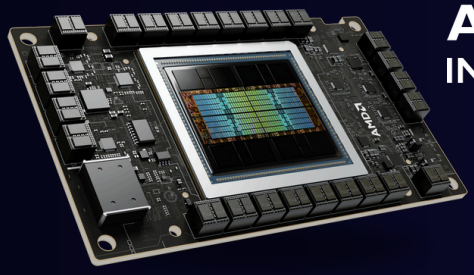
Multimodal RAG-Based Healthcare Assistant on Dell PowerEdge™ XE9680 Rack Server with AMD Instinct™ MI300X Accelerators

In this blog, Metrum AI introduces a healthcare assistant powered by AMD Instinct MI300X accelerators on Dell PowerEdge XE9680 servers. This assistant uses advanced vision-language models to analyze pathology images and generate detailed summaries, improving efficiency in healthcare workflows.

| September 2024



AMD INSTINCT

DELL Technologies AMD METRUM AI Hugging Face

The integration of AMD's Instinct MI300X accelerator into Dell Technologies' cutting-edge PowerEdge XE9680 server, represents a major advancement in diversifying the AI hardware ecosystem. Metrum AI has utilized this powerful hardware combination to develop an innovative Healthcare Assistant, showcasing the transformative potential of generative AI in reducing patient wait times, alleviating staff workloads, and improving overall patient outcomes by combining voice, language, and image data modalities.

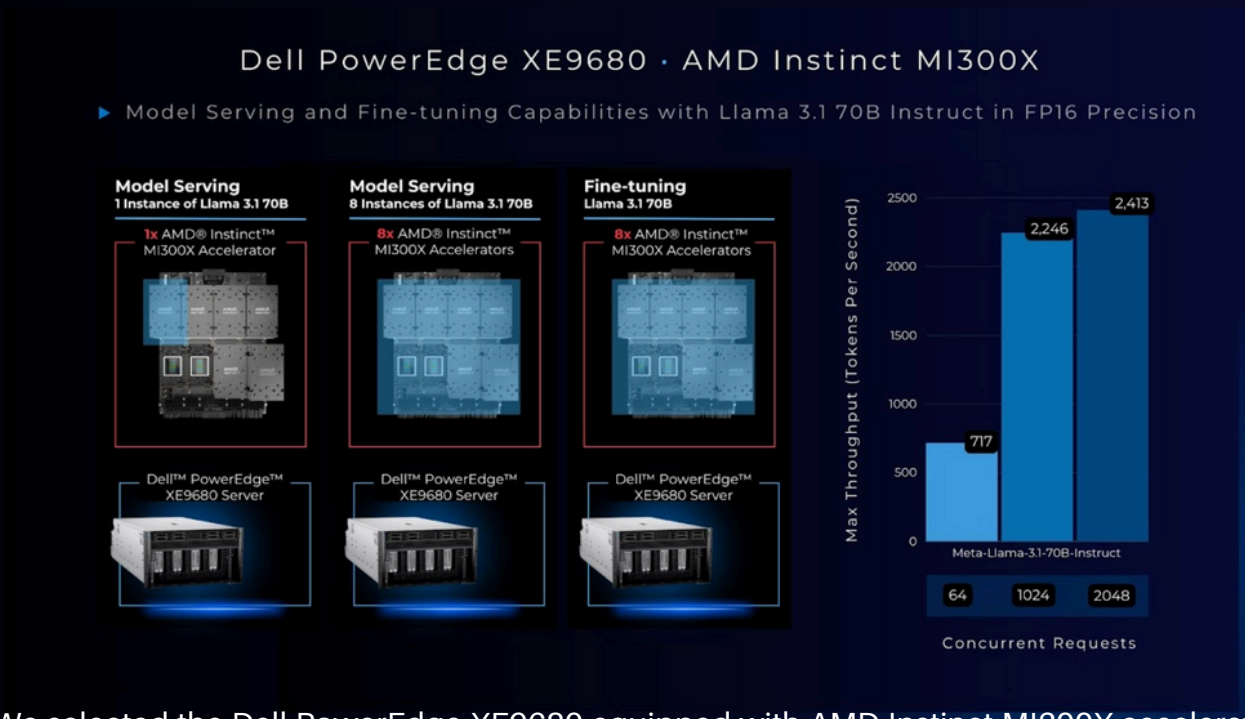
This solution leverages the expanded memory capacity of the MI300X by incorporating a vision-language model for pathology image analysis, retrieval-augmented generation (RAG) for summary creation, and audio session transcription (ASR) to enhance the efficiency and accuracy of clinical documentation, crucial for reducing the administrative overhead on healthcare providers. In this blog, Metrum AI provides insights into the solution architecture developed with industry-leading software and hardware components, and showcases the following:

- How to develop a multi-faceted healthcare solution leveraging AMD Instinct MI300X accelerators with Dell PowerEdge XE9680 server.
- How to deploy four distinct models (language, vision to language, text embeddings, and voice) on a single **Dell PowerEdge XE9680 server with eight AMD Instinct MI300X accelerators.**
- How to navigate an intuitive user interface to demonstrate the seamless integration of language, voice, and vision data, enhancing the process of gathering healthcare insights.

