AMPERE®

Llama 2 Serge-Chat on GCP C3A VMs

GCP C3A VMs, powered by Ampere® One CPU and high performance Ampere® AI inference engine, deliver best-in-class GPU-Free AI inference performance on standard AI frameworks, including PyTorch, TensorFlow, and ONNX-RT.

Ampere Altra Powered Generative AI Inference

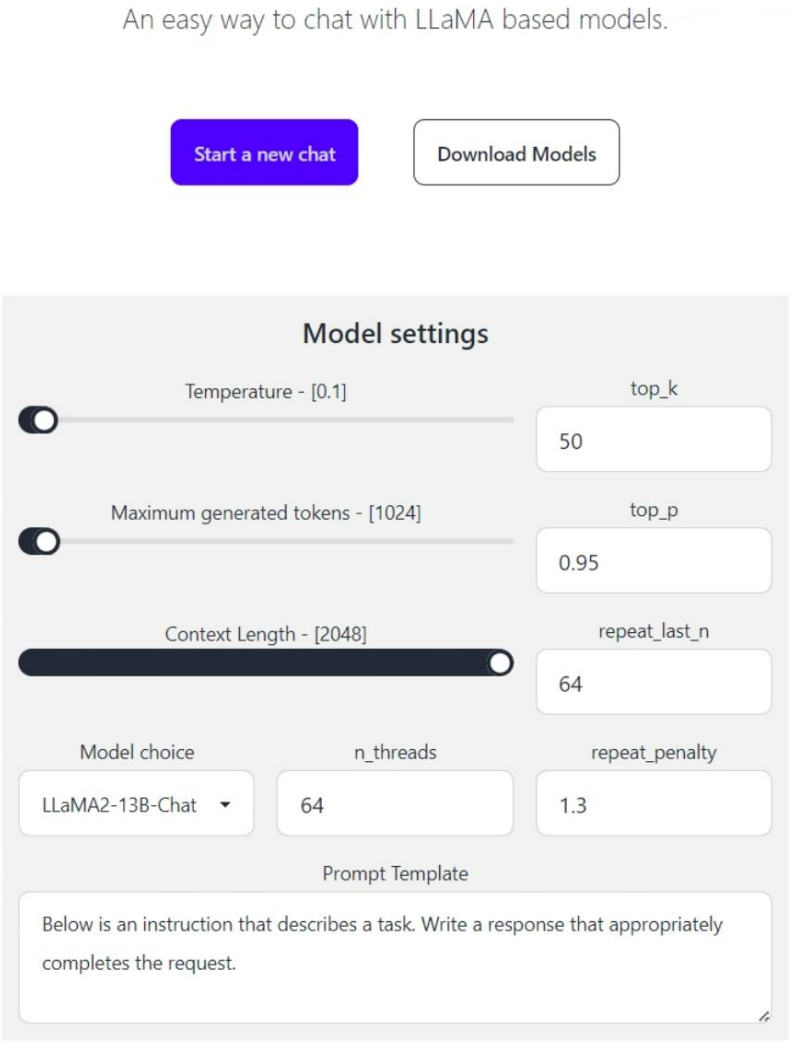
Ampere® **Cloud Native Processors** satisfy the performance requirements of widely used generative AI models such as Meta's Llama 2, especially for the smaller versions of the Llama model. Llama 2 is arguably the most popular open-source LLM (large language model) at the moment. Ampere CPUs provide the **best price-performance and optimized power draw** for all kinds of ML deployments.

Setup

Deployment of the fine-tuned version of the open-source **Llama 2** model running on Ampere-based **GCP C3A** cloud instance with **Ampere® Optimized AI Frameworks**. The demo showcases real-time chatbot interaction on internally developed Ampere chatbot called Serge. The chatbot supports a variety of models, e.g., Mistral, and allows for testing with different input or output token length and a chosen number of threads among other currently supported parameters.

Key Benefits Demonstrated

- Meets or exceeds the necessary **low latency** requirements for real-time generative AI chatbot interaction.
- Delivers the best **price-performance** in LLM inference in cloud deployment scenarios.
- Provides the token generation rate needed for quality end-user experience.
- Ampere Cloud Native Processors enable **easy scaling** and can be **dynamically provisioned** based on the performance requirements of the user's applications.



Real-time Voice-to-Text Chat

This demo shows a **generative AI chatbot deployment**. It processes incoming real-time input from in the form of a user prompt and generates the output per user's instructions. The demo runs at a **real-time performance level** with sufficient **low latency** and per second token generation rate to satisfy the user needs.

Resources

Ampere Optimized Pytorch, and other Ampere Optimized AI Frameworks, can be accessed directly from Google Cloud Marketplace.

The docker images are also available in the downloads section of Ampere AI Solutions web page. All software is available free of charge and runs straight out-of-the-box with no additional coding required.

Figure 2: The integration of Ampere Optimized AI Frameworks with Ampere Altra Cloud Native Processors