From RAGS to Riches | Industry First Multimodal RAG on 5th Gen Intel Xeon Processors without GPUs

| July 2024

In this blog, Metrum AI presents a multimodal RAG solution developed on 4th Gen Intel® Xeon® Scalable Processors based Azure Instances and validated on 5th Gen Intel® Xeon® Processors based Azure Instances, using Llama 3 Model & Phi-3 without the need for GPUs.

Industry First Multimodal RAG on 5th Gen Intel Xeon Processors without GPUs

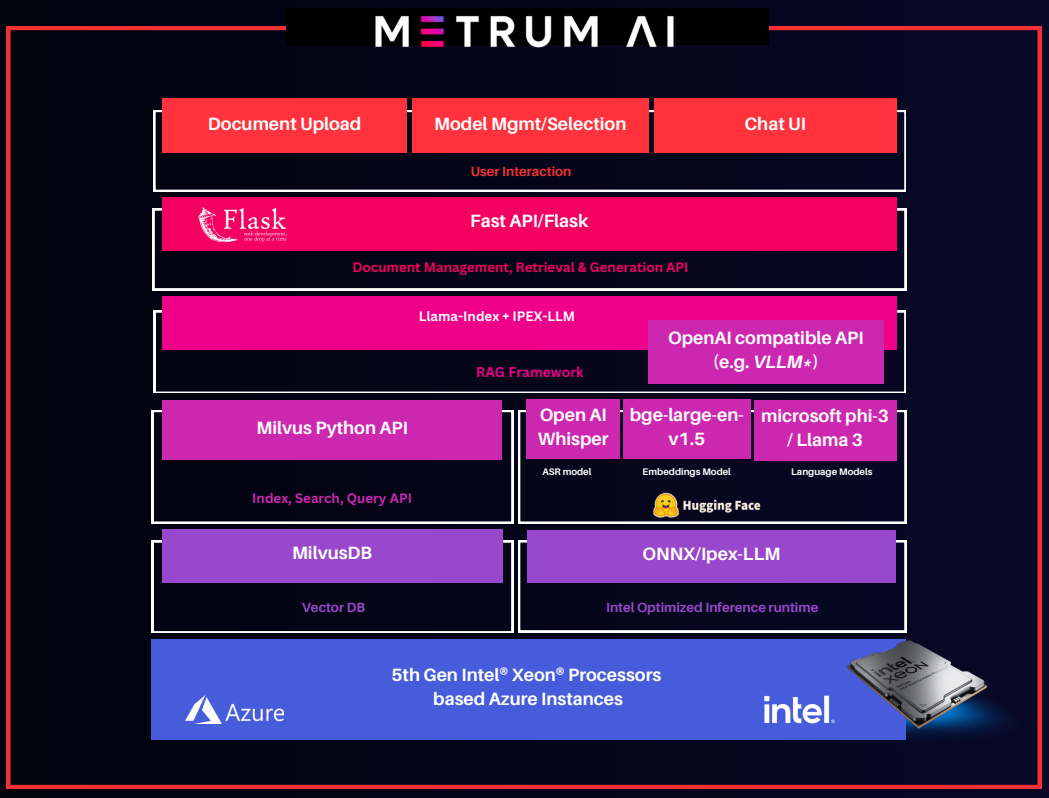
intel METRUM AI Azure

Today, most RAG (retrieval-augmented generation) solutions require GPUs to run generative AI models, leading to increased costs and lead times as well as incompatibility with existing infrastructure. Metrum AI is excited to showcase an industry first multimodal RAG solution that runs solely on 5th Gen Intel Xeon Processor based Azure Instances, without the need for GPUs, along with Intel DL Boost powered by Intel AMX acceleration. 5th Gen Intel Xeon Processors are engineered to seamlessly handle demanding AI workloads, including inference and fine-tuning on models containing up to 20 billion parameters, without an immediate need for additional hardware. In this blog, Metrum AI walks through an Earnings Webcast scenario, provides insights into the solution architecture and user interface, and uncovers the following critical value drivers:

- Developed multimodal RAG on **5th Gen Intel Xeon Processor based Azure Instances with Intel DL Boost powered by Intel AMX acceleration.**
- Deployed four models (language, image embeddings, text embeddings, and voice) on **5th Gen Intel Xeon Processor based Azure Instances, without the need for GPUs.**
- Showcased live at MS build ‘24 in an earnings call application demonstrating the use of language and voice for rapid analyst insights.

