

LM Studio Brings Phenomenal AI Power without the Cloud

cisco Live !

Hank Preston, Distinguished Architect Learn
with Cisco
ccie 38336 | devnet expert 20220001

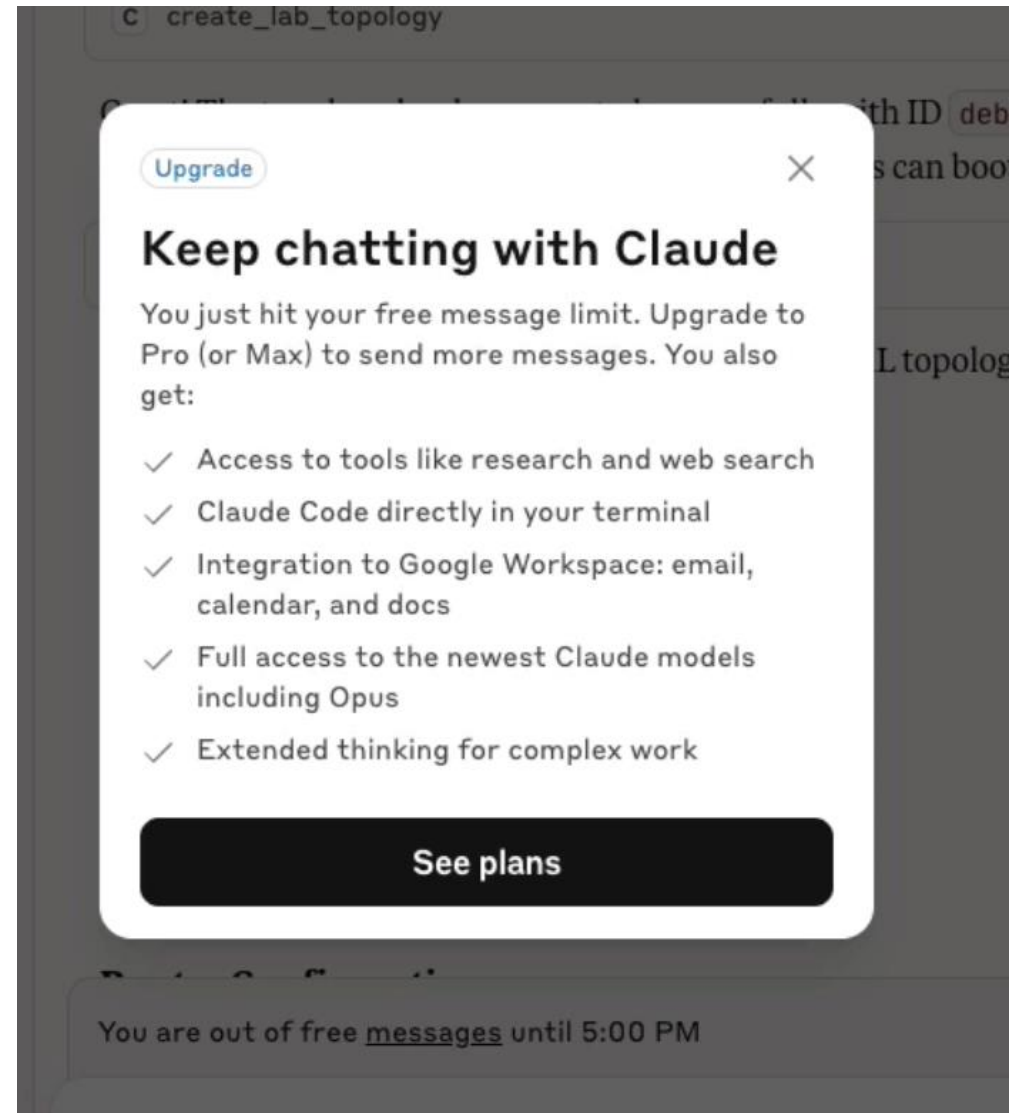
August 2025

Agenda

1. Why consider local LLMs?
2. Introduction to LM Studio
3. Finding and choosing models
4. Demo, Demo, Demo

Why consider local LLMs?

Avoid the limitations of "as a Service" offers





AI Generated Image

Security and Privacy... keep your prompts and responses to yourself

Experimentation and Learning...

You don't need a nuclear powered data center for hello- world.

You are an AI assist - Branched - Branched	6K tokens
You are an AI assist - Branched	5K tokens
You are an AI assist	7K tokens
You are an AI assist - Branched - Branched - Branch...	12K tokens
You are an AI assist - Branched - Branched	9K tokens
You are an AI assist - Branched	12K tokens
You are an AI assist	21K tokens
You are an AI assist	7K tokens
Unnamed Chat	1K tokens
Unnamed Chat	...
SSH Ports from NetBox - Branched	6K tokens
SSH Ports from NetBox	46K tokens
You are an AI assist	109K tokens
You are an AI assist	27K tokens
NetBox Site Overview	13K tokens
Unnamed Chat	0 tokens
CML Lab Status Inquiry	15K tokens
You are an AI assist	49K tokens
You are an AI assist	45K tokens



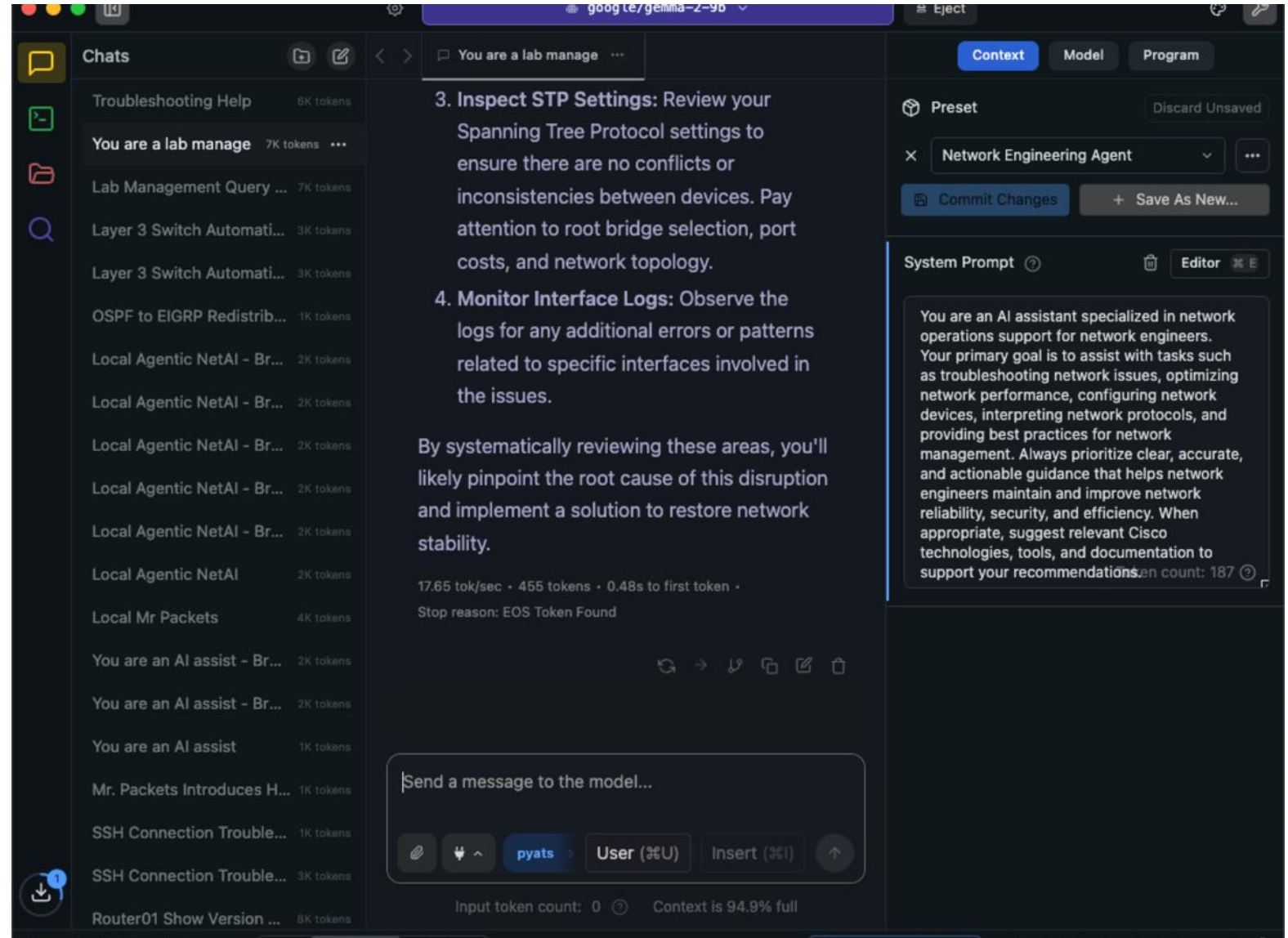
AI Generated Image

**Offline access... sometimes the Internet
just isn't available.**

Introduction to LM Studio

LM Studio Desktop

- Local application to run LLMs
- Easy to use interface
- Easy access to open models
- Experiment with AI nerd knobs and settings
- Build AI Agents with MCP Servers
- APIs and SDKs available for development use cases

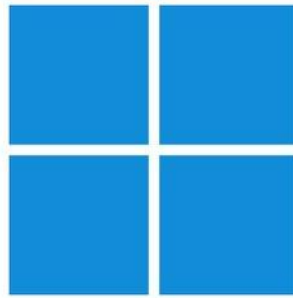


Minimum System Requirements



macOS

- Apple Silicon (M1-4)
- RAM: 16GB+ recommended



Windows

- CPU: x64 and ARM
- RAM: 16GB+ recommended
- GPU: 4GB dedicated VRM



Linux

- Ubuntu 20.04 or newer
- CPU: x64
- Distributed as Appliance

Some Details

- By Element Labs, a small New York City based company
- Free to use for both personal and work
- Main application is not Open Source, but some components are
- LLM prompts, responses, and history stay 100% local
- Does NOT include any LLM model
 - Interface to discover and download models from HuggingFace
- Uses runtimes from OpenAI, Llama, and native
 - CPU/GPU requirements
- Provides an OpenAI compatible API

Finding and choosing models



**When we have too many choices,
the very fact that there's a chance
to get it wrong makes us feel more
uncertain.**

Barry Schwartz, author of “The Paradox of Choice – Why More is Less”

[Main](#)
[Tasks](#)
[Libraries](#)
[Languages](#)
[Licenses](#)
[Other](#)

