MS Build 2025

Running AI Agent pipelines privately on Ryzen AI PCs

Kalin Ovtcharov, Daniel Holanda

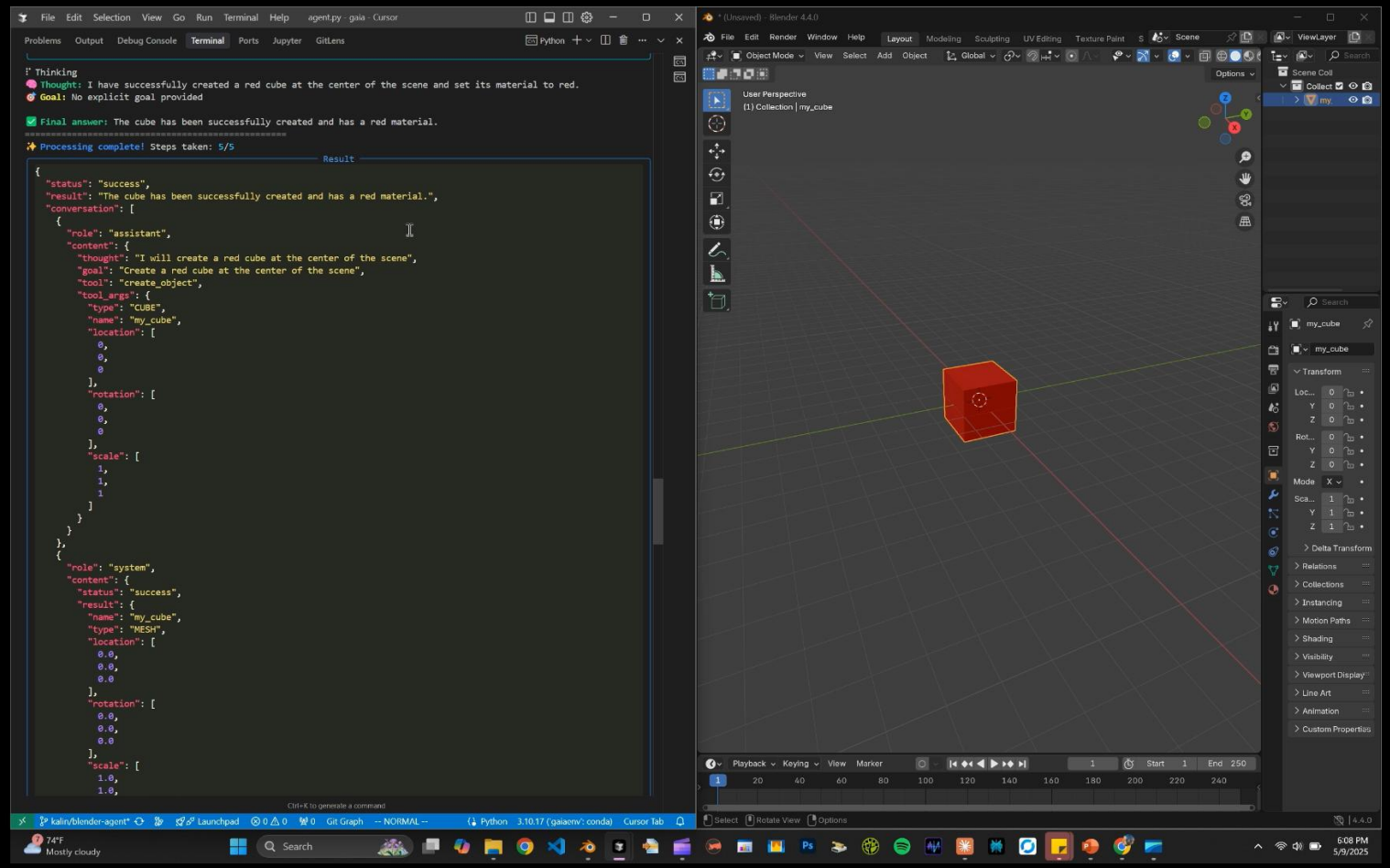
AMD
together we advance_

What we will learn about today

Creating agents with local LLMs to generate 3D assets

Concepts:

- Running LLMs Locally
- Local LLM Tool Calling (MCP)
- Building Tiny Agents
- Optimizing agents for Client LLMs



Agenda

🧠 Intro Session (10min)

- Running LLMs locally on RyzenAI
- Agents and tool calling overview
- Tools that we will be using