Problem Statement

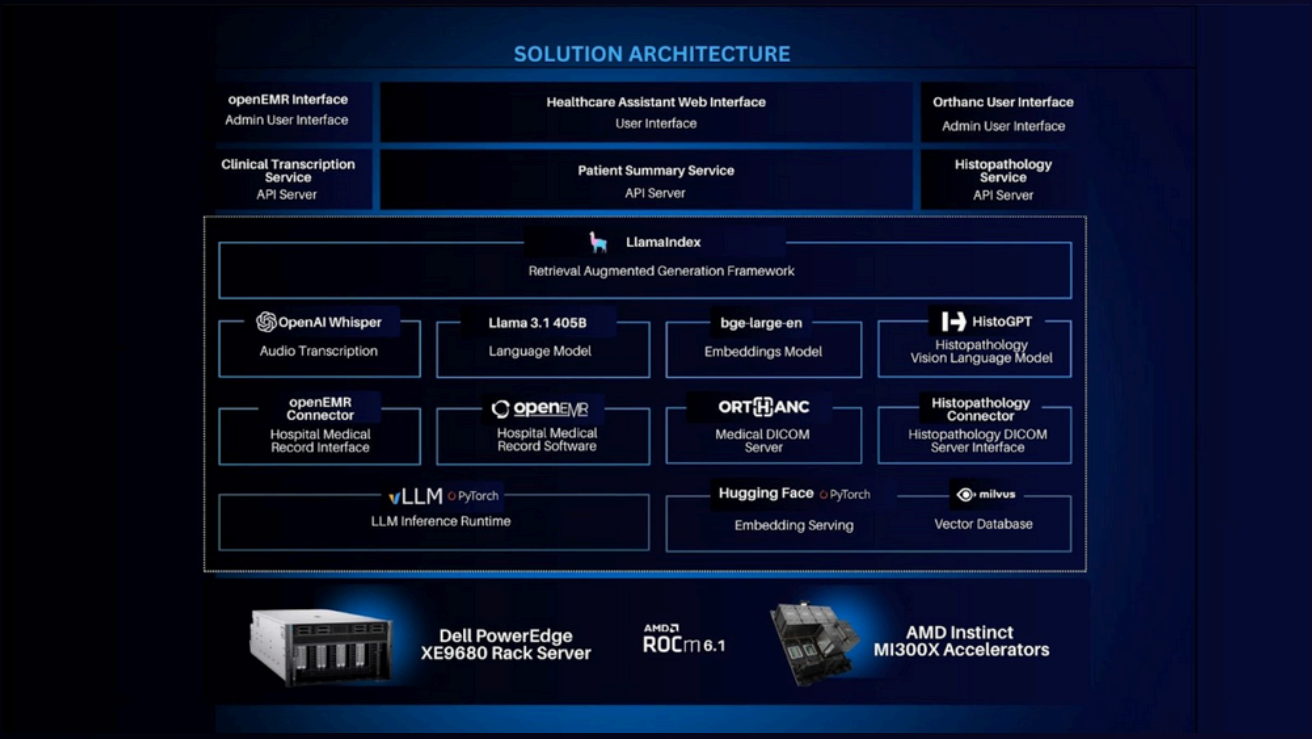
Healthcare professionals are under increasing pressure to deliver high-quality care while managing rising patient demand. Long wait times and administrative tasks reduce healthcare professionals' time with patients. This multimodal RAG-based healthcare assistant addresses these issues by automating time-consuming clinical documentation, such as updating patient records using voice-to-text, analyzing pathology images, and supporting clinical decision-making with AI-assisted analysis and diagnosis. By streamlining documentation and providing quick access to critical insights, this solution allows medical staff to see more patients, reduce wait times, and improve patient outcomes.