Tasks

[Text Generation](#)
[Any-to-Any](#)
[Image-Text-to-Text](#)
[Image-to-Text](#)
[Image-to-Image](#)
[Text-to-Image](#)
[Text-to-Video](#)
[Text-to-Speech](#)
[+ 42](#)

Parameters

[<1B](#)
[6B](#)
[12B](#)
[32B](#)
[128B](#)
[>500B](#)

Libraries

[PyTorch](#)
[TensorFlow](#)
[JAX](#)

[Transformers](#)
[Diffusers](#)
[Safetensors](#)

[ONNX](#)
[GGUF](#)
[MLX](#)
[Keras](#)
[+ 41](#)

Apps

[vLLM](#)
[TGI](#)
[llama.cpp](#)
[MLX LM](#)

[LM Studio](#)
[Ollama](#)
[Jan](#)
[+ 12](#)

Models 1,935,482

Filter by name

Full-text search

Sort: Trending

 **openai/gpt-oss-120b**

 Text Generation • 120B • Updated 4 days ago • 490k • 3.24k

 **openai/gpt-oss-20b**

 Text Generation • 22B • Updated 4 days ago • 2.37M • 2.81k

 **Qwen/Qwen-Image**

 Text-to-Image • Updated 6 days ago • 69.7k • 1.49k

 **rednote-hilab/dots.ocr**

 Image-Text-to-Text • 3B • Updated about 7 hours ago • 17.9k • 613

 **KittenML/kitten-tts-nano-0.1**

Updated 7 days ago • 33.7k • 401

 **MiniCPM-V-4**

 Image-Text-to-Text • Updated about 12 hours ago • 32

 **unsloth/gpt-oss-20b-GGUF**

 Text Generation • 21B • Updated 4 days ago • 420k • 276

 **zai-org/GLM-4.5V**

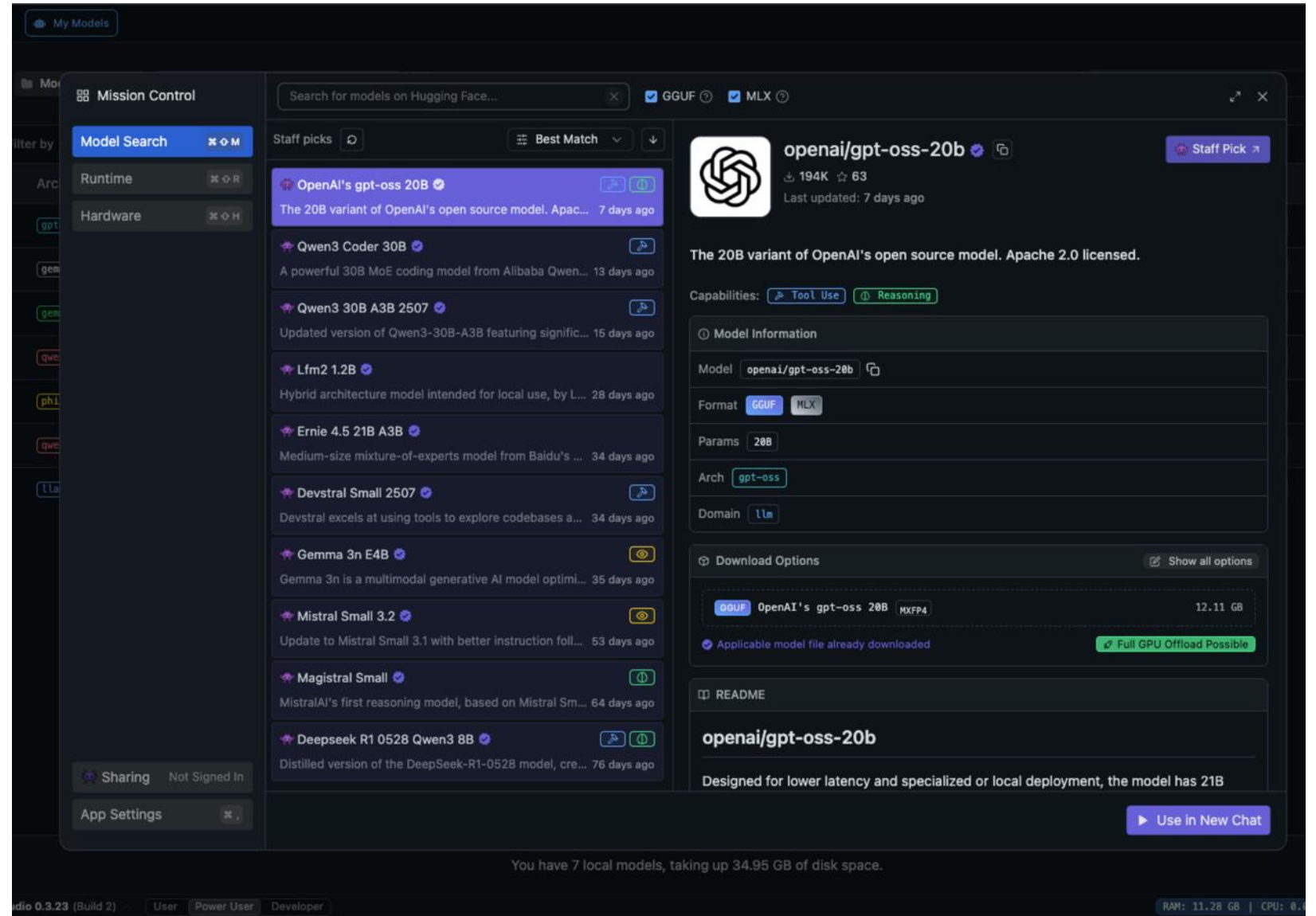
 Image-Text-to-Text • 108B • Updated about 12 hours ago • 622 • 269

Almost 2 MILLION models available on HuggingFace

Model Search

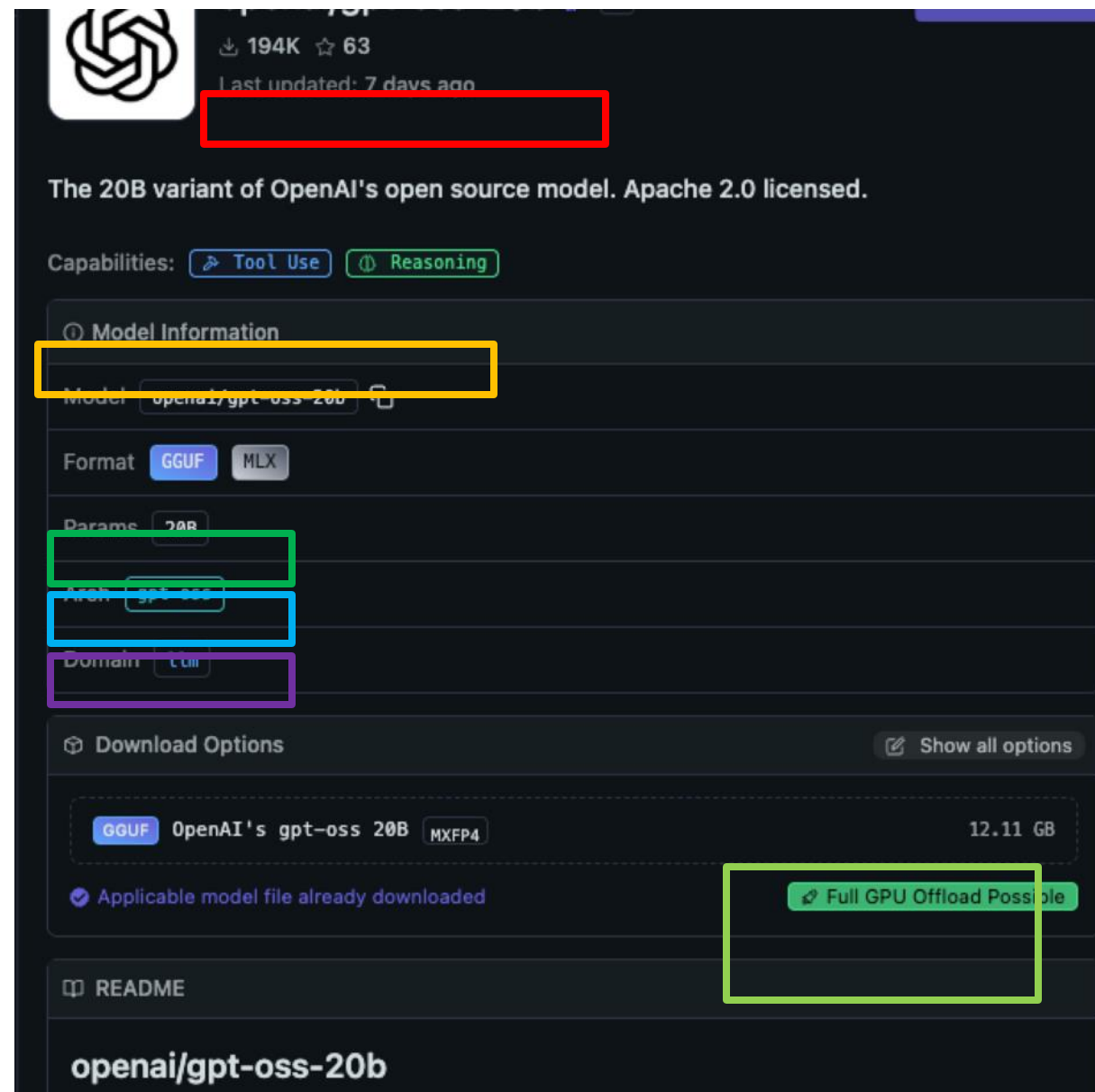
Use LM Studio's integrated model search to find and download models for use.

The “Staff picks” are a great place to start if you are new to local LLM use.



Understanding the Model Card

- **Model Provider and Name**
 - Know WHO trained and built the model
- **Capabilities**
 - Tool Use (aka MCP), Reasoning, Vision
- **Format**
 - GGUF – Cross platform, llama.cpp
 - MLX – Apple Silicon
- **Parameters**
 - Complexity of neural network
- **Architecture**
 - The “family” the model belongs to
- **Model Size**
 - Size of model, hardware requirement to run



LLM Architecture Cheat Sheet

Architecture	Origin	Main Focus / Strengths*
gpt-oss	OpenAI	Replicate (Chat)GPT in open weights/open source
Gemma	Google Deep Mind	Lightweight, efficient, broad application
Phi	Microsoft Research	Small footprint, reasoning. Edge/device based deployments
Granite	IBM	General thinking, multi-language
Mistral	Mistral AI	Efficiency and large context
LLaMA	Meta AI	High efficiency, used as base for many fine tuned models
Qwen	Alibaba	Strong multilingual, compete with western models
Deepseek	Deepseek	High performance reasoning model

* – As described by model providers

Some Hand-gestions...