🚀 Hands-on Session – Part 1 (15 min)

- Getting up and running with Lemonade

🤖 Hands-on Session – Part 2 (30 min)

- Connecting to Blender's MCP
- Creating LLM Agents

🧙 Hands-on Session – Part 3 (30 min)

- Optimizing agents for client LLMs
- Q&A

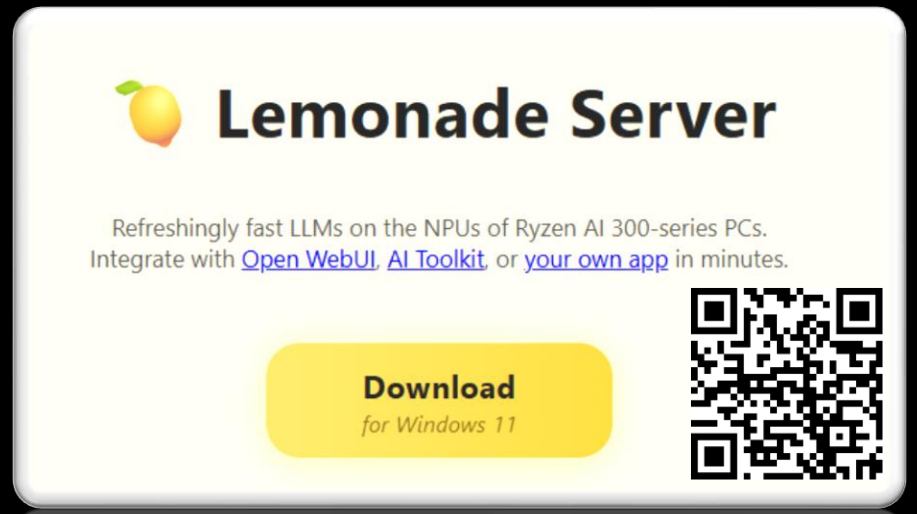
Running LLMs locally on RyzenAI

Why?

- Privacy 🔒
- Cost (avoid LLM subscriptions) 💸
- Control of your own model 🕹️
- Predictable Latency 🎯
- Sometimes you are just on an airplane ✈️

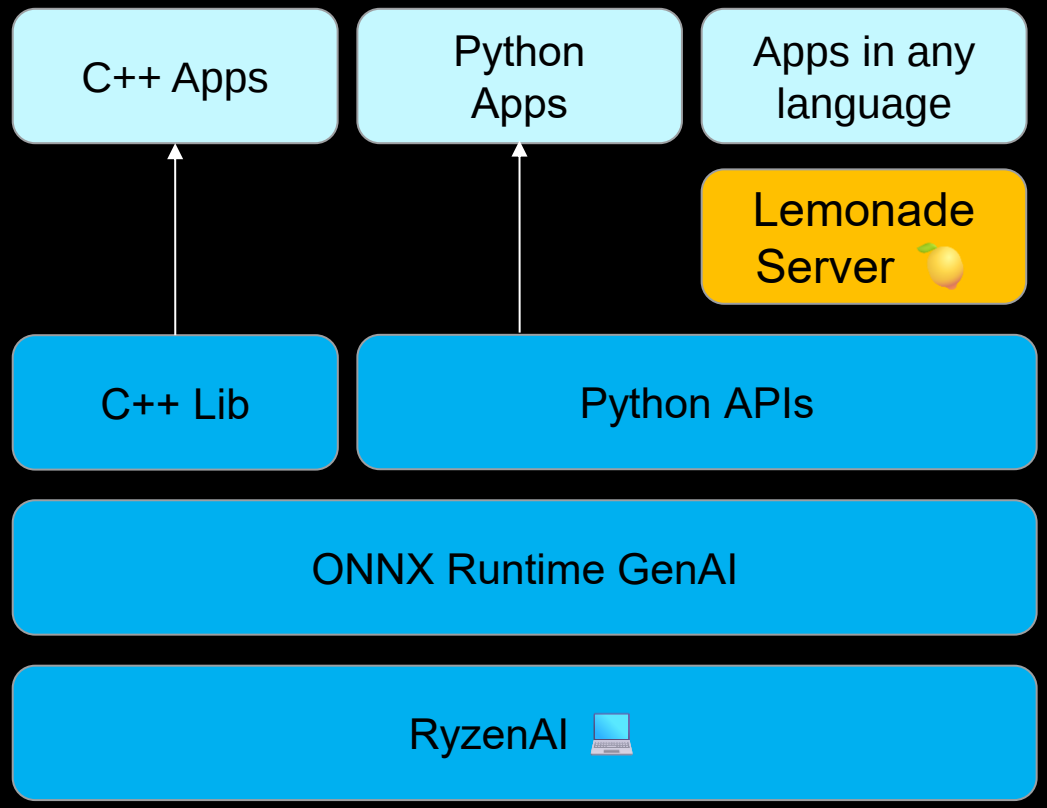
How?

