

AI: Easy to use local RAG and MCP with LM Studio

magnusviri (James Reynolds), October 22, 2025

About James Reynolds

- Bachelor of Music, 1999, University of Utah
- Marriott Library, 14 years (Mac admin)
- School of Biological Sciences, 11 years (Mac, Linux, network admin)
- College of Science (SBS), 3 months
- Security and AI are “hobbies” of mine



Slides Will Be At magnusviri.com

Computer Requirements

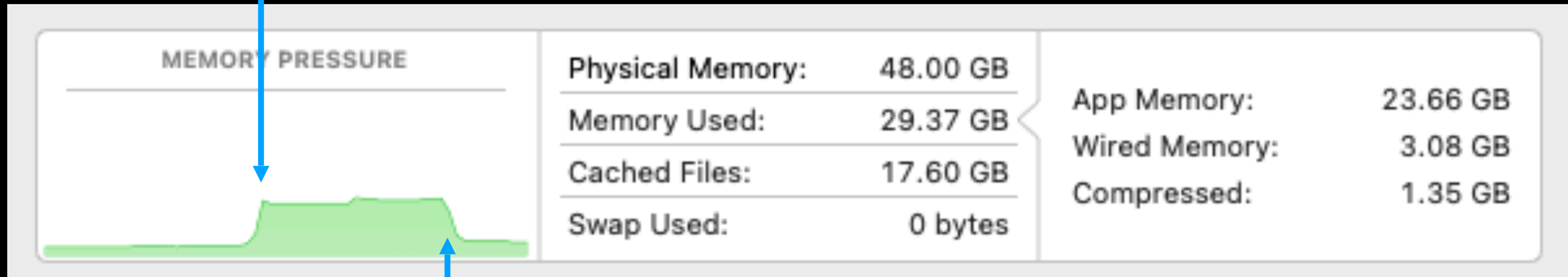
- Nvidia CUDA, AMD Ryzen, Mac M-Series processor, or a fast CPU
- Minimum of 8GB of GPU RAM (used to be 16GB but it's getting lower)

Why Use AI Locally?

- Using “dumb” AI is the best way to see what it’s really capable of
- Privacy
- Cloud AI might become extremely expensive
- People are losing loved ones to AI-Fueled Spiritual Fantasies

<https://www.rollingstone.com/culture/culture-features/ai-spiritual-delusions-destroying-human-relationships-1235330175/>

Model loaded into RAM



Model unloaded from RAM

Learn How AI Works

- “The moment we stopped understanding AI” <https://youtu.be/UZDiGooFs54>
- “The Essential Main Ideas of Neural Networks” <https://youtu.be/CqOfi41LfDw>
- “Watching Neural Networks Learn” <https://youtu.be/TkwXa7Cvfr8>
- “But what is a neural network?” <https://youtu.be/aircAruvnKk>
- <https://github.com/ianand/spreadsheets-are-all-you-need>
- “AI Is Dangerous, but Not for the Reasons You Think”
<https://youtu.be/eXdVDhOGqoE>
- “Follow the Funding of AI Doom” <https://youtu.be/Zrs3uD3xLoo>

AI Sucks

S

what is the world record for crossing the english channel entirely on foot



The world record for crossing the English Channel entirely on foot is held by Christof Wandratsch of Germany, who completed the crossing in 14 hours and 51 minutes on August 14, 2020. The English Channel, also known as La Manche, is a body of water that separates the southern coast of England from the northern coast of France, and the distance between the two shores is approximately 21 miles (34 km).

Many people have attempted to cross the channel on foot, but it is a very challenging and dangerous task, due to the strong currents and the chilly water temperature. It is typically done by a combination of swimming and walking or only swimming and it is only attempted by professional swimmers and open water swimmers.