- Avoid “community models” from sources you don’t know
 - You don’t know what fine-tuning was done
- Test different models and architectures
 - There can be VASTLY different results for use cases
- Try both “large” and “small” models.
 - Larger is not always “better”
- Open Models != Open Source
 - Worth looking at the actual license on models for anything beyond pure experimentation

Demo, Demo, Demo

What we're going to see

Basic AI Chat Features

- Chatting and Model Tuning
- System Prompt Libraries for Custom Assistants

Some Network Use Cases

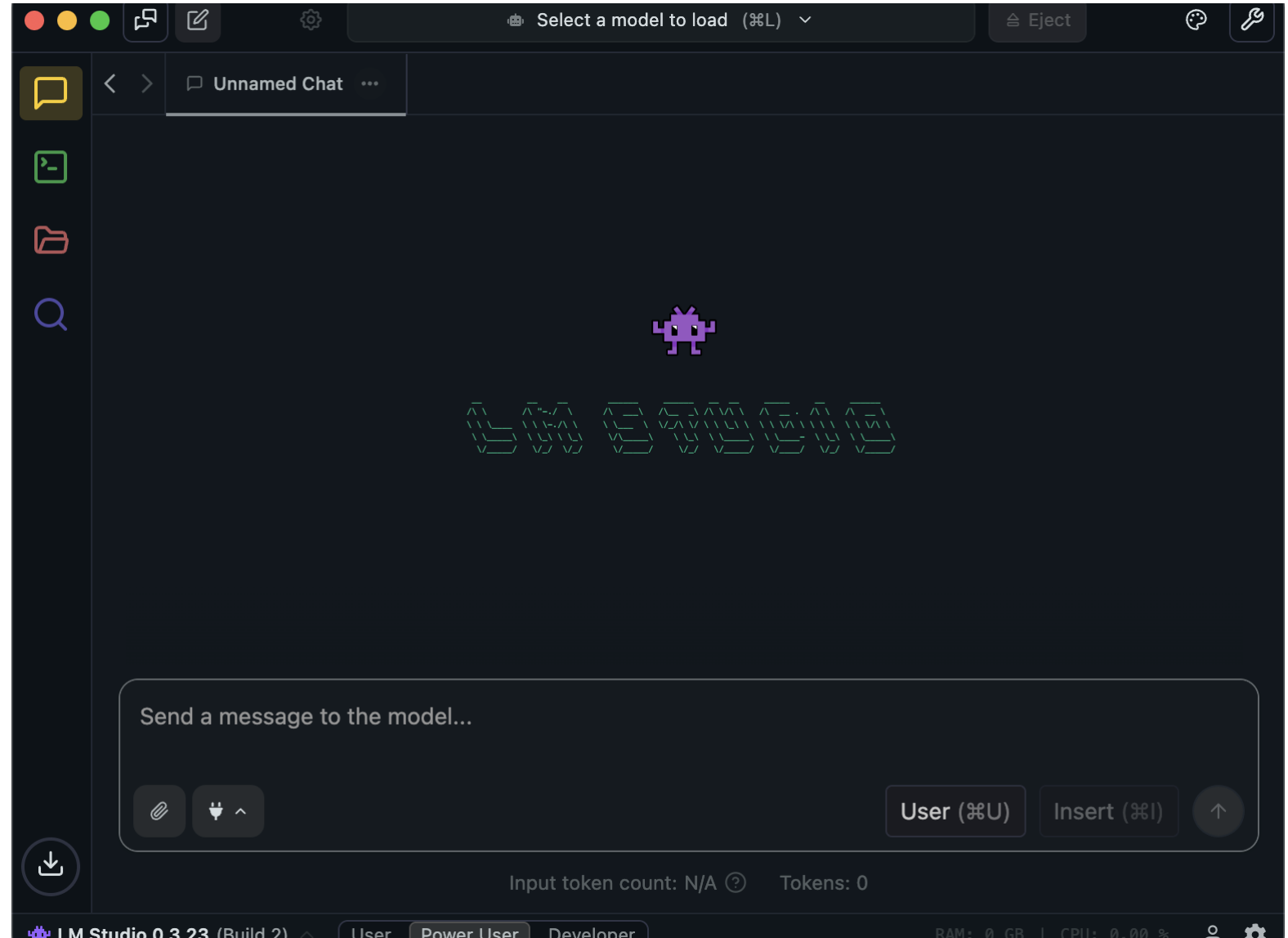
- Network Configuration Analysis with RAG
- Analyzing Network Diagrams
- Network Operations Possibilities

Agentic AI and Development

- Using and Testing MCP Servers
- AI APIs and Python Development

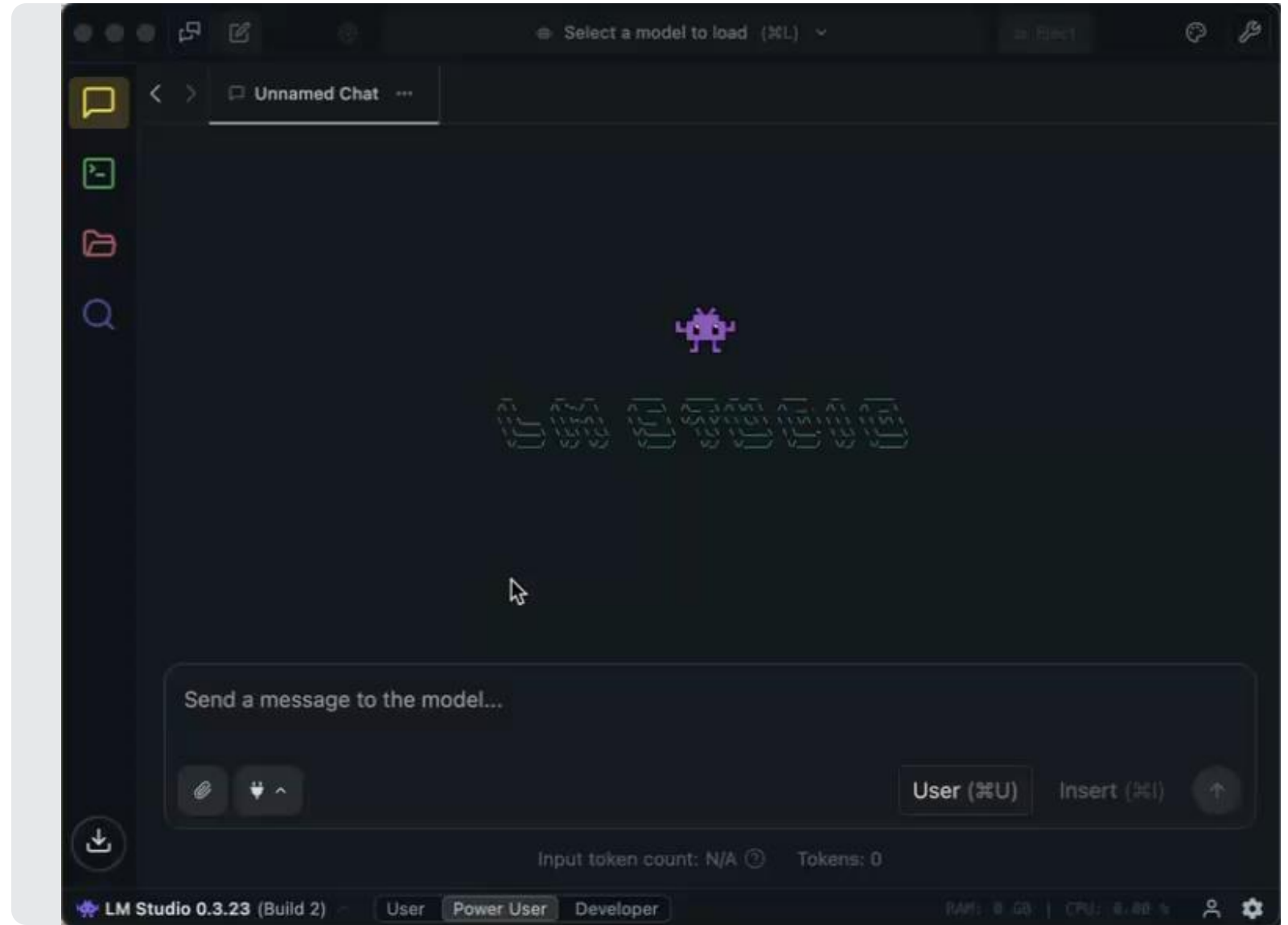
Conversation and Model Tuning

LM Studio allows you to download and work with multiple language models with different capabilities.