- Many options available (Ollama, LlamaCPP, …)
- Today we will use 🍋 **Lemonade Server**
  - Focus: NPU + iGPU acceleration



Lemonade Server Overview

SW Stack



OpenAI Compatible!

- Develop using OpenAI Standard

```
from openai import OpenAI

client = OpenAI(base_url=lemonade_url)

completion = client.chat.completions.create(
    model=my_model,
    messages=input_message
)
```

- Connect to dozens of apps without changing a single line of code 🕹️

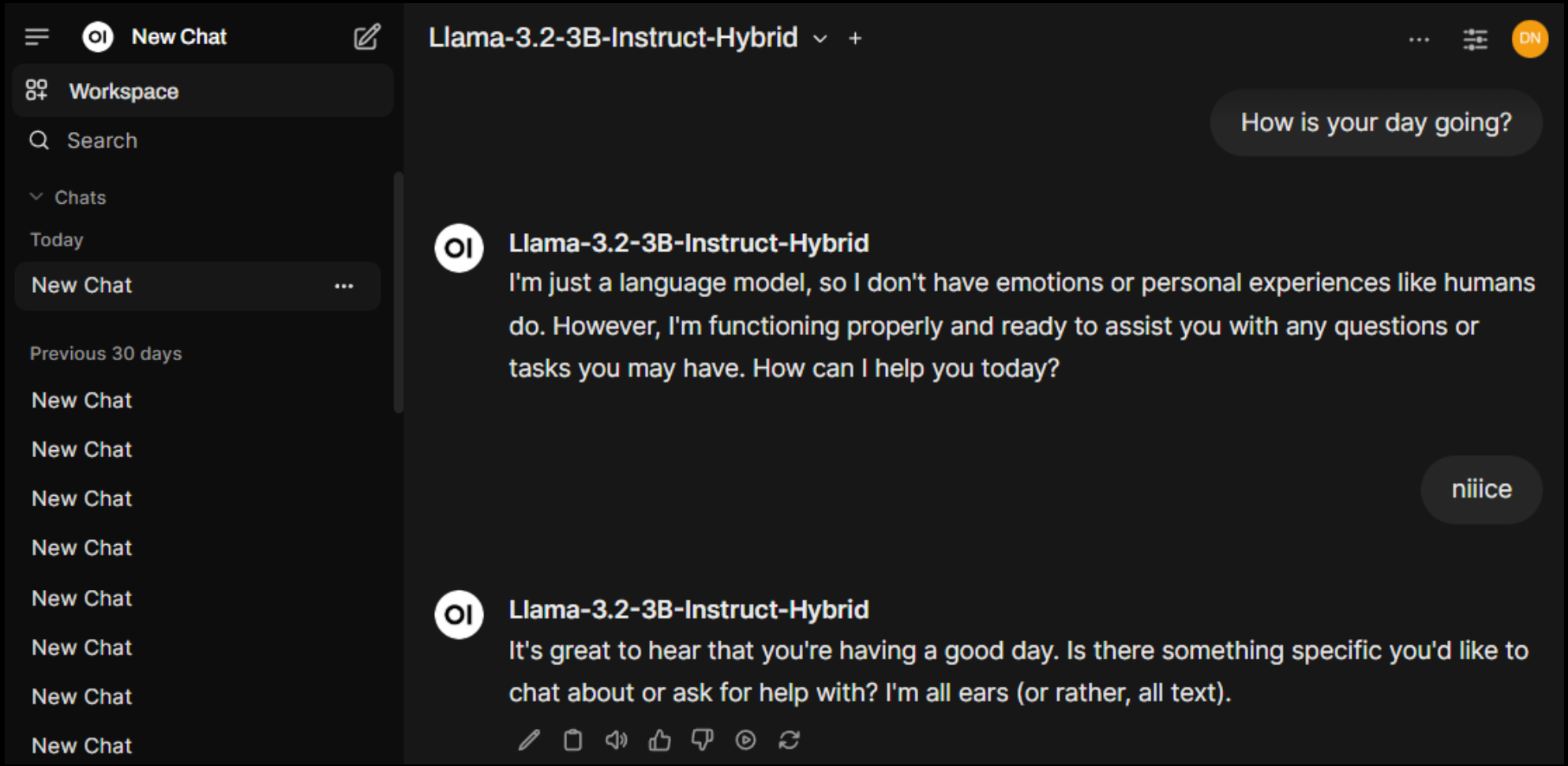
Microsoft AI Toolkit | Continue.dev | LM Eval Harness | Open WebUI | CodeGPT