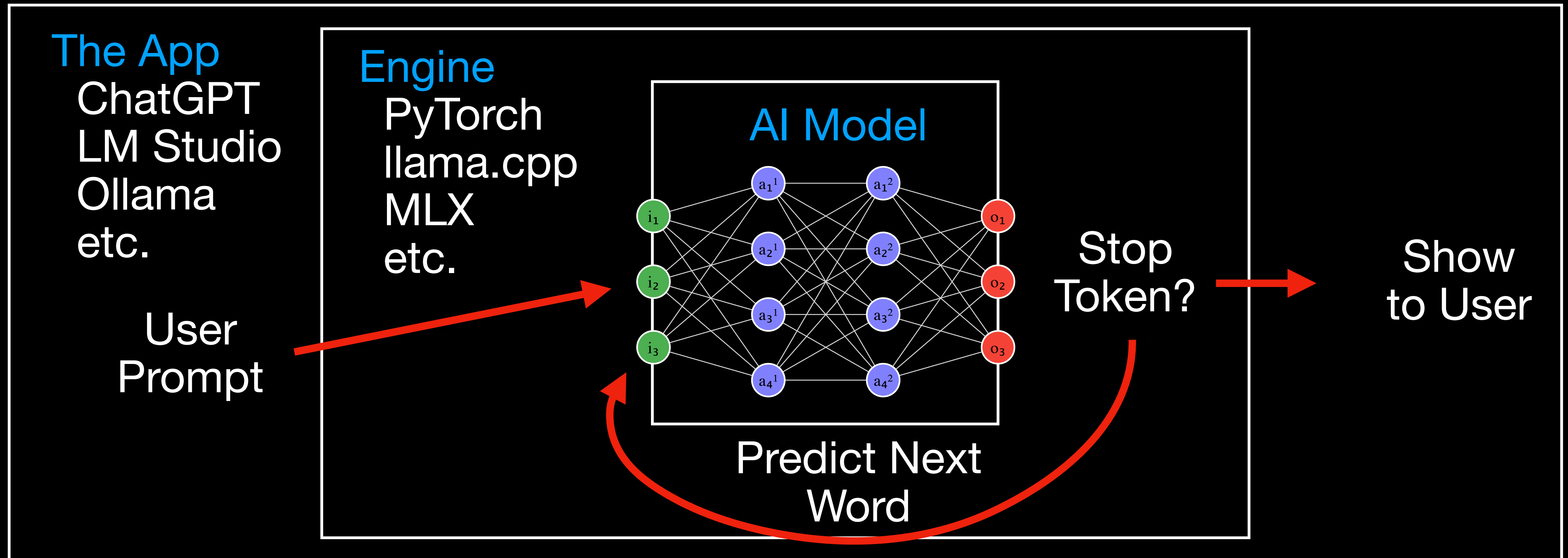
↻ Regenerate response



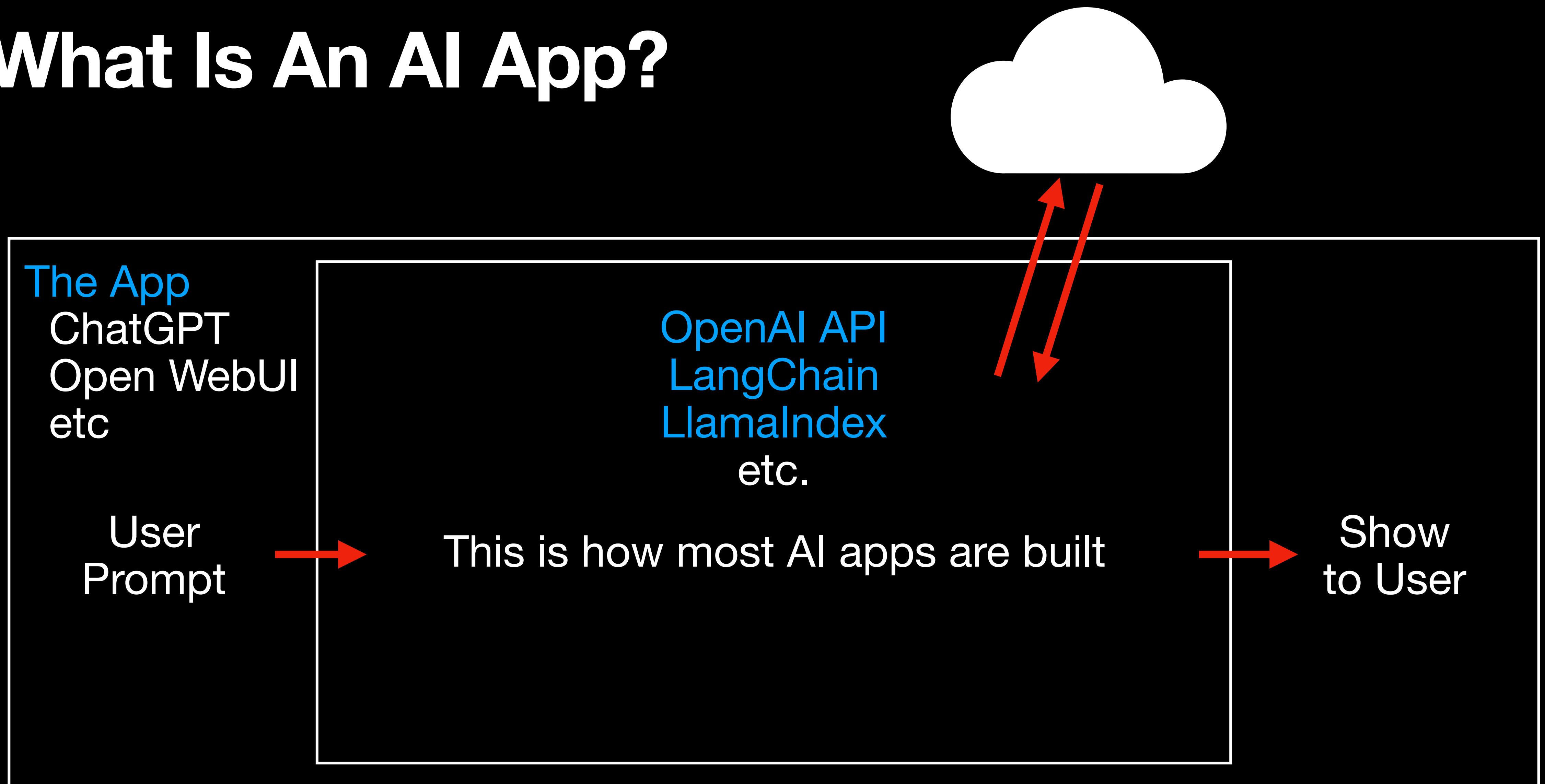
How to Improve AI?

- Fine Tuning
 - LORAs
- Change the seed, temperature, Top K Sampling, etc
- Structured Output
- Change the system or user prompt
- Manage the context (including RAG)
- Function calling (including MCP)

What Is An AI App?

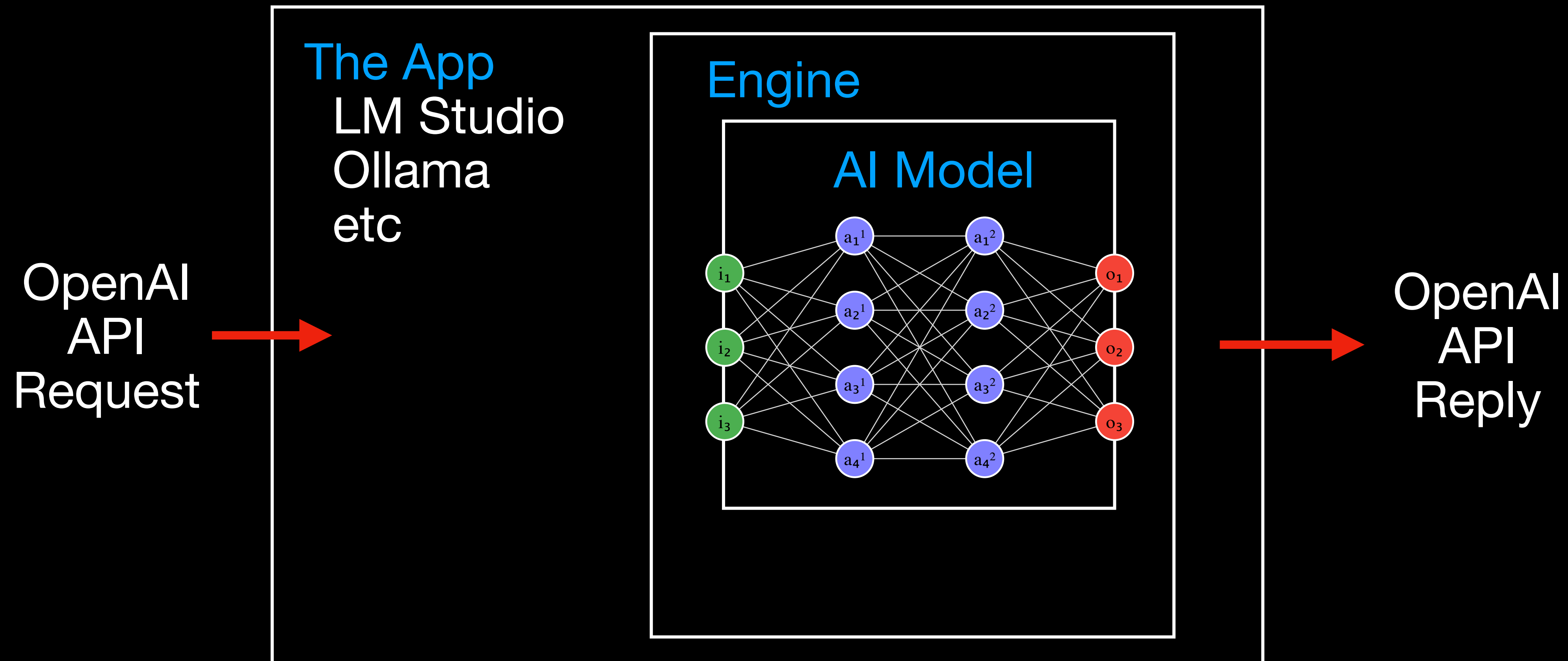


What Is An AI App?

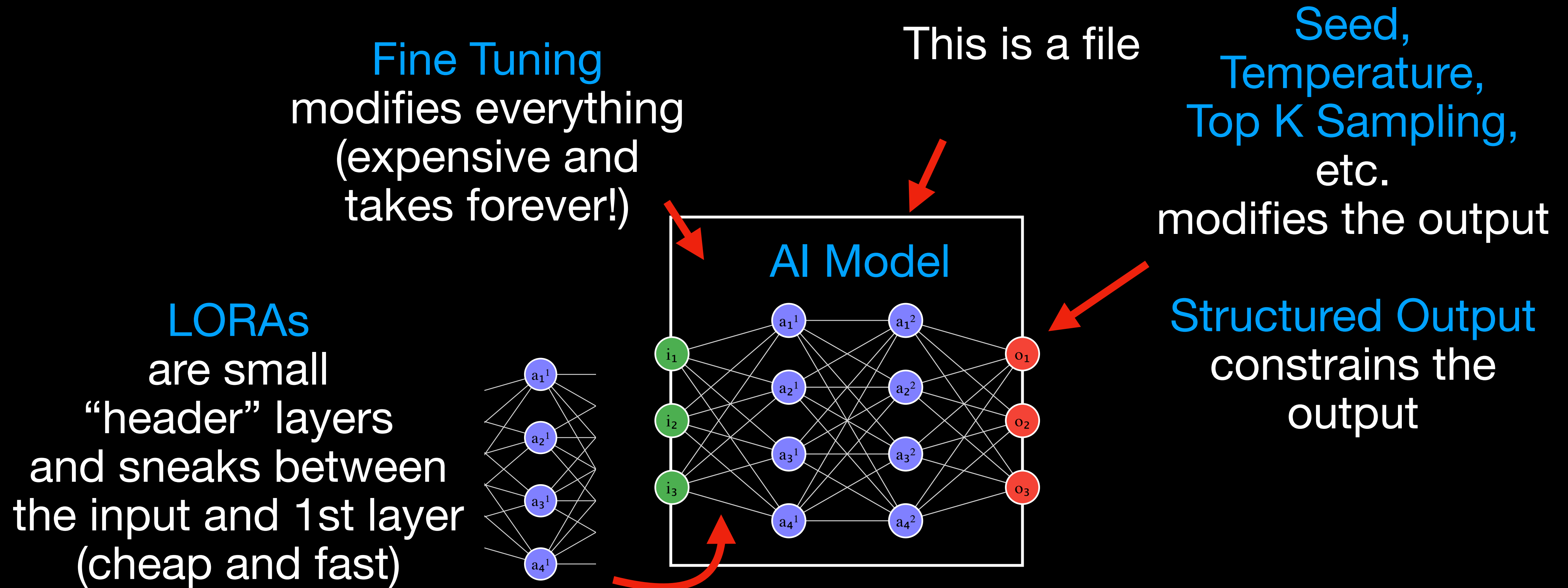


What Is An AI App?