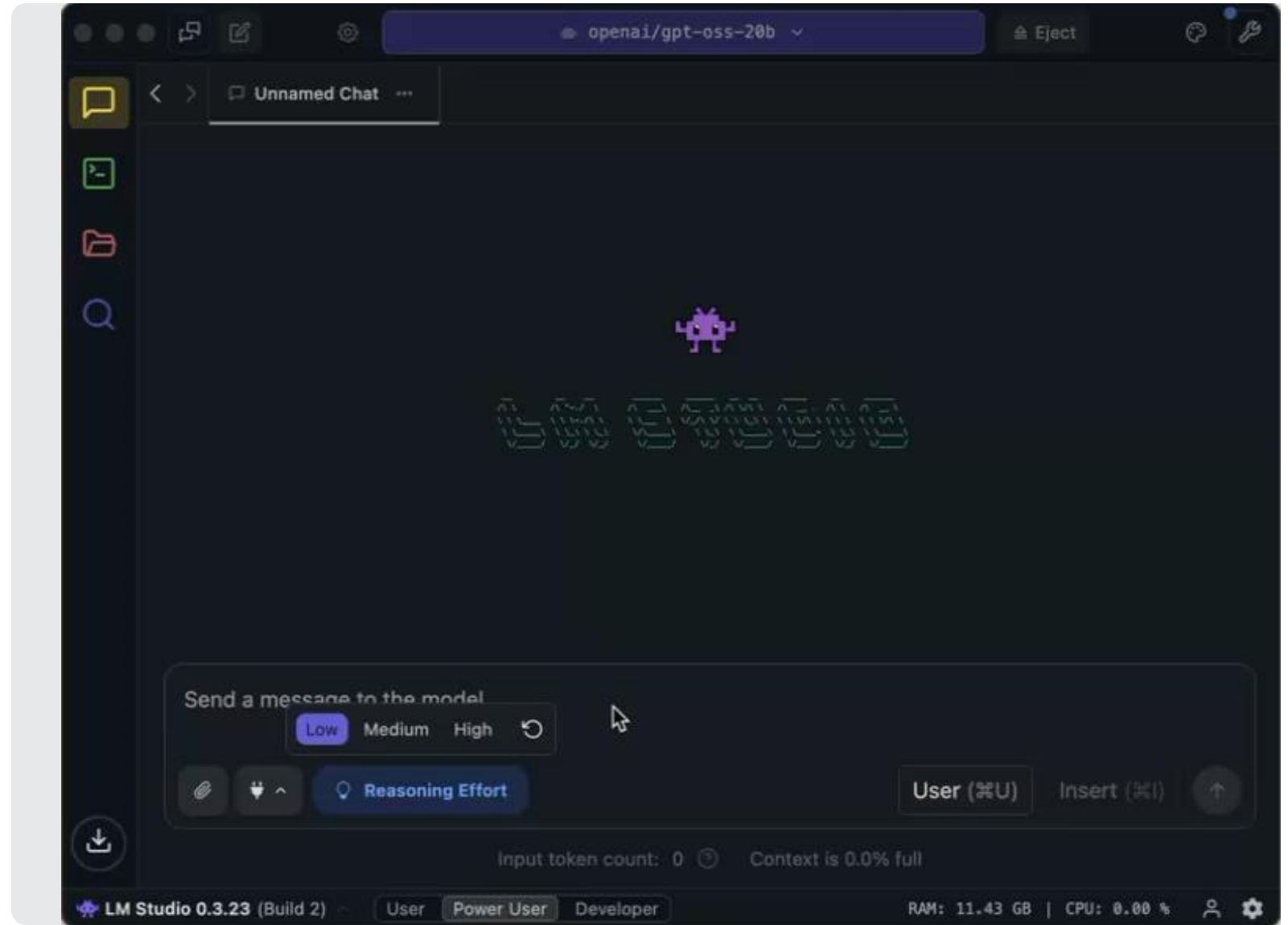
Conversation and Model Tuning

LM Studio allows you to download and work with multiple language models with different capabilities.



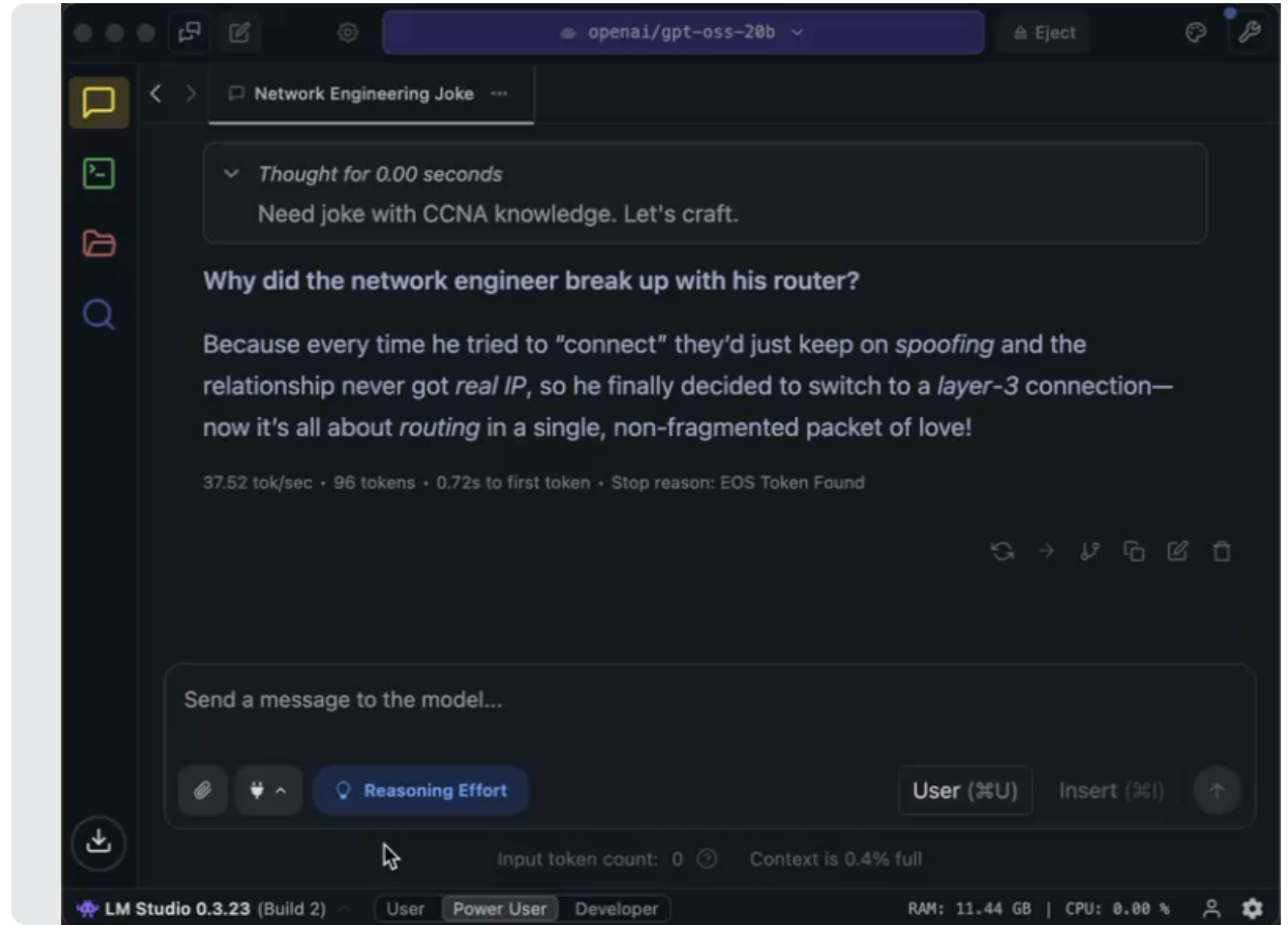
Conversation and Model Tuning

LM Studio allows you to download and work with multiple language models with different capabilities.



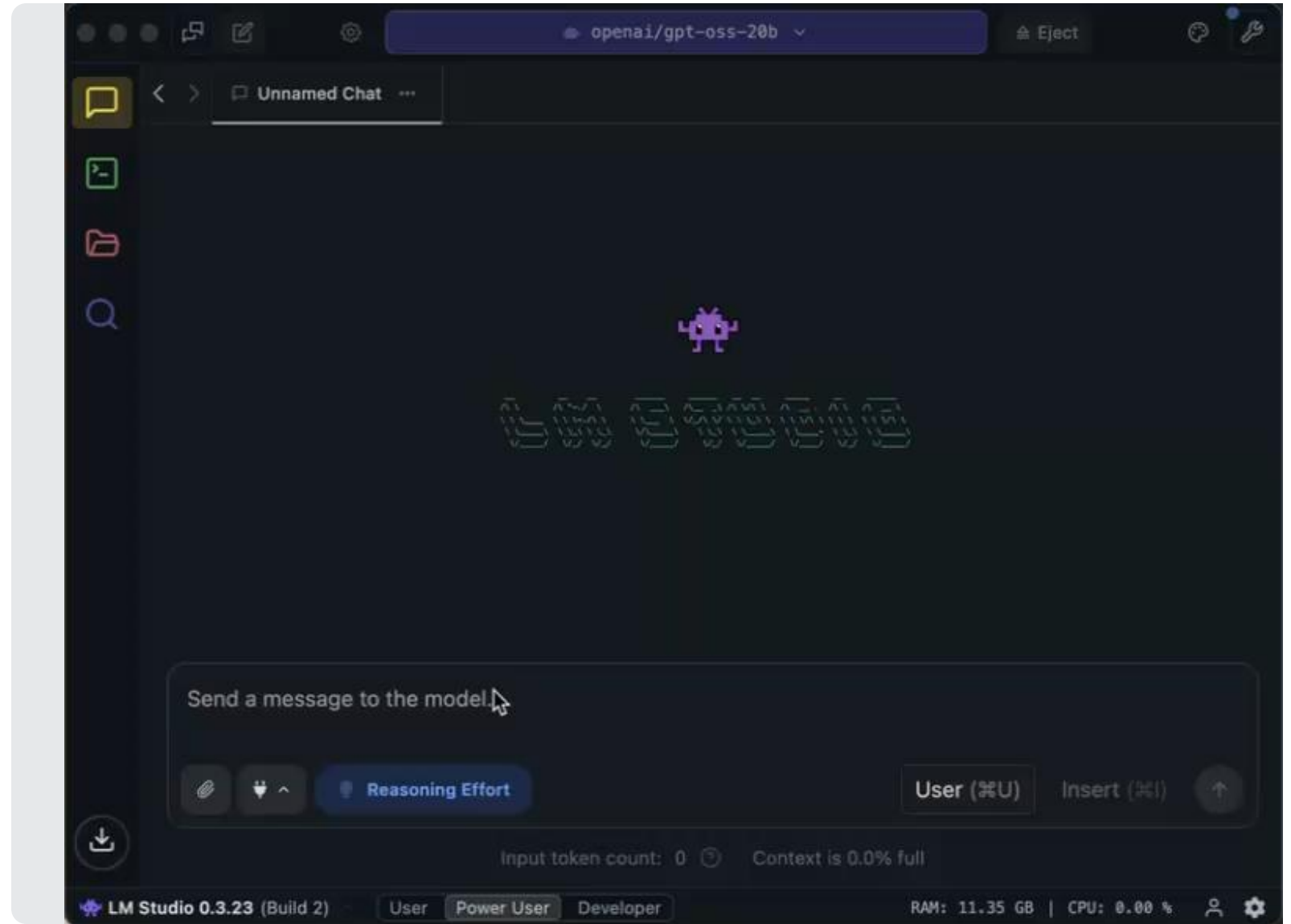
Conversation and Model Tuning

LM Studio allows you to download and work with multiple language models with different capabilities.



