



# Voice to Text Chat with Lampi Fine-Tuned Llama 2 7B and Whisper Models

Ampere<sup>®</sup> Cloud Native Processors with Ampere<sup>®</sup> Optimized Al Frameworks, deliver best GPU-Free Al inference performance for applications developed in PyTorch, TensorFlow, and ONNX-RT.

# Ampere Altra Powered Generative Al Inference

Ampere **Cloud Native Processors** satisfy the performance requirements of widely used generative AI models such as Whisper, originally developed by Open AI and used especially for ASR (automatic speech recognition), and Meta's Llama 2, especially for the smaller versions of the Llama model. Ampere CPUs **provide the best price-performance and optimize power draw for all kinds of ML deployments**.

## About Lampi

Lampi provides AI workflow copilot oriented to business use cases agnostic and fully modular adapted on the fly to client's business needs. Lampi deploys multimodal AI on the fly allowing for a consistency in results to reach the desired quality and compliance levels.

#### Setup

Deployment of the dynamic Lampi fine-tuned versions of the opensource generative AI Llama 2 7B and Whisper models running on Ampere - based Scaleway COP-ARM cloud instance with **Ampere® Optimized AI Frameworks**. The demo utilizes 32 cores with assigned memory of 128 GB for real-time voice-to-text chatbot interaction with the training set of Ampere AI and energy efficiency content chosen to inform the audience visiting the Ampere Cloudfest booth.

## Key Benefits Demonstrated

- Meets or exceeds the necessary low latency requirements for real-time generative AI chatbot interaction. (voice-to-text; also voiceto-voice not presented at the event due to technical limitations posed by the show floor environment)
- Ampere delivers best throughput for Whisper model inference outcompeting some of the most commonly used GPUs.
- Delivers the best **price-performance** in GPU-Free AI inference in both cloud and edge deployment scenarios.
- Ampere Optimized AI Frameworks work right out-of-the-box and can be downloaded for free from the Ampere AI solutions page.
- Ampere Altra<sup>®</sup> Family of Cloud Native Processor enables 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 partner solution of a generative AI voice-to-chat deployment. It incoming real-time input from a dedicated external microphone available to the visitors of the Ampere Cloudfest booth. The demo runs at a **real-time performance level** with latency low enough and a per second token generation rate high enough to satisfy the user needs. Lampi also developed a voice-to-voice demo for display, which might be displayed in the future pending a relevant setting.

#### Resources

The docker image of Ampere Optimized PyTorch is available in the downloads section of Ampere AI Solutions web page. Other Ampere® Optimized Frameworks can also be accessed from the same location.

Ampere Optimized TensorFlow, PyTorch, ONNX-RT can also be downloaded and installed free of charge on any edge workstation or server through Ampere AI Solutions web page.

#### Figure 2. Integration of Ampere Optimized Frameworks with Ampere Cloud Native Processors



Ampere Computing reserves the right to make changes to its products, its datasheets, or related documentation, without notice and warrants its products solely pursuant to its terms and conditions of sale, only to substantially comply with the latest available datasheet.

Ampere, Ampere Computing, the Ampere Computing and 'A' logos, and Altra are registered trademarks of Ampere Computing.

Arm is a registered trademark of Arm Limited (or its subsidiaries) in the US and/or elsewhere. All other trademarks are the property of their respective holders. Copyright © 2024 Ampere Computing. All Rights Reserved.

#### Ampere Computing® / 4655 Great America Parkway, Suite 601 / Santa Clara, CA 95054 / www.amperecomputing.com