Local apps can also serve the API



Let's Look at the AI Model



Fine Tuning and LORAs are the only times the inside of the models are messed with. **Control vectors** can mess with the math in the model but I don't know of an app that allows us non-scientists to play with them.

HuggingFace has 2,151,499 models

The screenshot displays the Hugging Face website interface. At the top, the Hugging Face logo and a search bar are visible. Navigation tabs for 'Main', 'Tasks', 'Libraries', 'Languages', 'Licenses', and 'Other' are present. The 'Main' tab is active, showing a 'Tasks' section with buttons for Text Generation, Any-to-Any, Image-Text-to-Text, Image-to-Text, Image-to-Image, Text-to-Image, Text-to-Video, Text-to-Speech, and a '+ 42' button. Below this is a 'Parameters' section with a slider ranging from '< 1B' to '> 500B'. The 'Libraries' section features buttons for PyTorch, TensorFlow, JAX, Transformers, Diffusers, Safetensors, ONNX, GGUF, Transformers.js, MLX, and Keras, along with a '+ 41' button. The 'Apps' section includes buttons for vLLM, TGI, llama.cpp, MLX LM, LM Studio, Ollama, Jan, and a '+ 13' button. On the right, the 'Models' section shows a total of 2,151,499 models and a 'Filter by name' button. A list of featured models is displayed, including:

- inclusionAI/Ling-1T**: Text Generation • 1000B • Updated 1 day ago • 1.64k downloads • 393 likes
- neuphonic/neutts-air**: Text-to-Speech • 0.7B • Updated 5 days ago • 18.5k downloads • 576 likes
- nanonets/Nanonets-OCR2-3B**: Image-Text-to-Text • 4B • Updated 1 day ago • 314 downloads • 209 likes
- Phr00t/Qwen-Image-Edit-Rapid-AIO**: Text-to-Image • Updated 1 day ago • 269 likes
- Kwaipilot/KAT-Dev-72B-Exp**: Text Generation • 73B • Updated 2 days ago • 1.78k downloads • 129 likes
- LiquidAI/LFM2-8B-A1B**: Text Generation • 8B • Updated 1 day ago • 5.47k downloads • 205 likes
- Qwen/Qwen3-VL-4B-Instruct**: Image-Text-to-Text • 4B • Updated about 2 hours ago • 156 downloads • 92 likes

What's Actually Sent To the Model?

Engine

Llama Prompt Format aka Prompt Template

<|begin_of_text|><|start_header_id|>system<|end_header_id|>

You are a helpful AI assistant for travel tips and
recommendations<|eot_id|>

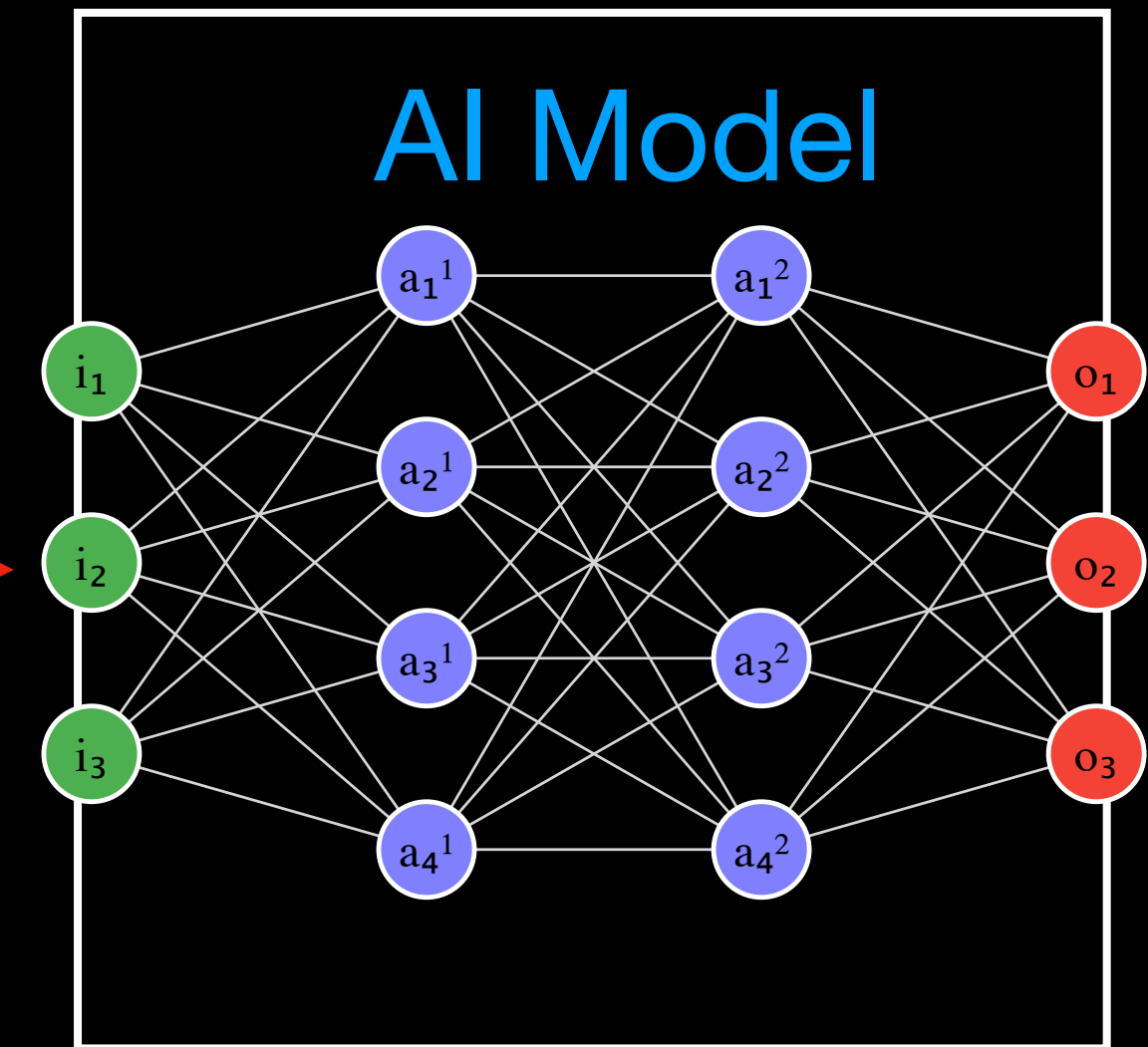
<|start_header_id|>user<|end_header_id|>

What can you help me with?<|eot_id|>

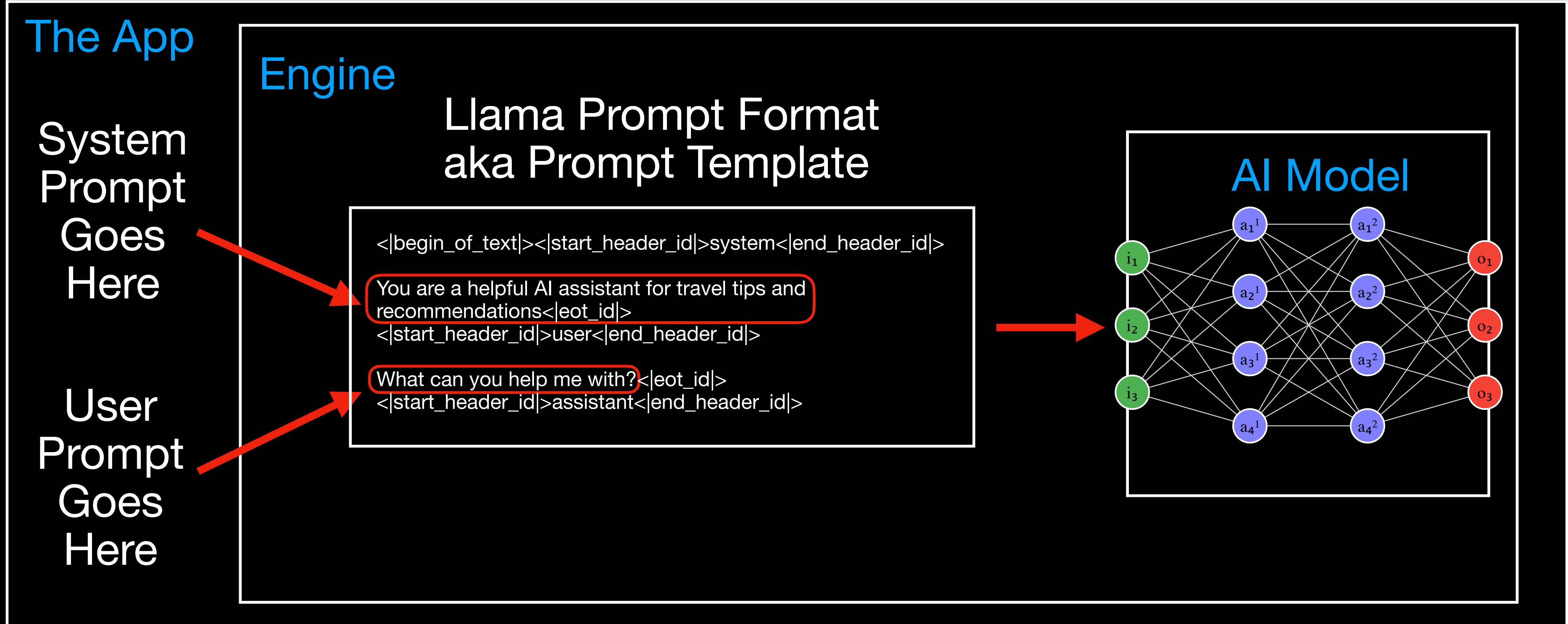
<|start_header_id|>assistant<|end_header_id|>