Solution Architecture



We selected the Dell PowerEdge XE9680 equipped with AMD Instinct MI300X accelerators for our solution due to its exceptional performance and memory capacity, crucial for handling the latest large language models. With 192GB of HBM3 memory per accelerator, we can comfortably run the entire Llama 3.1 70B model on a single accelerator. Memory and compute-intensive workloads, such as serving multiple model instances and fine-tuning are also possible using only one hardware system with eight accelerators. As shown in the chart above, max token throughput with vLLM model serving of Llama 3.1 70B scales by a factor of 3 with an increase in concurrent requests, achievable due to the unparalleled memory capacity of the AMD Instinct MI300X accelerator.

To deliver an industry specific solution, we paired a large language model with critical software components, such as a vision-language model, voice-to-text model, text embeddings, large language model, and vector database. The memory and performance capabilities of Dell PowerEdge XE9680 with AMD Instinct MI300X accelerators make it possible to support this extensive software stack without compromising accuracy or efficiency.

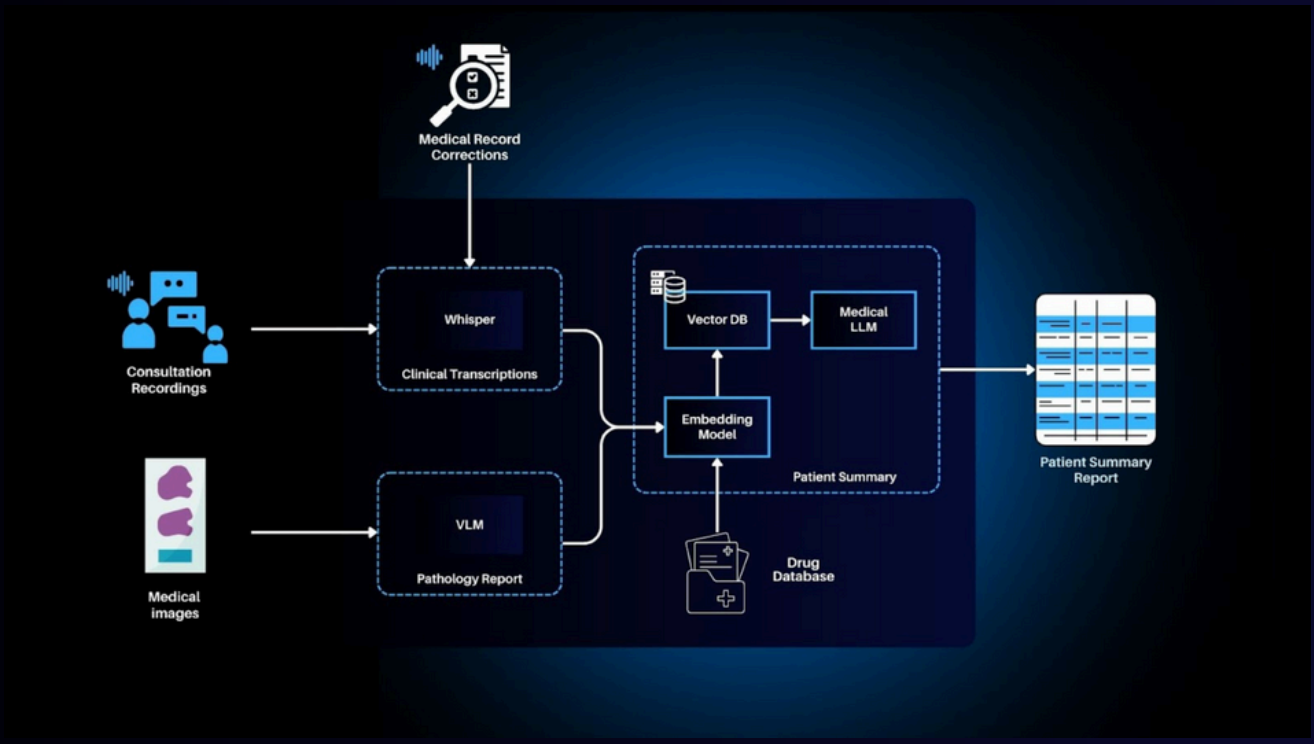


The software stack includes the following key components:

- **HistoGPT Vision Language Model (VLM),** a vision language model trained with histopathology whole slide images to generate detailed disease information.
- **Orthanc**, a lightweight, open-source DICOM server designed for medical imaging. It provides a user-friendly web interface and robust DICOM functionality, allowing healthcare institutions to store, retrieve, and manage medical images efficiently.
- **OpenEMR**, a popular open-source electronic health record (EHR) and medical practice management solution. It provides a comprehensive suite of tools for managing patient records, scheduling, billing, and clinical operations.
- **OpenAI Whisper Transcription Model,** an open-source large multilingual speech recognition model.
- **gte-large embeddings model**, one of the top ranked text embeddings models available on Hugging Face.
- **Llama 3.1 405B Model**, the world's largest open-weight language model with 405 billion parameters, served using vLLM with AMD ROCm optimizations.