Solution Architecture

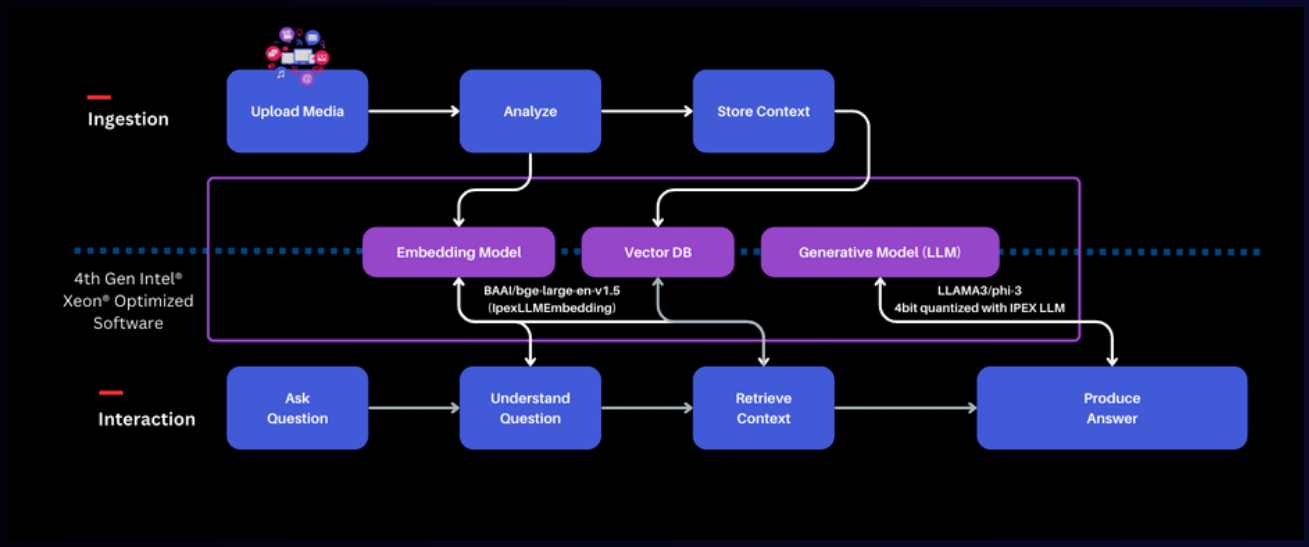
The software stack includes a suite of models, including OpenAI Whisper, an automatic speech recognition system (ASR), bge-large-en-v1.5, a text embedding model, and Llama 3 model with vLLM. It also utilizes LlamaIndex with IPEX-LLM as a RAG Framework and MilvusDB as a vector database, along with Fast API and Flask to interface with user interactions including input file uploads, model selection, and using the chat console.



