voicechat2: Local AI Voice Chat

AMD Pervasive AI Developer Contest
(Finalist Demo)

Leonard Lin
https://leonardlin.com/

https://github.com/lhl/voicechat2

About Me

- Grew up on the PC/early-online era (BBS's, Usenet, etc)
- Technologist and Developer
  - https://linkedin.com/in/randomfoo
- Currently based in Tokyo
- Training Multilingual and Multimodal LLMs (Shisa.AI)
- See also, recent TokyoAI presentation:
  - Lessons Learned Training Open Source Japanese LLMs

Why voicechat2?

- Last year I built a web-based voice-to-voice conversational AI system for fun
- This turned into a backend system system for a 3D virtual avatar client
- I wanted to revisit it at some point, but got busy with other projects
- Clean-room, open source re-write for this Hackathon
  - WebRTC or Websockets
  - Compressed audio (Opus)
  - Incorporate interleaving tricks
  - Fully open source

Why voicechat2? (continued)

- Modular - can plug in any SRT, LLM, TTS framework and model
  - whisper.cpp, WhisperX, FasterWhisper
  - llama.cpp, ExLlamaV2, HF Transformers, vLLM
  - xTTS, StyleTTS, Melo, Piper
- With interleaving, can be as fast as native voice-decoder/encoder models
- Dev project, but documented, Apache 2.0 licensed

W7900 / RDNA3

Compatibility testing

- https://llm-tracker.info/W7900-Pervasive-Computing-Project
- https://llm-tracker.info/howto/AMD-GPUs
- https://www.reddit.com/r/LocalLLaMA/comments/191srof/amd_radeon_7900_xtxtx_inference_performance/
- https://www.reddit.com/r/LocalLLaMA/comments/1atvxu2/current_state_of_training_on_amd_radeon_7900_xtx/
- https://wandb.ai/augmxnt/train-bench/reports/Trainer-performance-comparison-torchtune-vs-axolotl-vs-Unsloth---Vmlldzo4MzU3NTAx
  - https://github.com/AUGMXNT/speed-benchmarking/tree/main/train-bench

W7900 / RDNA3

PROJECT	STATUS
PyTorch	GOOD
Triton	OK
Flash Attention	BAD
bitsandbytes	GOOD
xformers	BROKEN

W7900 / RDNA3

PROJECT	STATUS
llama.cpp	GOOD
ExLlamaV2	OK
MLC	GOOD
vLLM	OK (no FA)

W7900 / RDNA3

PROJECT	STATUS
Whisper	GOOD
whisper.cpp	GOOD
FasterWhisper	BROKEN
WhisperX	BROKEN
StyleTTS2	GOOD
xTTS	GOOD

W7900 / RDNA3

PROJECT	STATUS
torchtune	GOOD
axolotl	OK
Unsloth	BROKEN
accelerate	BAD
DeepSpeed	BAD

DEMO

- https://github.com/lhl/voicechat2
- https://www.hackster.io/lhl/voicechat2-local-ai-voice-chat-4c48f2
- https://www.reddit.com/r/LocalLLaMA/comments/1eju211/voicechat2_an_open_source_fast_fully_local_ai/
- https://ai.shisa.ai/voicechat2/ (limited time devbox tunnel demo)