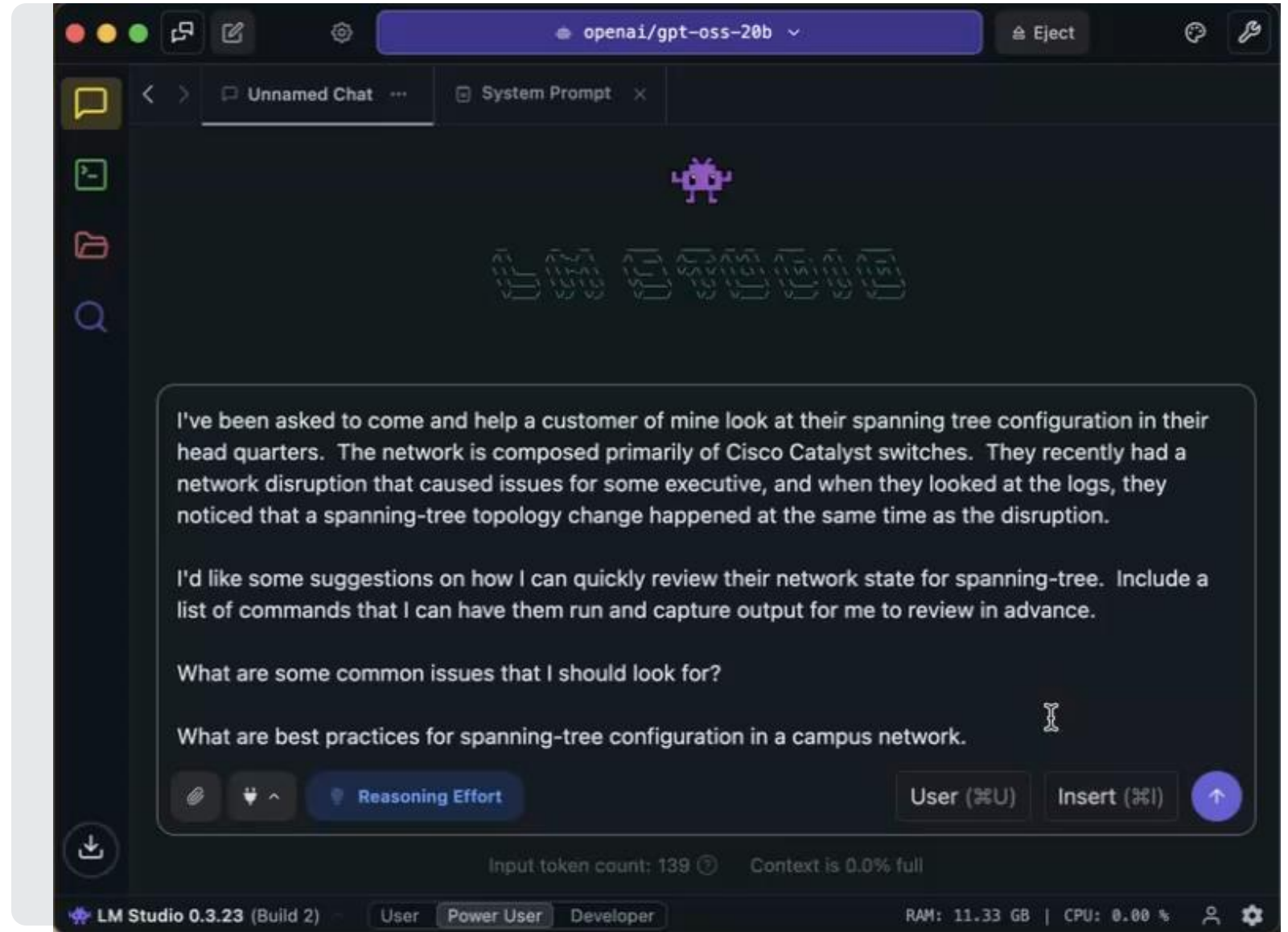
Creating an Assistant with a System Prompt

A single LLM can be a teacher, student, therapist, comedian, or teammate. The difference is the System Prompt that establishes the persona, or role that the LLM will take.



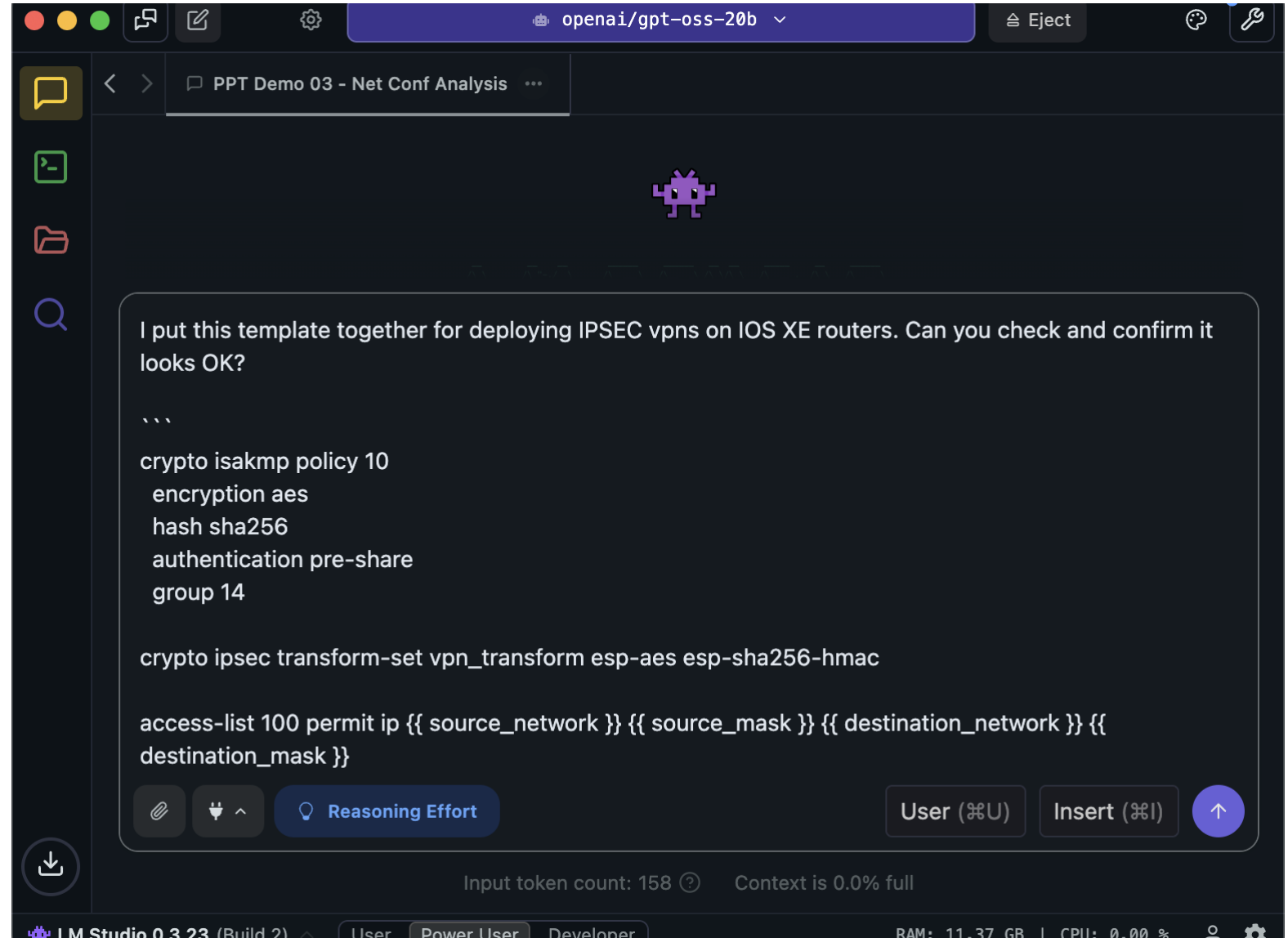
Creating an Assistant with a System Prompt

A single LLM can be a teacher, student, therapist, comedian, or teammate. The difference is the System Prompt that establishes the persona, or role that the LLM will take.



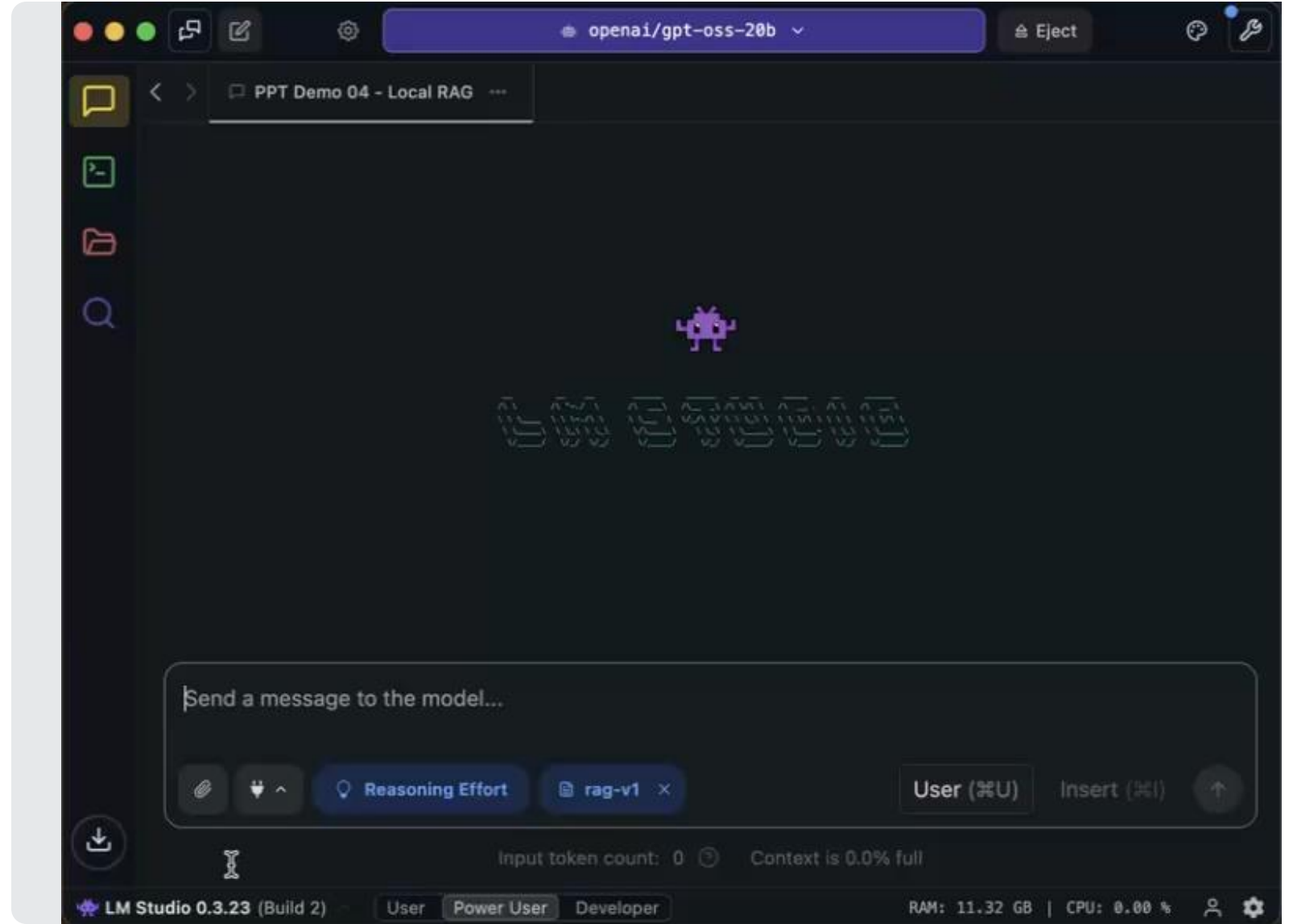
Network Configuration Analysis

We've all done it, we ask a fellow network engineer to review a configuration template before we apply it. But sometimes you don't have another engineer available. AI can step in and help out.