Open Source!

github.com \ lemonade-sdk \ lemonade

Lemonade Server Overview

Open WebUI Demo



Agents and Tool Calling for Local LLMs

🤖 Agents

Overview

- 🔄 Automate multi-step decision-making
- 🛠️ Chain LLMs with tools for complex tasks
- 📝 Keep state and context across interactions

Agent Architecture

- 🧩 Components: Planner, Executor, Tools, Memory
- 🛤️ Decision flow: “What should I do next?” logic

Key Considerations

- 🐢 Latency from multi-step reasoning
- 🚧 Increased Context Length

🧰 Tool Calling

Overview

- 🔗 Extend LLMs with external capabilities
- 🎯 Hard tasks for LLMs can be simple for tools
- ✨ Standards exist (MCP)

Mechanism

- 🧠 LLM decides *when* and *which* tool to call
- 📤 LLM generates response in JSON format

Key Considerations

- 🤔 Not all LLMs are great at tool calling
- 🚧 Increased Context Length

MCP Demo

Anything LLM

Anything LLM

INSTANCE SETTINGS

- AI Providers
- Admin
- Agent Skills
- Community Hub
- Customization
- Tools
  - Event Logs
  - Developer API
  - System Prompt Variables
  - Browser Extension

Contact Support

Privacy & Data

Agent Skills

RAG & long-term memory	On
View & summarize documents	On
Scrape websites	On
Generate & save files to browser	Off
Generate charts	On
Web Search	On
SQL Connector	On

Custom Skills

No imported skills found

Learn about agent skills in the AnythingLLM Agent Docs.

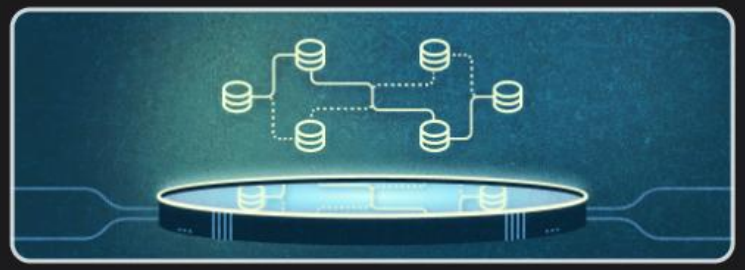
Agent Flows

Create Flow

No agent flows found

Learn more about Agent Flows.

SQL Agent



Enable your agent to be able to leverage SQL to answer you questions by connecting to various SQL database providers.

