Capabilities that Go Beyond Text

The leading edge of AI research has recently pivoted in another promising direction: multi-modal AI. With the ability to connect different modalities (e.g., text, speech, image, video), multi-modal AI makes AI systems more like humans in their ability to employ multiple sensations to process and understand information.

Multi-modal AI hinges on the complementarity of these forms of data to help build more accurate representations and help generate nuanced content by using the content provided through more than one modality; an AI system can be implemented to help create more accurate representations of information, as well as generate more nuanced content. By using multiple modalities, an AI system can be implemented to help capture a broader cross-section of the world, replicating some of the multi-sensory experiences we have as humans.

Multi-modal AI can be implemented to provide improved advantages over single-modality approaches. Firstly, it combines different modalities, leveraging their complementary strengths. This integration allows for better predictions with higher confidence levels than isolated modalities. For example, object detection can be enhanced by analyzing both the visual and audio feeds of a particular video. Multi-modal AI agents can be implemented to help bridge gaps in human communication by helping to emulate how we naturally exchange both verbal and visual information. It is a step towards advanced artificial general intelligence (AGI).

The are numerous potential applications for multi-modal AI and related use cases. Multi-modal AI can be employed with computer vision for helping with tasks such as image captioning, visual question answering, and video understanding; in natural language processing for sentiment analysis, dialog systems, and machine translation; and in other areas such as robotics, medicine, and entertainment.

The Chance to Maximize Potential

Leaders can consider providing the essential computational resources to help harness AI's potential fully. Innovations like RAFT and multi-modal AI can be implemented to help facilitate more sophisticated and accurate applications, while GenAI-driven tools can be implemented to help with code generation, documentation, and testing which can result in improved productivity.