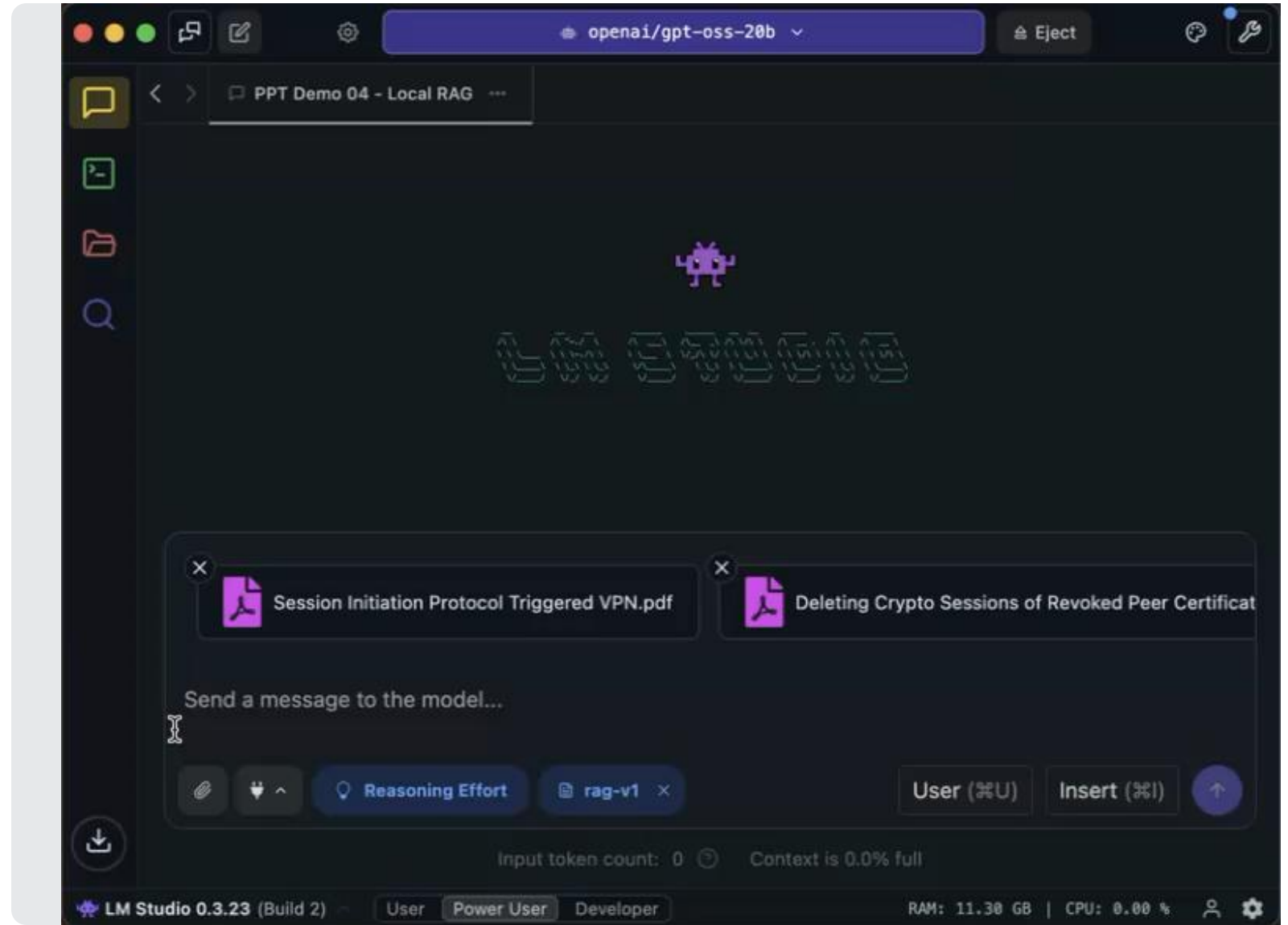
Providing Source Documentation as Part of Queries

LLMs were trained on a LOT of data. Some good, some bad. When you want to ensure good data is used, you can supply documentation to be used for the response. This called RAG, Retrieval Augmented Generation.



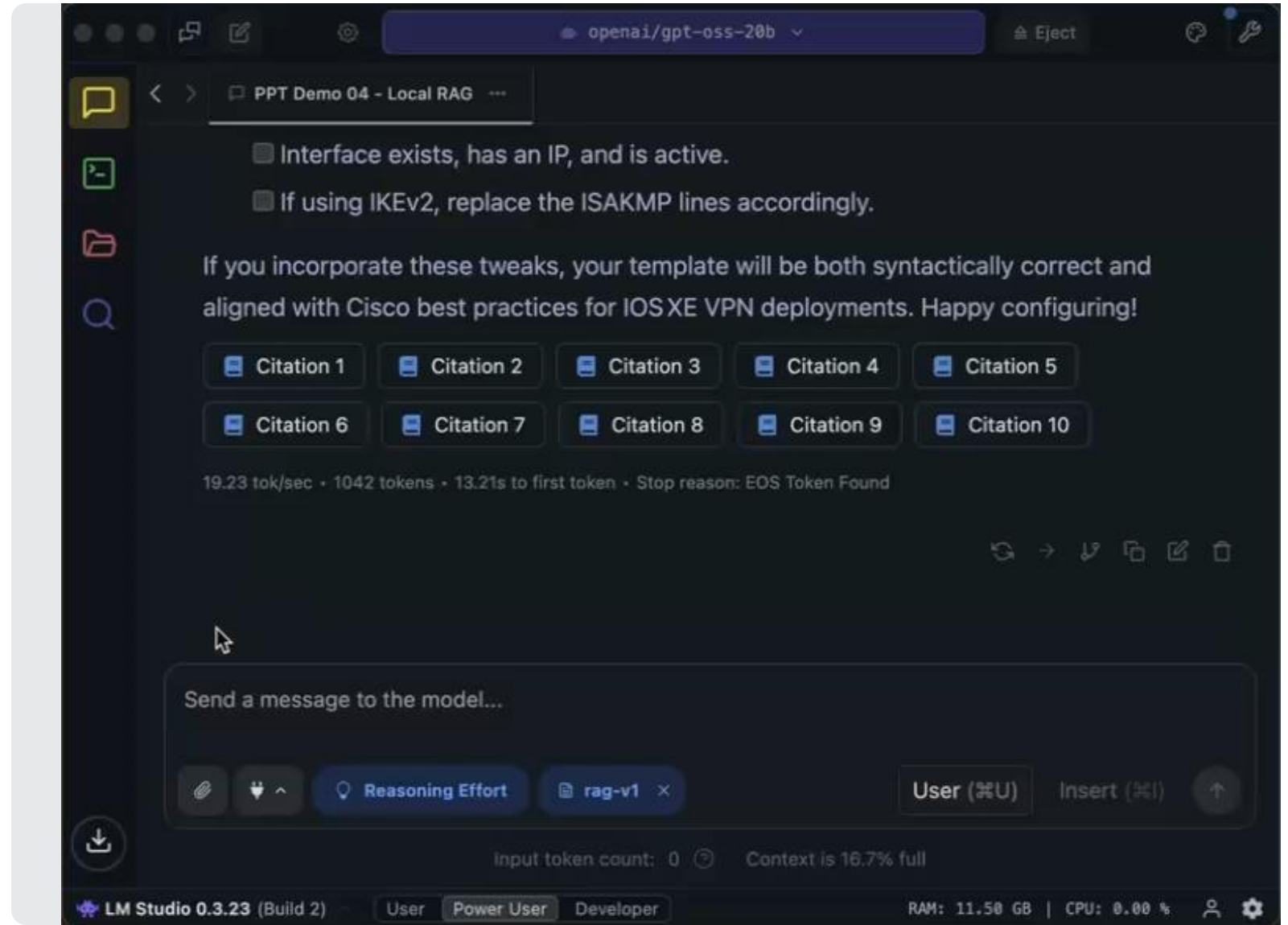
Providing Source Documentation as Part of Queries

LLMs were trained on a LOT of data. Some good, some bad. When you want to ensure good data is used, you can supply documentation to be used for the response. This called RAG, Retrieval Augmented Generation.