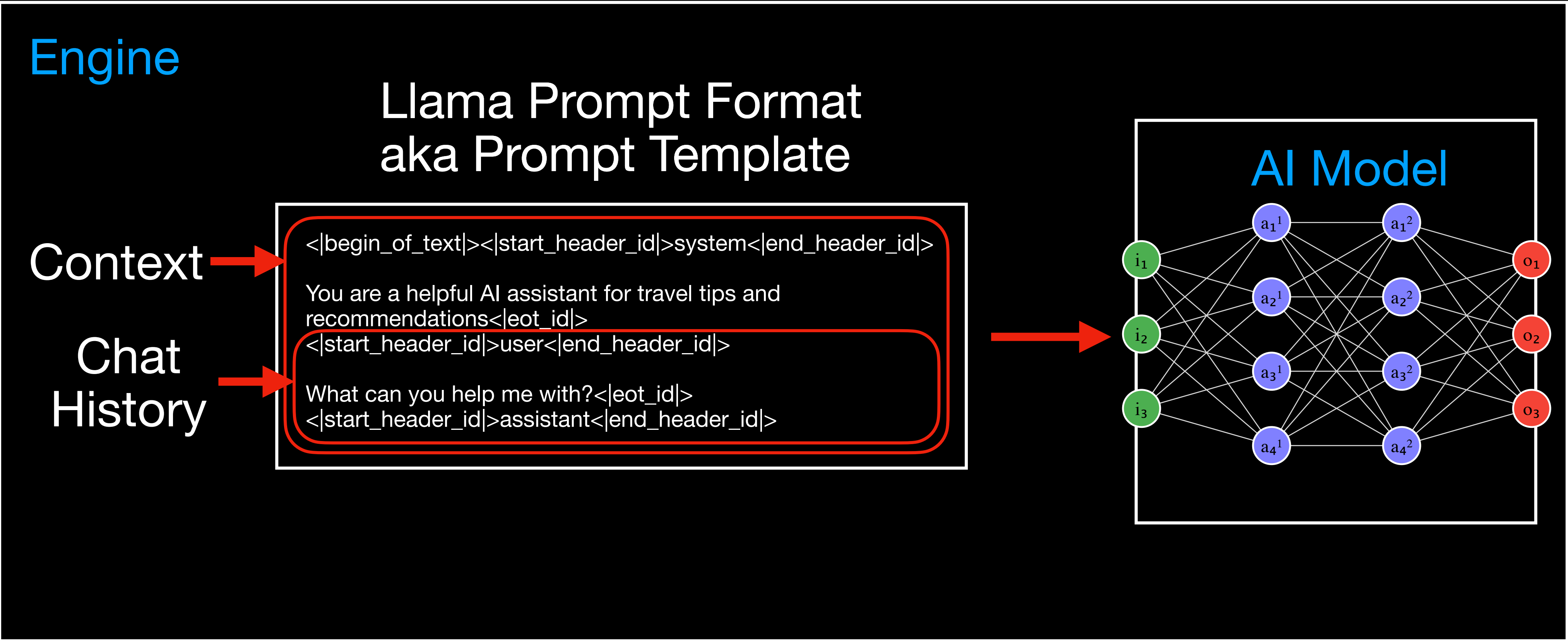
AI Model



Prompt Engineering?



Context vs Chat History



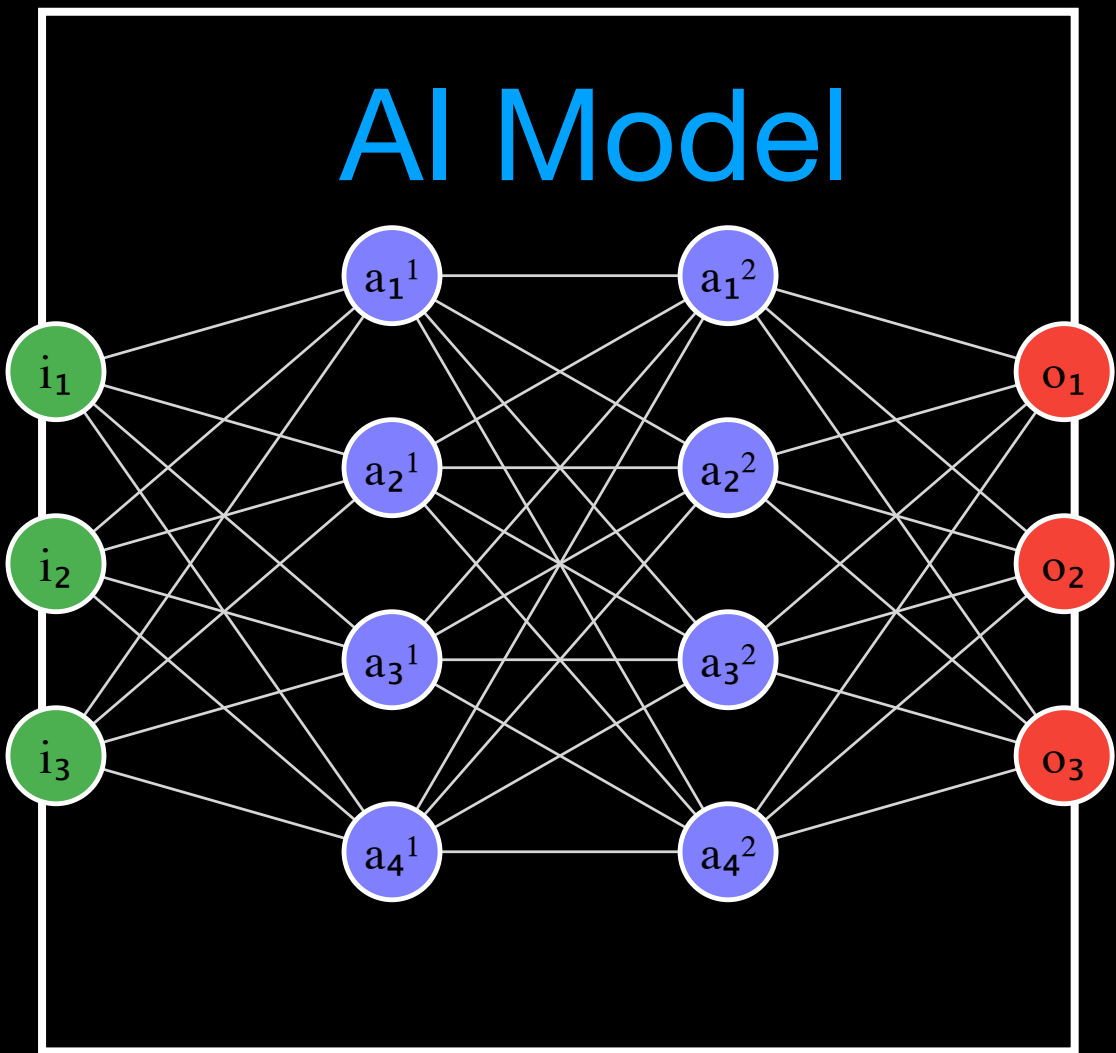
Cue the AI To Start Next Word Prediction