- **LlamaIndex**, a popular open-source retrieval augmented generation framework.
- **MilvusDB**, an open-source vector database with high performance embedding and similarity search.
- **vLLM** (**v0.5.3.post1**), an industry-standard library for optimized open-source large language model (LLM) serving, with support for AMD ROCm 6.1.

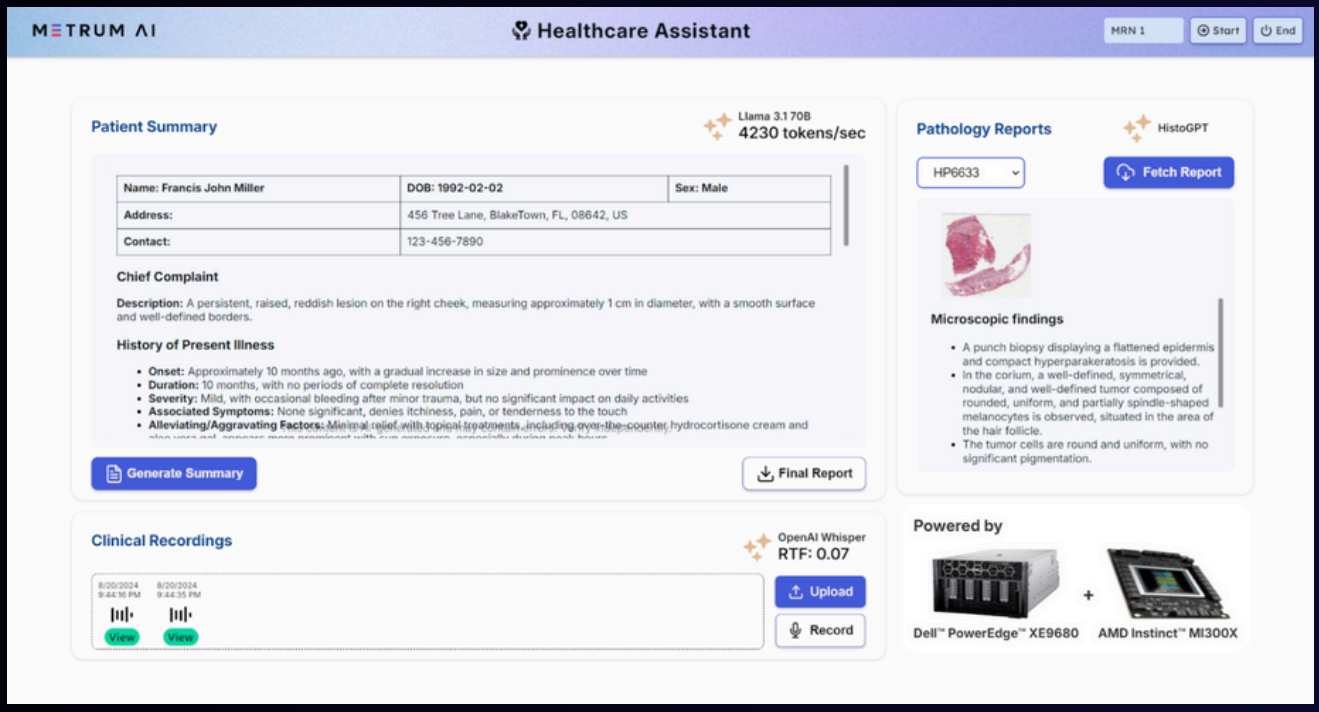
Step-by-Step Walkthrough



The flow diagram illustrates how the system processes multiple data modalities, from embeddings generation to AI model interactions with vector databases. Key features of the user interface include:

- Clinical Recordings: Upload and manage clinical audio files for transcription.
- Pathology Reports: Retrieve detailed AI-generated reports based on pathology images.
- Patient Summaries: Automatically generate and export comprehensive patient summaries, integrating transcribed notes and pathology reports.

The interface is designed for seamless interaction, enabling healthcare professionals to upload recordings, view transcriptions, and generate summaries efficiently.