Providing Source Documentation as Part of Queries

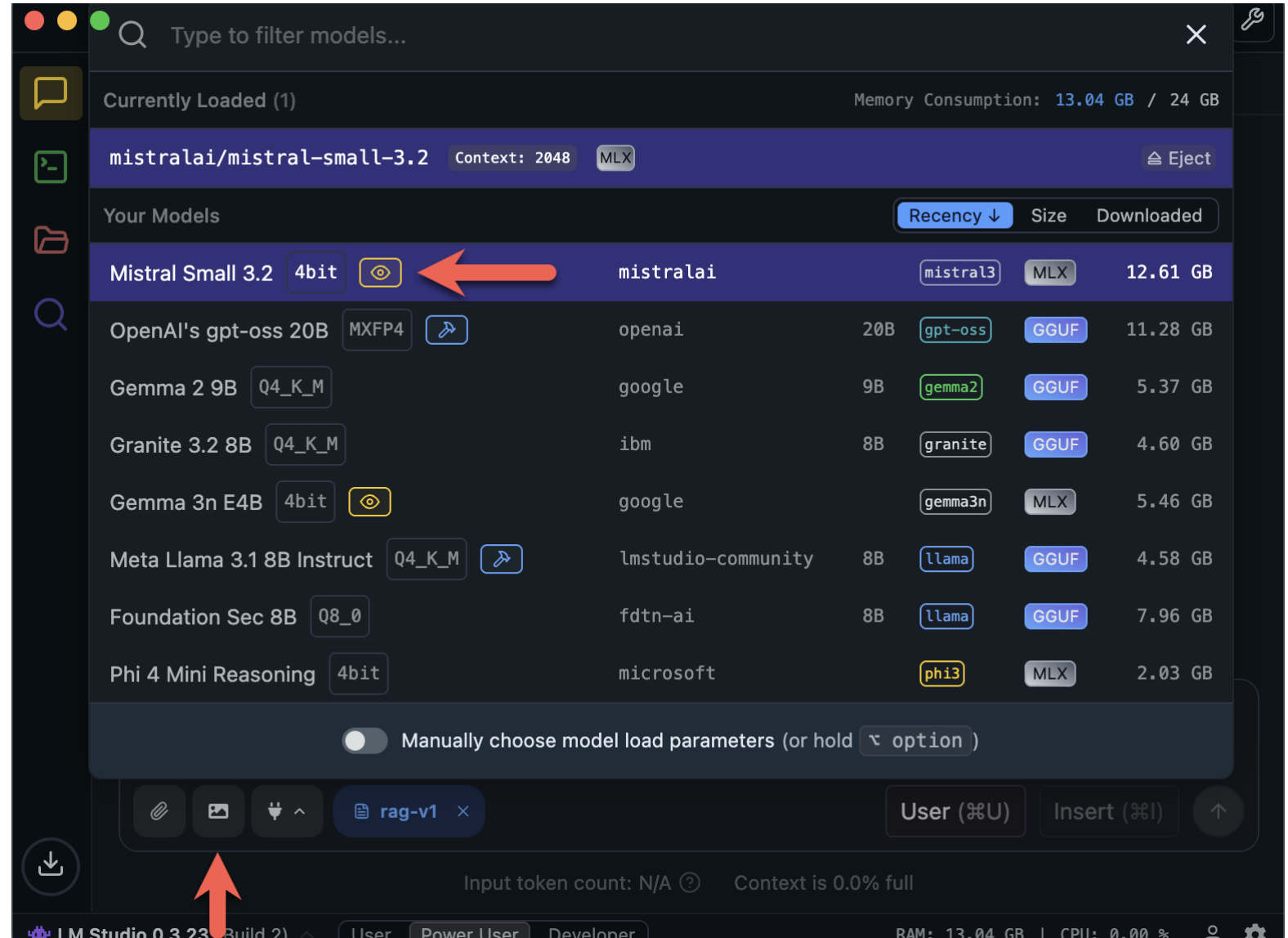
LLMs were trained on a LOT of data. Some good, some bad. When you want to ensure good data is used, you can supply documentation to be used for the response. This called RAG, Retrieval Augmented Generation.



Network Diagram Analysis

LLMs can process more than just words, some can work with images as well. Allowing you to upload a network diagram and get information from it.

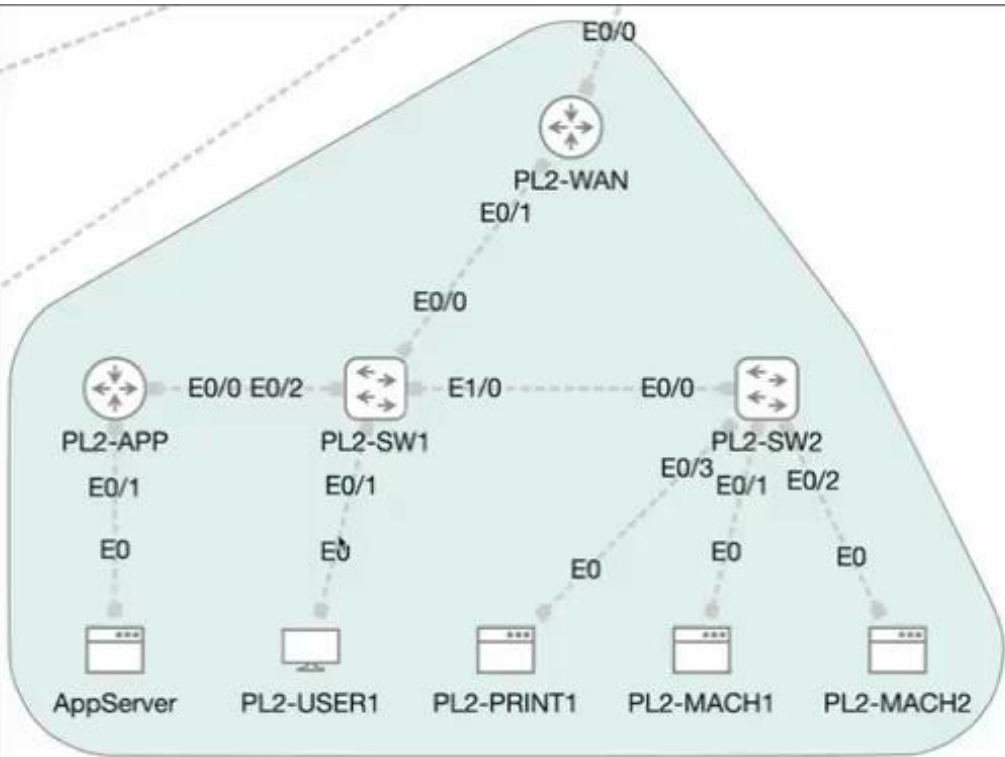
But be aware, processing complex images maybe too much for your local LLM.



Network Diagram Analysis

LLMs can process more than just words, some can work with images as well. Allowing you to upload a network diagram and get information from it.

But be aware, processing complex images maybe too much for your local LLM.



Widget Woods Plant

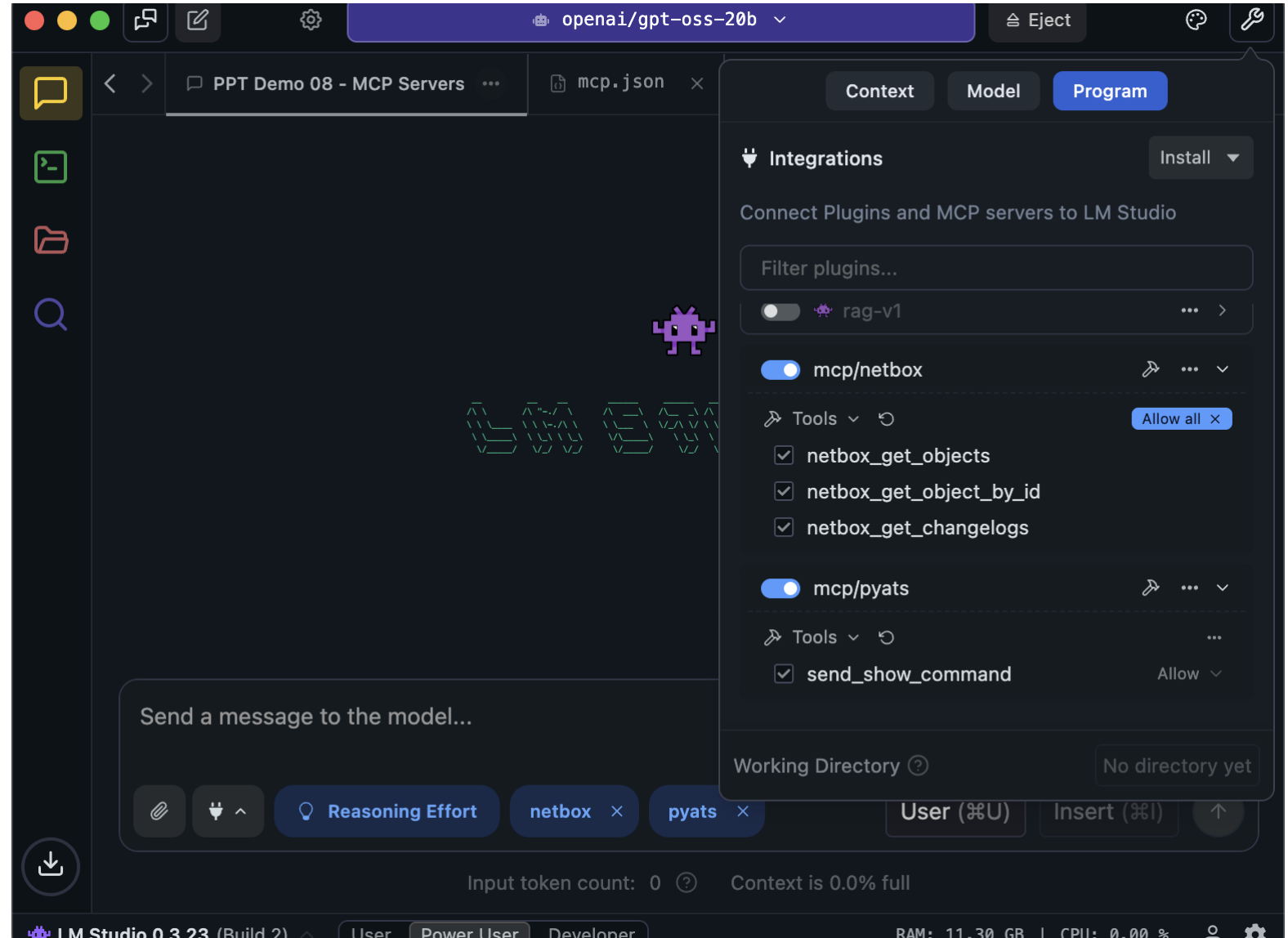
Widget Woods Network: 10.132.0.0/15		
Name	VLAN ID	Subnet
Management	10	10.132.0.0/22
Users	101	10.133.200.0/24
Printers	121	10.133.201.0/24
LegacyApp	999	10.133.204.0/24
Plant1	666	10.132.64.0/18
Plant2	667	10.132.128.0/20

MCP Server Development and Testing

MCP (Model Context Protocol) is the current “standard” for developing tools for Agentic AI systems.

MCP servers send data about your systems to the LLM to process, so using cloud LLMs requires a high degree of trust.

With LM Studio, you can test MCP servers 100% locally.