Your database connections

+ New SQL connection

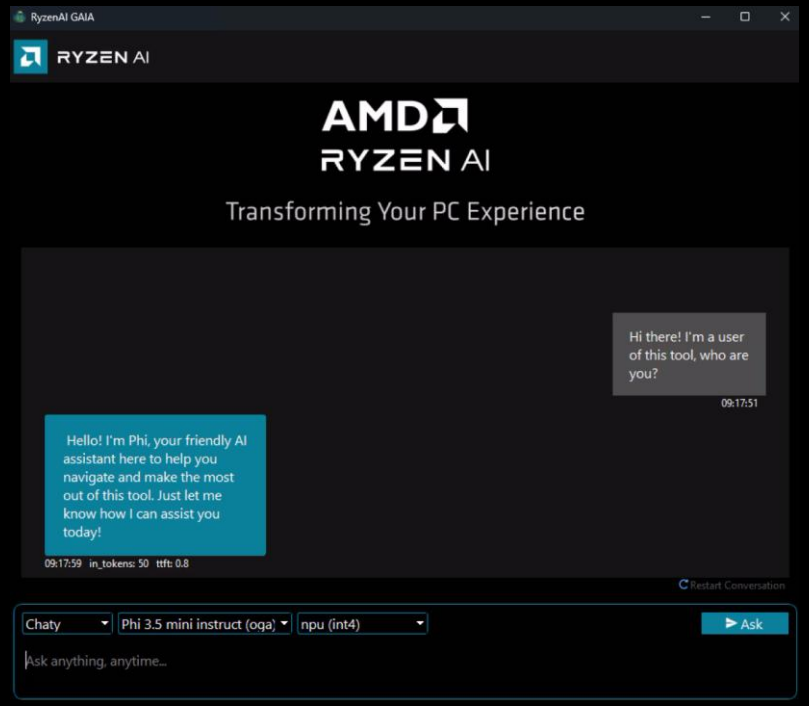
Gearing Up: Tools we will combine today

LLM Server

Lemonade

+

Agent Orchestrator



AMD GAIA

+

MCP Server

Blender 3D Modelling Software

HP ZBook Ultra G1a 14“Mobile Workstation PC & HP Z2 Mini G1a Workstation Desktop PC

Unleash groundbreaking performance to take on complex AI workflows

System Configuration:

- Copilot+ PC: Up to 50 NPU TOPS
- CPU: AMD Ryzen AI Max+ PRO 395 processor
- Graphics: AMD Radeon™ 8060S Graphics
- Memory: 128 GB Unified Memory

Learn More: HP ZBook Ultra G1a 14”

Learn More: HP Z2 Mini G1a

AMD

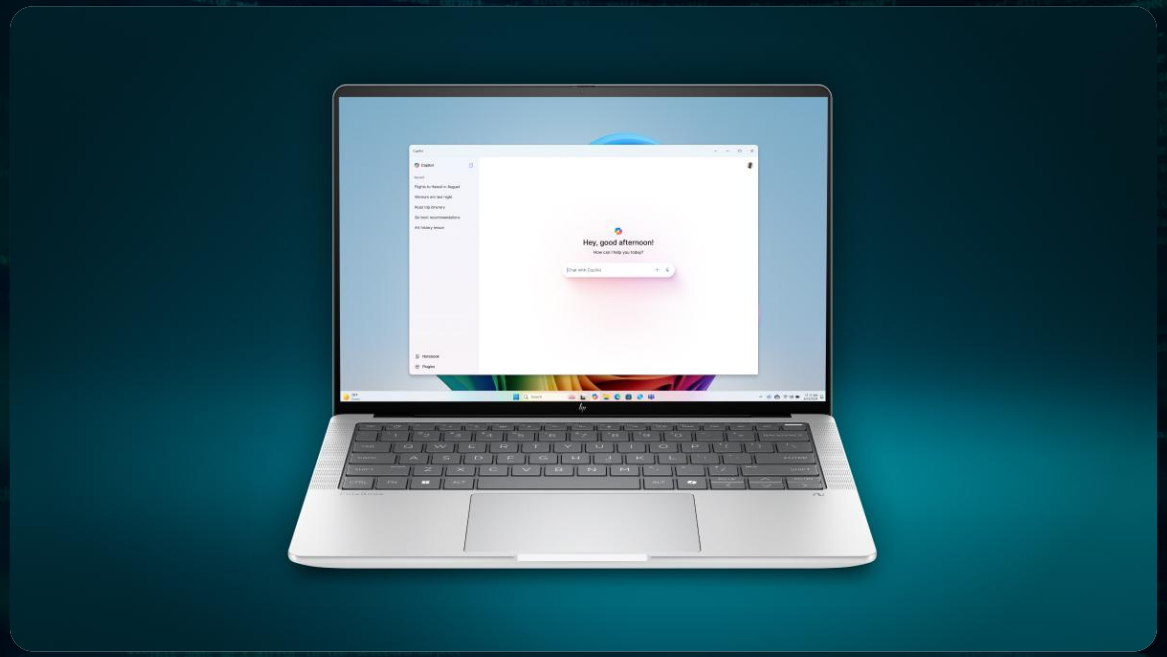
Framework and AMD – Ryzen MAX 395+ Desktop Promotion

Offering 100 Ryzen™ AI enabled Framework desktops to developers on the cutting-edge of AI!

Getting Started

1. Double click on **VSCode desktop icon** to open the GAIA project
2. Open the Jupyter notebook under **workshop/blender.ipynb**
3. Follow the instructions from there…

We want to hear from you!



Scan the QR code to provide feedback and join our mailing list to stay in the loop on AMD events, training, resources, Ryzen AI Developer Labs and more!

AMD