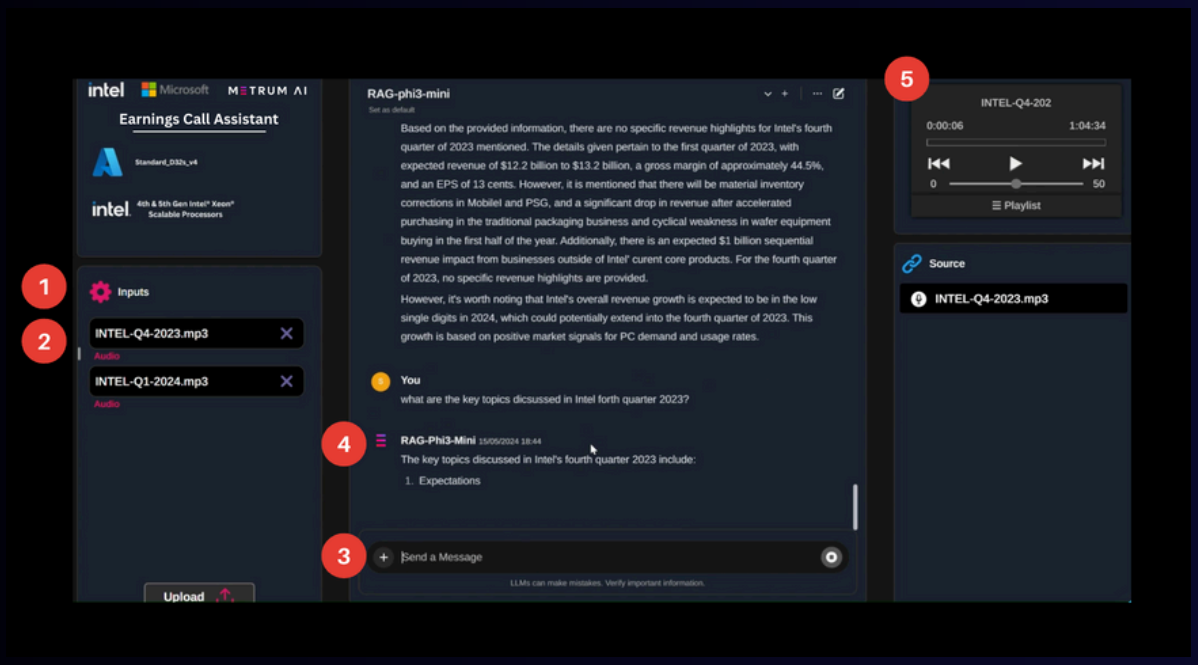
Solution Workflow

Earnings webcasts are a critical tool for publicly traded companies to communicate with their investors, analysts, and the broader market, and typically involve large amounts of data collected from the company's knowledge base. RAG solutions enhance the accuracy, efficiency, and effectiveness of the Earnings webcast process by extracting insights from company proprietary data, providing significant benefits to enterprises in terms of communication, decision-making, and stakeholder engagement. The image provided below illustrates the solution workflow, which details how users can begin by uploading relevant input media files and query the application, which will then compile relevant information and provide references to the uploaded files from which the answers were extracted.

See diagram on next page.



The user interface shown below allows the user to monitor file uploads, query the application, monitor the conversation console, and refer to audio or video snippets relevant to user queries. The following steps expand on the solution workflow, with each step marked on the user interface.



1. User ingests input audio/text through file.
2. Solution creates text embedding from input video frames, audio snippets, or documents.
3. User converses with the application through the conversation console to ask questions about information presented in the input file.
4. Solution generates text responses on the conversation console.
5. Solution retrieves a relevant file snippet from the list of input files. This provides the user with reference content from which the text response was generated.

Summary

5th Gen Intel Xeon Processors offer industry leading features and capabilities that support enterprises creating custom Retrieval-Augmented Generation (RAG) solutions using their own proprietary data. In this blog, we showcased how enterprises deploying applied AI can take advantage of RAG capabilities in the context of an Earnings Call Assistant.

Additional Criteria for IT Decision Makers

What is RAG, and why is it critical for enterprises?

RAG, which stands for Retrieval-Augmented Generation, is a method in natural language processing (NLP) that enhances the generation of responses or information by incorporating external knowledge retrieved from a large corpus or database. This approach combines the strengths of retrieval-based models and generative models to provide more accurate, informative, and contextually relevant outputs.

The key advantage of RAG is that it leverages a large amount of external knowledge dynamically, enabling the model to generate responses that are not just based on its training data but also on up-to-date and detailed information from the retrieval phase. This makes RAG particularly useful in applications where factual accuracy and detail are crucial, such as in customer support, academic research, and other domains requiring precise information. Ultimately, RAG provides enterprises with a robust tool for improving the accuracy, relevance, and efficiency of their information systems, leading to better customer service, cost savings, and competitive advantages.

***DISCLAIMER - Performance varies by hardware and software configurations, including testing conditions, system settings, application complexity, the quantity of data, batch sizes, software versions, libraries used, and other factors. The results of performance testing provided are intended for informational purposes only and should not be considered as a guarantee of actual performance.

***DISCLAIMER - Performance varies by hardware and software configurations, including testing conditions, system settings, application complexity, the quantity of data, batch sizes, software versions, libraries used, and other factors. The results of performance testing provided are intended for informational purposes only and should not be considered as a guarantee of actual performance.