Engine

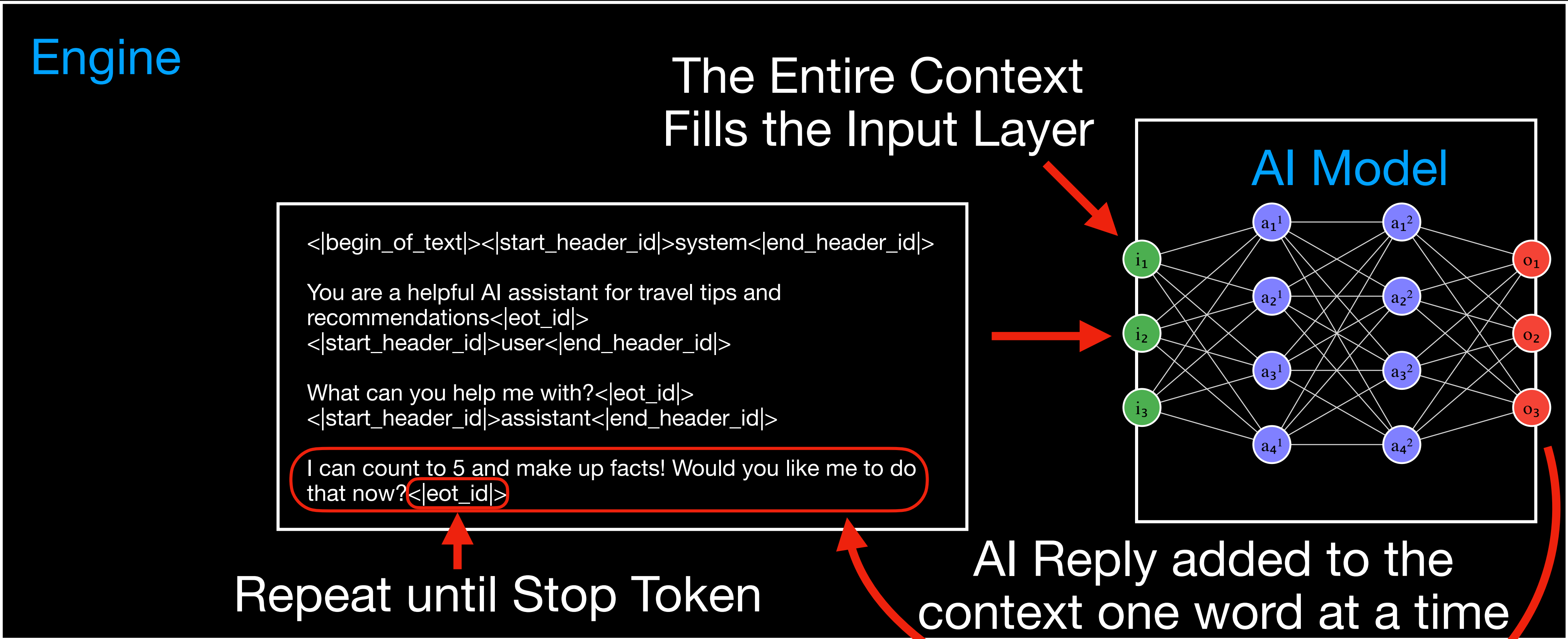
Llama Prompt Format
aka Prompt Template

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|>  
You are a helpful AI assistant for travel tips and  
recommendations<|eot_id|>  
<|start_header_id|>user<|end_header_id|>  
What can you help me with?<|eot_id|>  
<|start_header_id|>assistant<|end_header_id|>
```

Cue
the AI