In this particular solution, we showcase a dermatology use case inclusive of pathology. With more than 9500 patients diagnosed with skin cancer every day in the United States, dermatologists could leverage this solution to streamline the process of identifying and addressing the ailment with simplified note transcribing, assisted tumor pathology image analysis, and support for rapidly generating appointment summaries. This solution can be extended to a variety of other medical use cases involving voice and image data modalities. The following user guide outlines the steps for using the Healthcare Assistant Interface:

- **Select a Patient:**
  - From the Dropdown At the top of the dashboard, select a patient *MRN 1* from the dropdown list.
  - Start Session: Click on the *Start* button to initiate the session for the selected patient.
- **Upload a Recording**
  - Upload Audio: Use the *Upload* button in the Clinical Recording Panel to upload a recording for transcription and analysis.
- **View Transcription**
  - Transcription Completion: Once the transcription is completed, a *View* button will appear next to the audio note.
  - View Transcription: Click *View* to read the transcription of the recording.

- **Generate Summary**
  - Generate Summary: Click on the *Generate Summary* button to create a comprehensive summary of the patient's medical history, conditions, and current diagnosis.
  - Ongoing Updates: You can keep uploading more recordings and generate new summaries as the patient's data evolves.
- **View Histopathology Reports**
  - Select a Test: On the right-hand side of the dashboard, select a test from the dropdown menu.
  - Fetch Report: Click on *Fetch Report* to view the corresponding histopathology report.
- **Final Report**
  - View and Save: After generating summaries and reviewing reports, click on *Final Report* to view the comprehensive report. You can also save this report for future reference.
- **End Session**
  - End Session: Click *End* to close the session. You can start a new session with a different patient or select another patient to start a new session without saving the current data.

Summary

Healthcare providers can now harness Gen AI to integrate and leverage various data sources, including voice, images, and text, to scale administrative tasks and improve patient outcomes, while maintaining the privacy of their proprietary data and workflows. Dell's flagship PowerEdge XE9680 server featuring eight AMD Instinct MI300X accelerators provides the memory capacity needed to support these rich multimodal data and model-intensive use cases.

In this blog, we demonstrated how enterprises deploying applied AI can leverage their proprietary data to benefit from multimodal RAG capabilities in the context of a healthcare assistant solution. We also explored the capabilities of the Dell PowerEdge XE9680 server equipped with AMD Instinct MI300X accelerators, achieving the following milestones:

- Developed a comprehensive healthcare solution using AMD Instinct MI300X accelerators on the Dell PowerEdge XE9680 server.
- Successfully deployed four different models (language, vision to language, text embeddings and voice) on a single Dell PowerEdge XE9680 server with eight AMD Instinct MI300X accelerators.
- Created an intuitive user interface to demonstrate the integration of language, voice, and vision data in streamlining the collection of healthcare insights.

To learn more, please request access to our reference code by contacting us at contact@metrum.ai.

Additional Criteria for IT Decision Makers

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

Retrieval-Augmented Generation (RAG), 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 deliver more accurate, informative, and contextually relevant outputs.

The key advantage of RAG is its ability to dynamically leverage a large amount of external knowledge, allowing the model to generate responses that are informed not only based on its training data but also by up-to-date and detailed information from the retrieval phase. This makes RAG particularly valuable in applications where factual accuracy and comprehensive details are essential, such as in customer support, academic research, and other fields that require precise information.

Ultimately, RAG provides enterprises with a powerful tool for improving the accuracy, relevance, and efficiency of their information systems, leading to better customer service, cost savings, and competitive advantages.

Why is the Dell PowerEdge XE9680 Server with AMD Instinct MI300X Accelerators well-suited for RAG Solutions?

Designed especially for AI tasks, the Dell PowerEdge XE9680 server is a powerful data-processing server equipped with eight AMD Instinct MI300X accelerators, making it well-suited for AI-workloads, especially for those involving training, fine-tuning, and conducting inference with Large Language Models (LLMs).

Effectively implementing Retrieval-Augmented Generation (RAG) solutions requires a robust hardware infrastructure that can handle both the retrieval and generation components. Key hardware features for RAG solutions include high-performance accelerator units and large RAM and storage capacity. With 192 GB of GPU memory, a single AMD Instinct MI300X accelerator can host an entire Llama 3 70B parameter model for inference. Optimized for generative AI, the AMD Instinct MI300X accelerator can deliver up to 10.4 Petaflops of performance (BF16/FP16), and provides 1.5TB of total HBM3 memory in a group of eight accelerators.

Resources

https://www.skincancer.org/skin-cancer-information/skin-cancer-facts/

AMD images: AMD Library, https://library.amd.com/account/dashboard/

Dell images: Dell.com

***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.