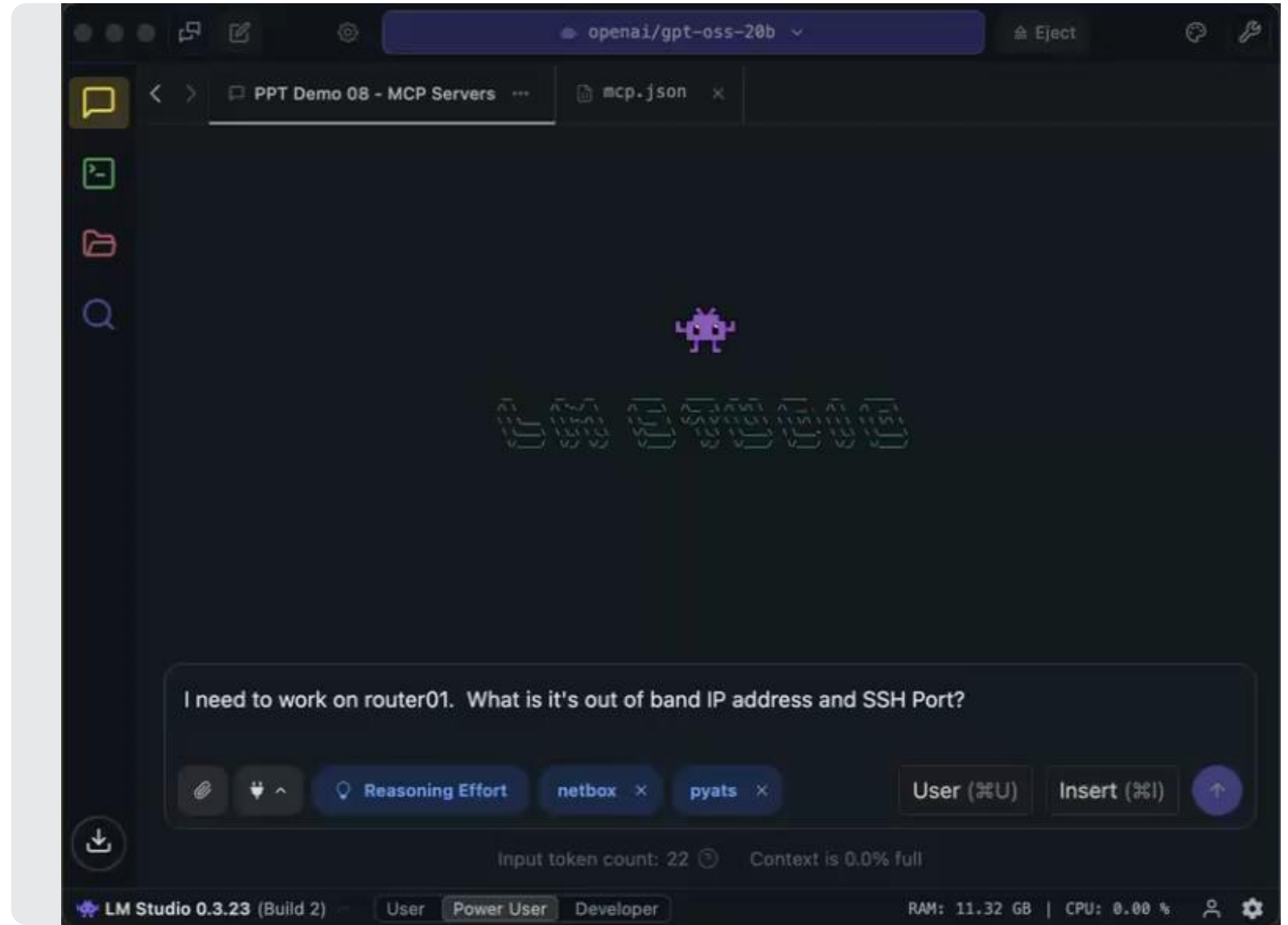


MCP Server Development and Testing

MCP (Model Context Protocol) is the current “standard” for developing tools for Agentic AI systems.

MCP servers send data about your systems to the LLM to process, so using cloud LLMs requires a high degree of trust.

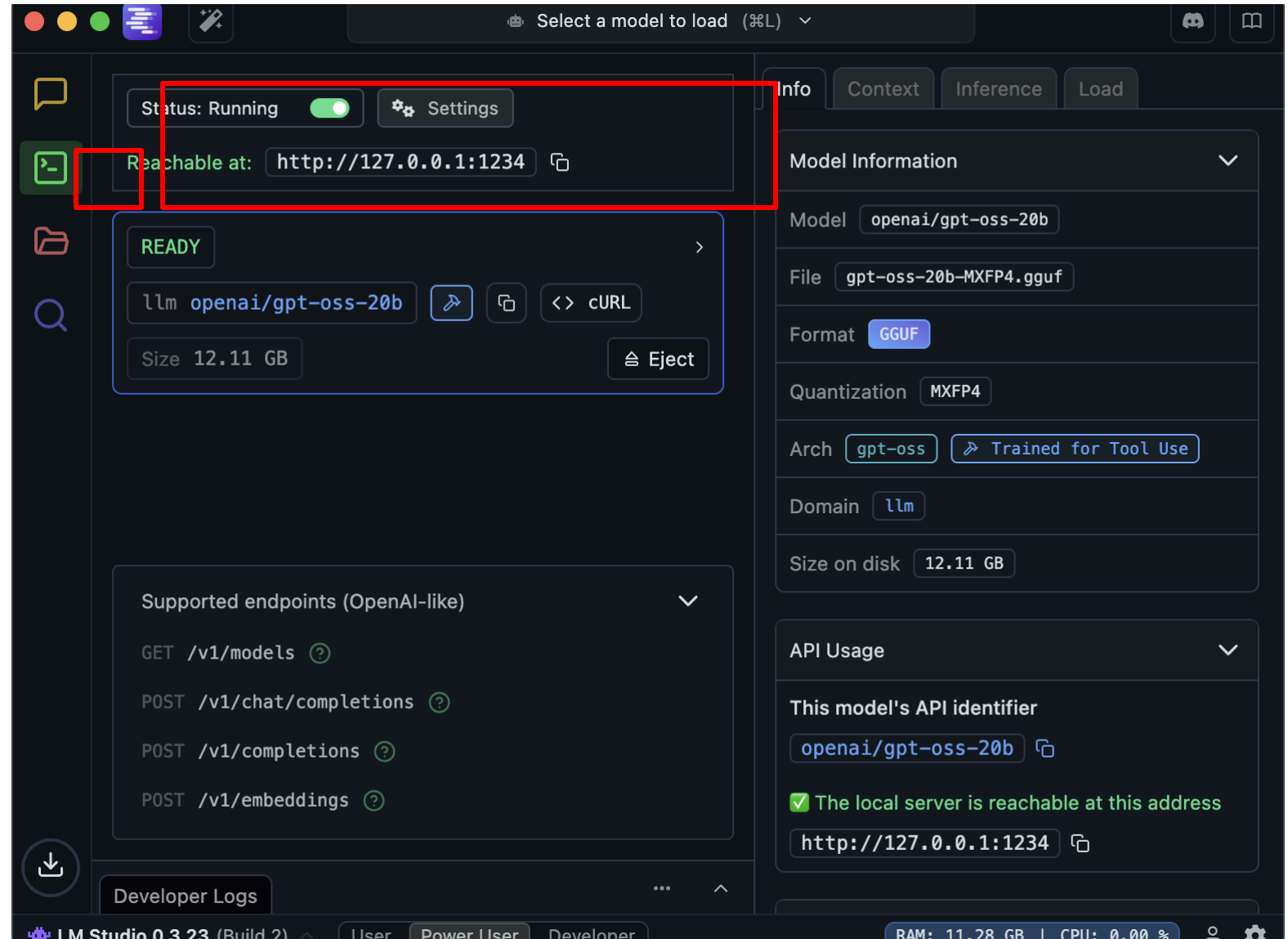
With LM Studio, you can test MCP servers 100% locally.



AI Development in Python

Access to LLMs through APIs is a powerful new tool for our network automation work.

LM Studio can enable OpenAI compatible REST APIs, and offers Python SDKs for building your own apps and scripts that leverage the local LLMs it manages and runs.



AI Development in Python

Access to LLMs through APIs is a powerful new tool for our network automation work.

LM Studio can enable OpenAI compatible REST APIs, and offers Python SDKs for building your own apps and scripts that leverage the local LLMs it manages and runs.

```
studio > 🐍 lmstudio_openai_example01.py > ...
# Example: reuse your existing OpenAI setup
from openai import OpenAI

# Point to the local server
client = OpenAI(base_url="http://localhost:1234/v1", api_key="lm-studio")

completion = client.chat.completions.create(
    ... model="model-identifier",
    ... messages=[
    ...     {"role": "system", "content": "Always answer like a pirate."},
    ...     {
    ...         "role": "user",
    ...         "content": "Why should network engineers be excited about AI
    ...     },
    ... ],
    ... temperature=0.7,
)

print(completion.choices[0].message.content)
```

AI Development in Python

Access to LLMs through APIs is a powerful new tool for our network automation work.

LM Studio can enable OpenAI compatible REST APIs, and offers Python SDKs for building your own apps and scripts that leverage the local LLMs it manages and runs.

```
lmstudio git:(main) x python lmstudio_openai_example01.py
```

r 'round, ye wired buccaneers of the net, and lend an ear to this here tale o' w
minds ought ta be jumpin' for joy at the thought o' Artificial Intelligence! 1.
the Uncharted** AI can sift through terabytes o' logs faster than a cannonba
a. It'll spot patterns, sniff out bottlenecks, and even predict where storms (a.l
comin' afore they hit yer crew. No more squintin' at data like it's buried treas
n – The New Black Spot** Ye can automate routine tasks—config changes, firmw
tch management—with the ease o' a deckhand rollin' in the breeze. That frees up y
he grander quests: designing networks that be as swift as a frigate and as resil
rced with steel. 3. **Dynamic Routing & Traffic Management** Think of AI as
artermaster on board, constantly re-routing traffic through the healthiest lanes,
oads, and keepin' latency low like a well-aimed harpoon. It can even adapt in rea
surges or rogue waves (those sudden spikes). 4. **Security – The Black Spot of t
s** AI-powered threat detection spots malicious packets faster than a lookou
arns new attack patterns, flags anomalies, and can auto-mitigate threats—so ye ca
safe from the scourge o' cyber-piracy. 5. **Capacity Planning & Forecasting**
analytics, AI tells ye when to raise the sails (add bandwidth) or lower them (tra
It'll help avoid the dreaded “bandwidth black hole” where demand outpaces supply
floundering. 6. **Cost-Cutting Treasure Trove** By optimizing resource use-
rdware—AI turns a costly ship into a lean, mean, profit-generatin' vessel. No mo
for unused bandwidth or buying new gear that never sees the light o' day. 7.
s Learning & Self-Healing Networks** Some AI systems can learn from past inc
ically patch known weaknesses—think of it as a crew that trains every day, improv
fought on the high seas of data. 8. **Innovation – The New Horizon** With
, ye can experiment with cutting-edge concepts like SD-WAN, intent-based network
m-ready protocols. It's like discoverin' a new continent—full o' untold riches!
oger, network engineers! Embrace AI, and let it steer yer ship through calm water
alike, makin' yer operations smoother, safer, and more profitable than ever befo

Closing

Useful Links

- LM Studio Home Page
- Cisco U Tutorial: [Run Your Own LLM Locally For Free and with Ease](#)
- Blog: Creating a NetAI Playground for Agentic AI Experimentation
- Blog: Bringing Some Source of Truth to Your NetAI Agentic Playground

Stay in touch!

Hank Preston

Webex/Email: hapresto@cisco.com

Blogs: <https://blogs.cisco.com/author/hankpreston>



Thank you

CISCO Live !