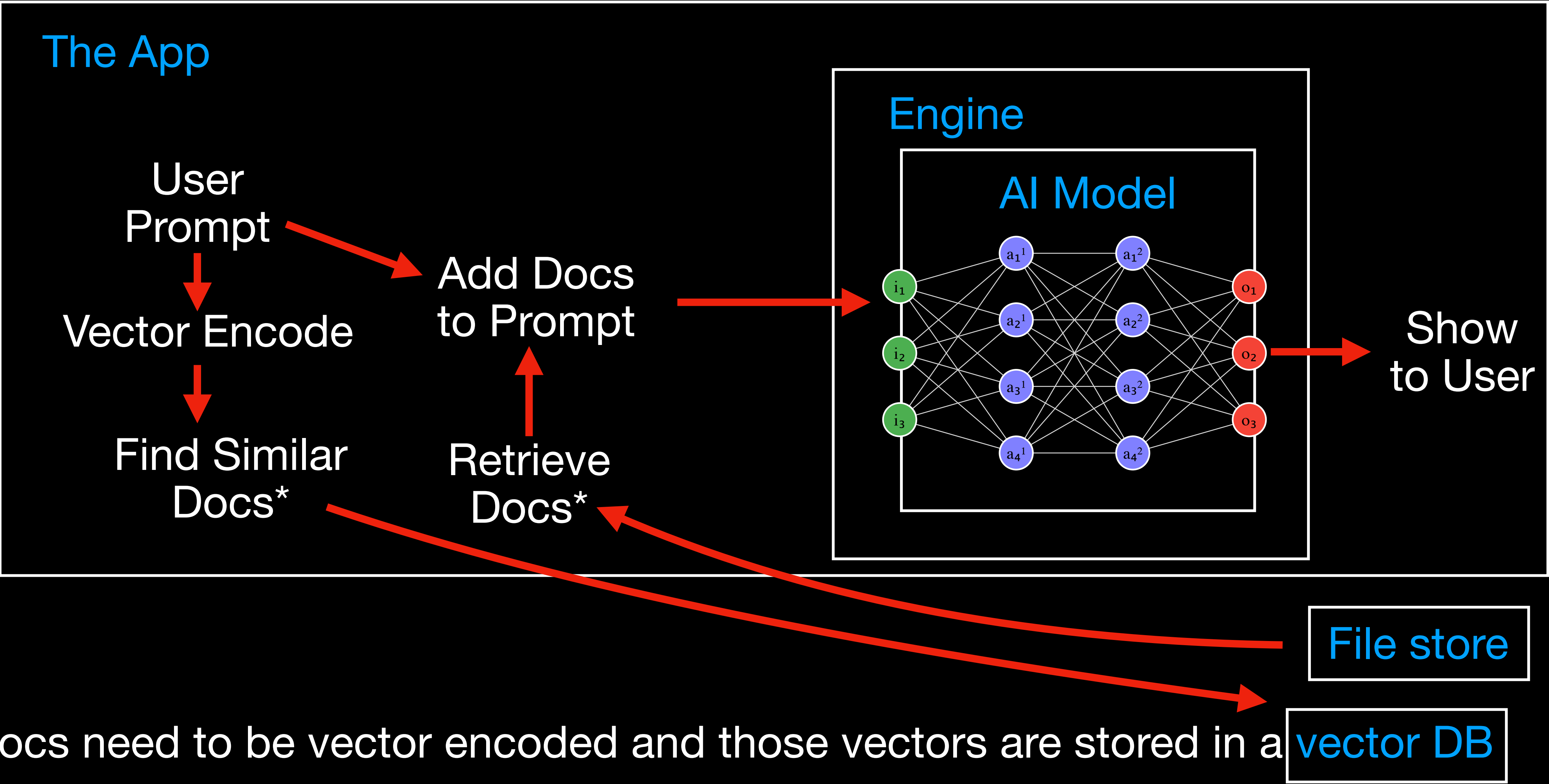
The AI Replies



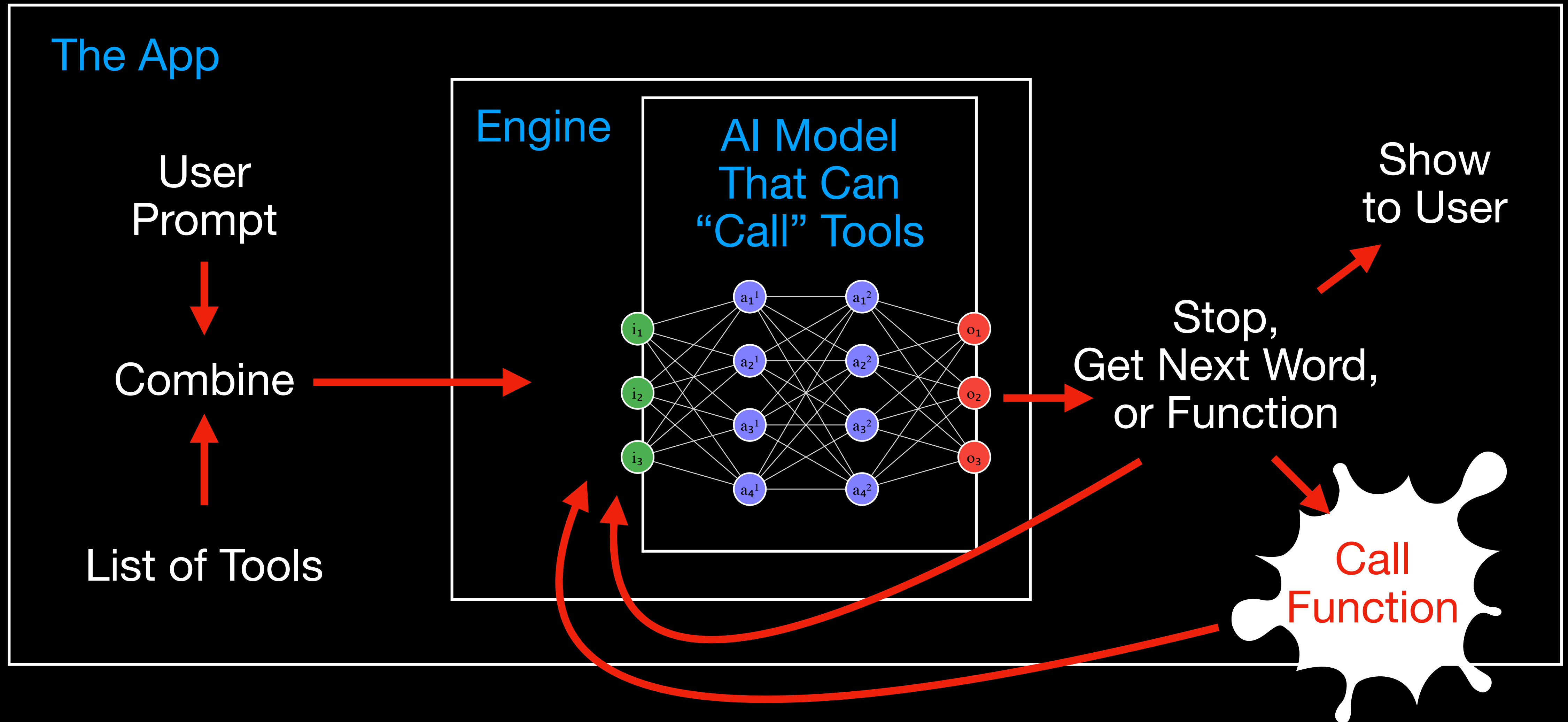
Context Limitations

- Soft maximum size
 - When it's above this limit it forgets stuff at the beginning or middle
- Hard maximum size
 - If it's too big it just gets cut off
- Mitigations
 - Rolling context window (forgets the beginning of the chat)
 - Summarize the past chat history (takes extra time to summarize)

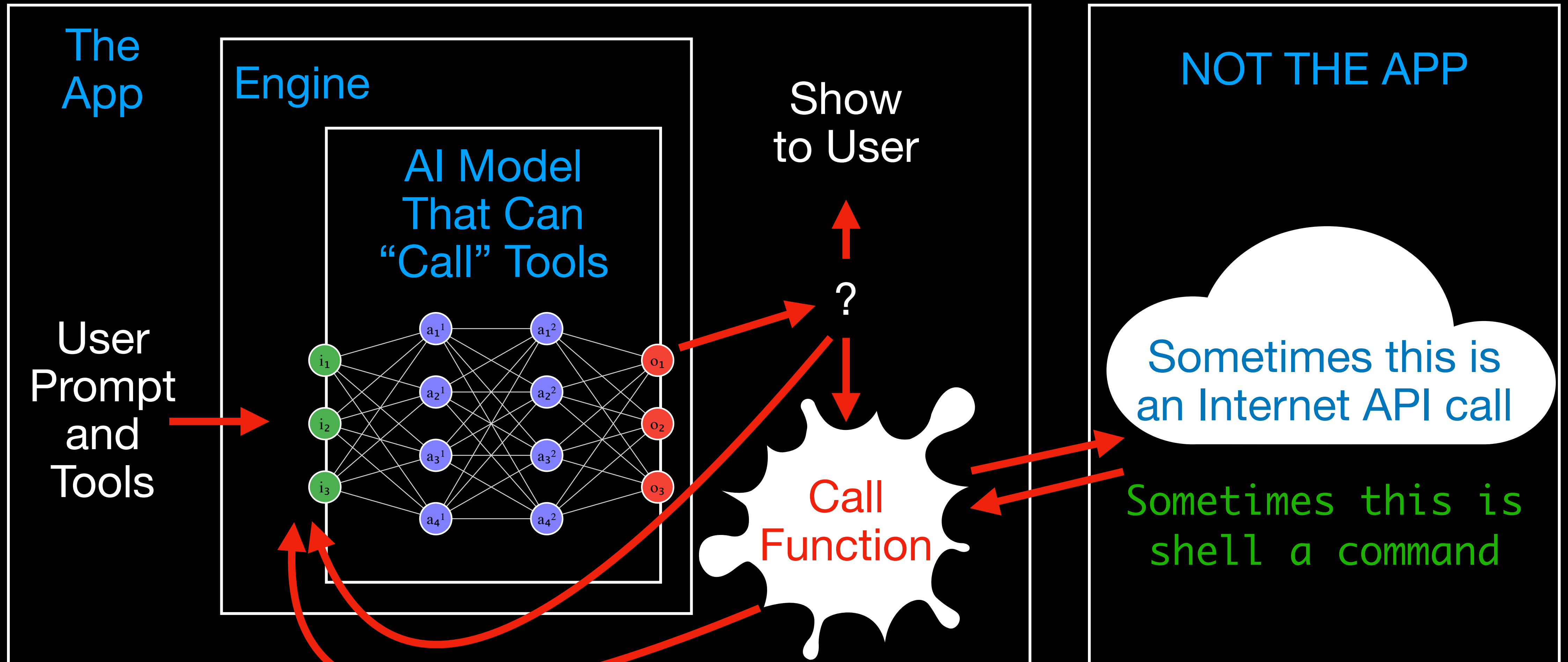
Retrieval Augmented Generation (RAG)



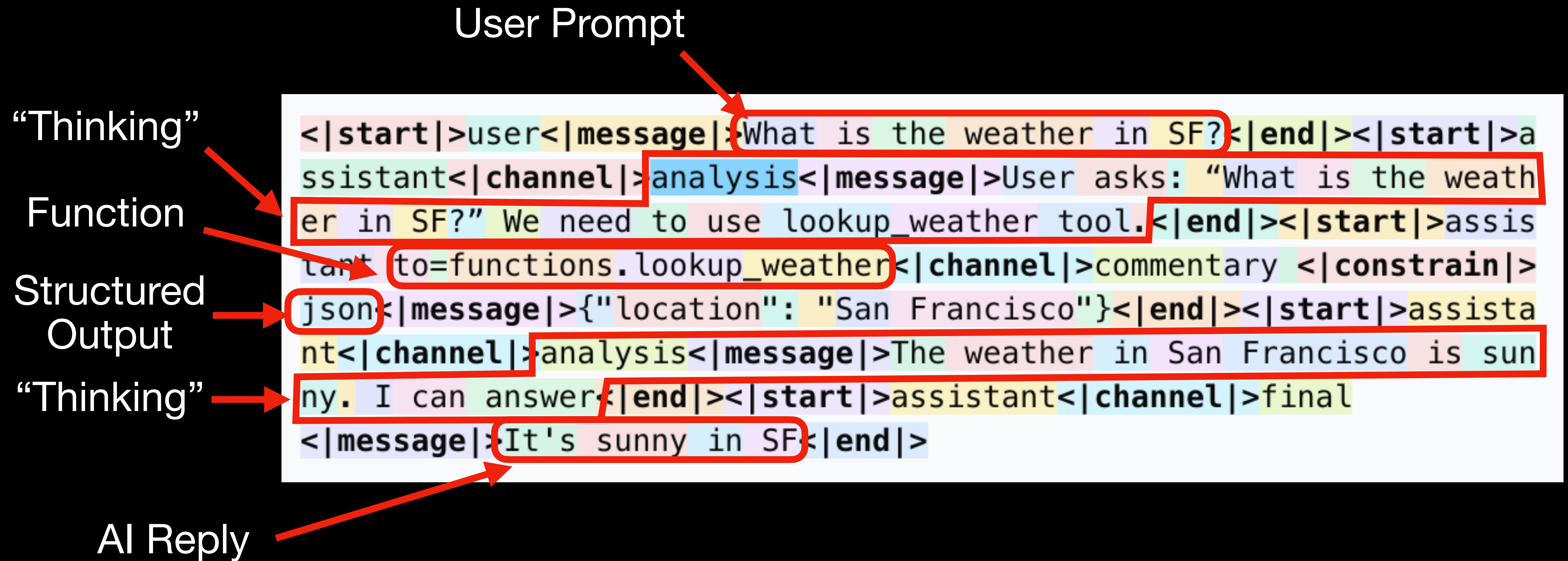
What Else Can We Do Here?



About That Function Call...



OpenAI's Harmony Template (gpt-oss)



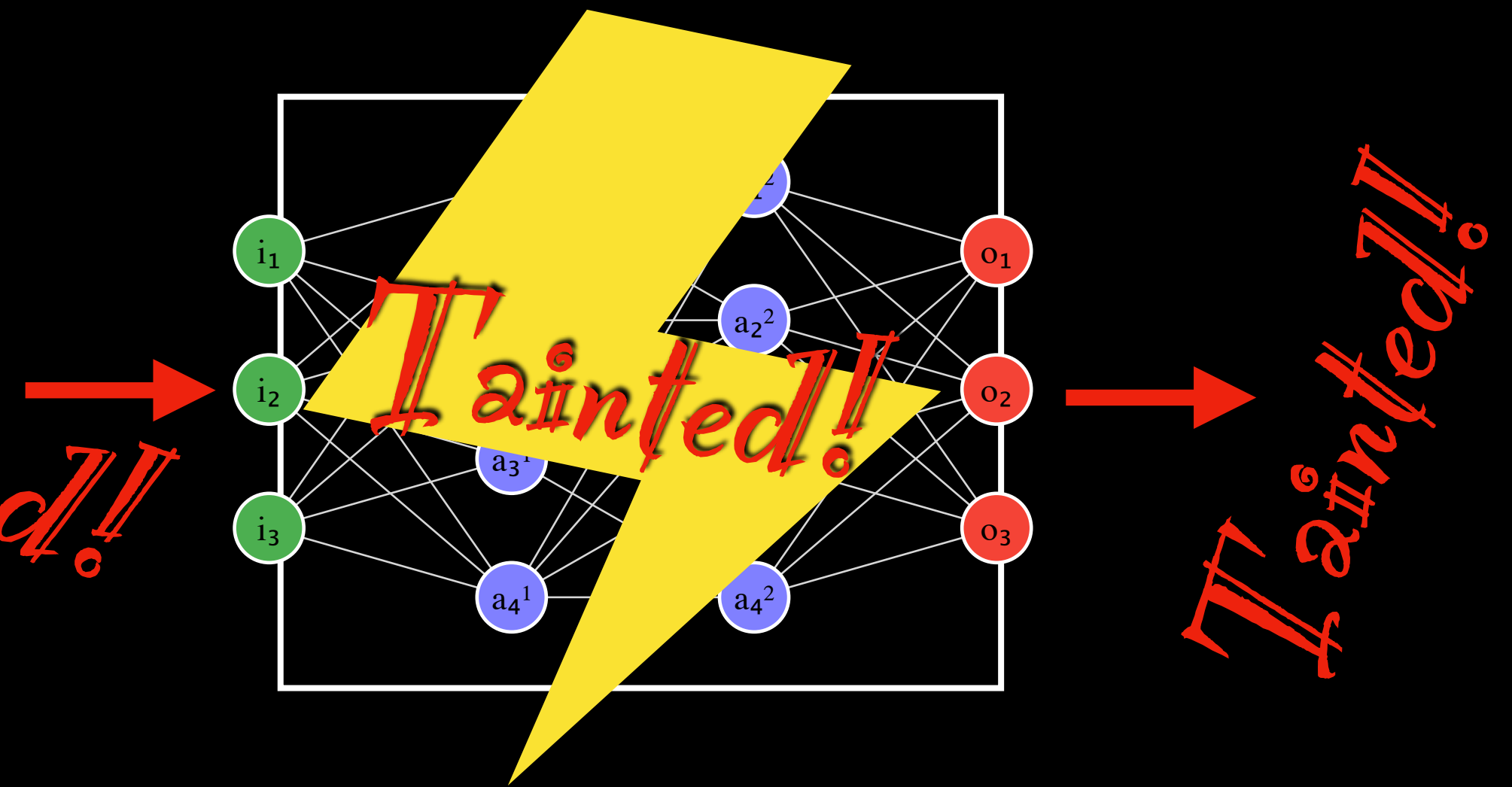
<https://github.com/openai/harmony>

It's All Tainted

Tainted!!

```
<|start|>user<|message|>What is the weather in SF?<|end|><|start|>assistant<|channel|>analysis<|message|>user asks: "What is the weather in SF?" We need to use lookup_weather tool.<|end|><|start|>assistant to=functions.lookup_weather<|channel|>commentary <|constrain|>json<|message|>{"location": "San Francisco"}<|end|><|start|>assistant<|channel|>analysis<|message|>The weather in San Francisco is sunny. I can answer<|end|><|start|>assistant<|channel|>final<|message|>It's sunny in SF<|end|>
```

Tainted!!

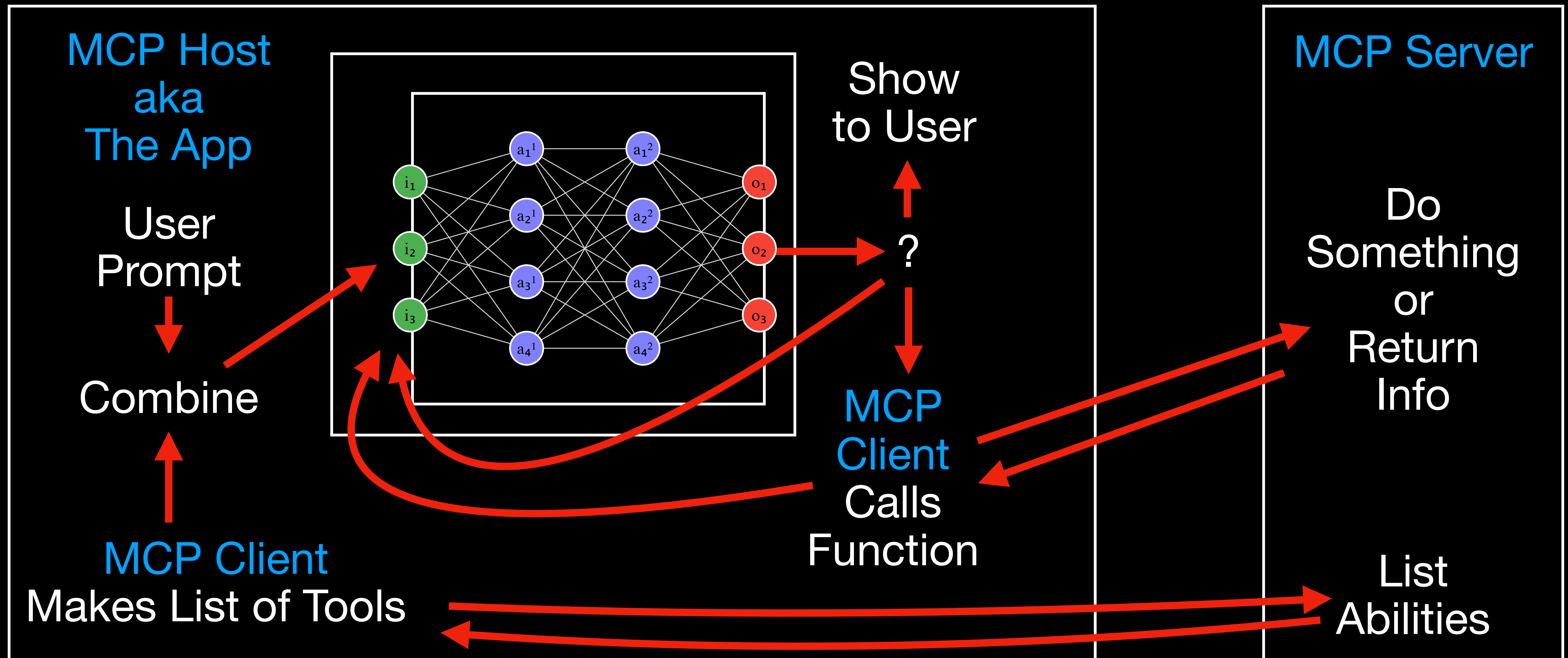


Each of the `<|...|>` tokens tell the LLM what to do or it's a stop token and gives the app a chance to intervene.

This works because of **model training**.
It can theoretically be **jailbroken** because it's **tainted**!

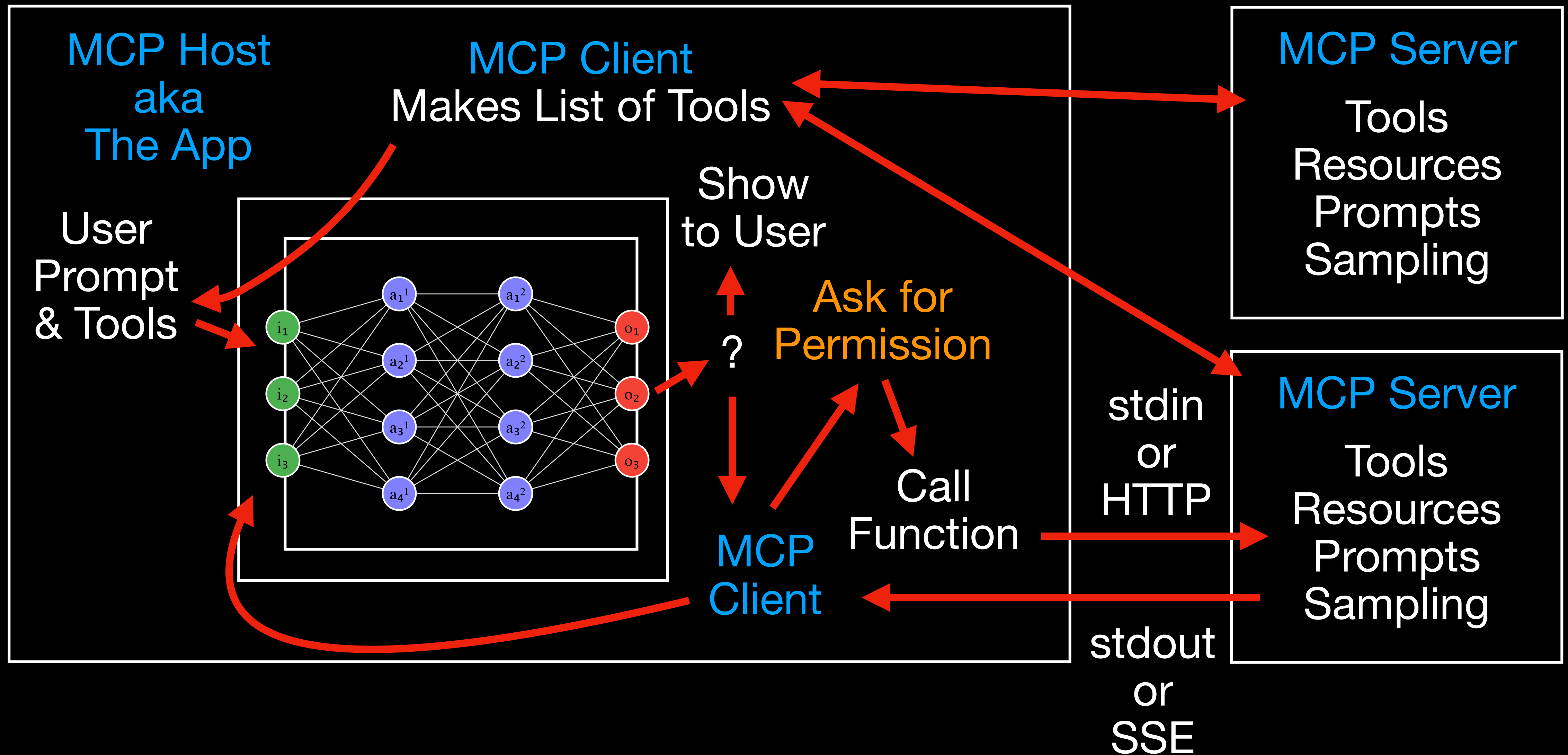
https://en.wikipedia.org/wiki/Taint_checking

Anthropic's Model Context Protocol (MCP)

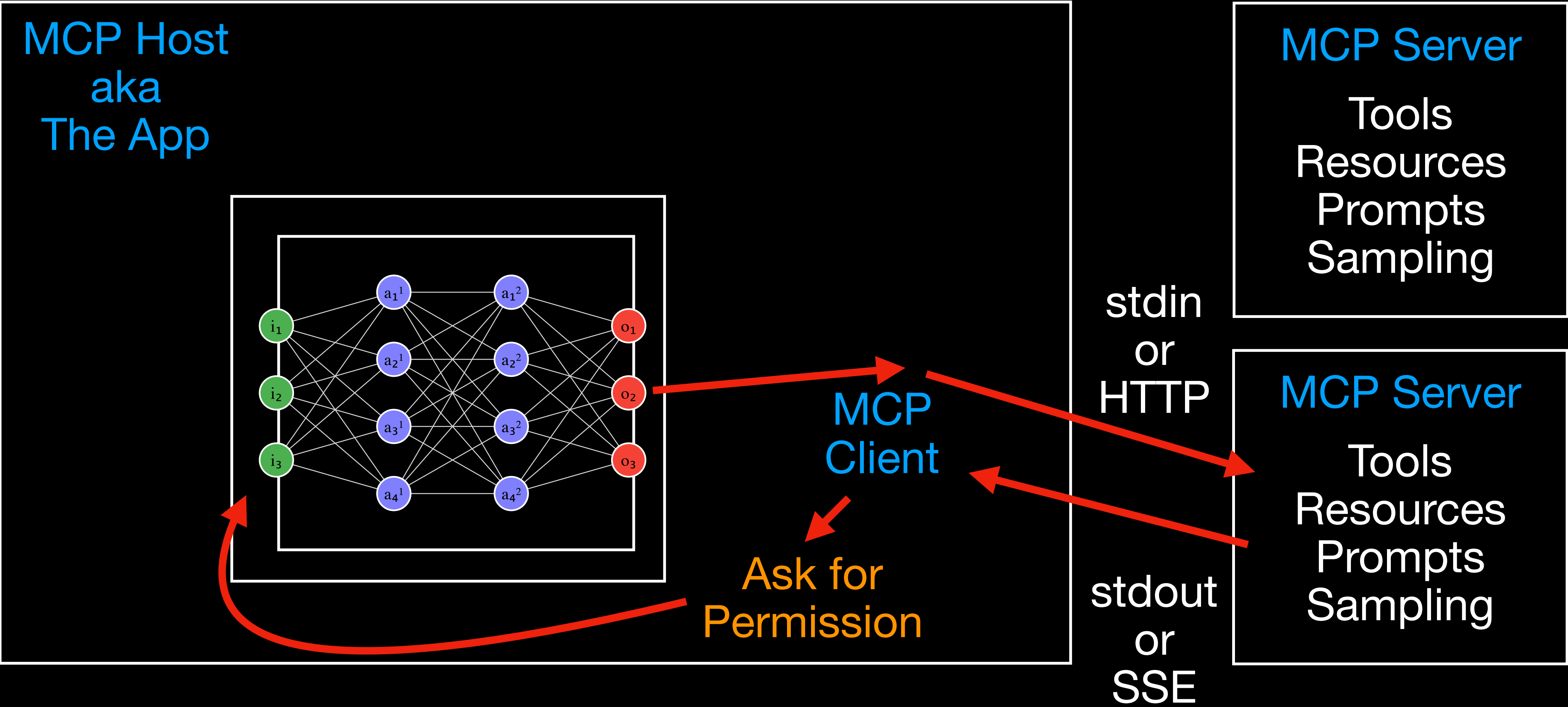


<https://huggingface.co/learn/mcp-course>

Anthropic's Model Context Protocol (MCP)



MCP Sampling



**Yes, MCP can be combined with RAG
but I'm not going to diagram it**

Demo

Questions

Slides Will Be At magnusviri.com