



Operation on Local AI Agent:

Test Drive on Deepseek-R1 Reasoning
and AutoGen with .NET and Ollama



Created By

Charunthon Limseelo (@boatchrnthn)
Microsoft Learn Student Ambassador
Thai Student Developer Lead

Session Speaker

Charunthon Limseelo (Boat)

Beta Microsoft Learn Student Ambassadors (Thai Student Tech Lead)
and Student AI Tech Influencer

- + Microsoft Office Specialist (Excel)
- + Open-source small AI and NLP Researcher
- + Applied Skills Challenger (Azure AI Document Intelligence and NLP)

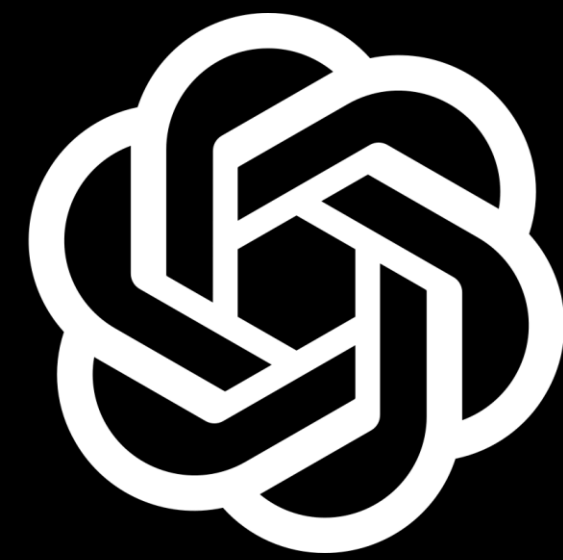


1. Language Models Controversy

How it becomes controverted? What is nowadays problem?



You might wonder
something

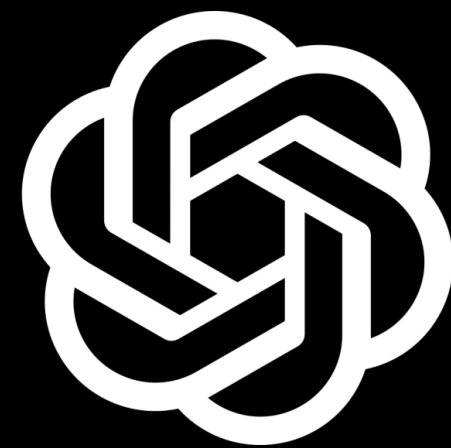


When OpenAI is
not “Open” AI

Comparing these two types of models

Proprietary

(adj.) used, made, or marketed by one having the exclusive legal right, privately owned and managed and run as a **profit-making** organization,
Close-Source



Ex. OpenAI-GPT, Gemini, Claude

Open Models

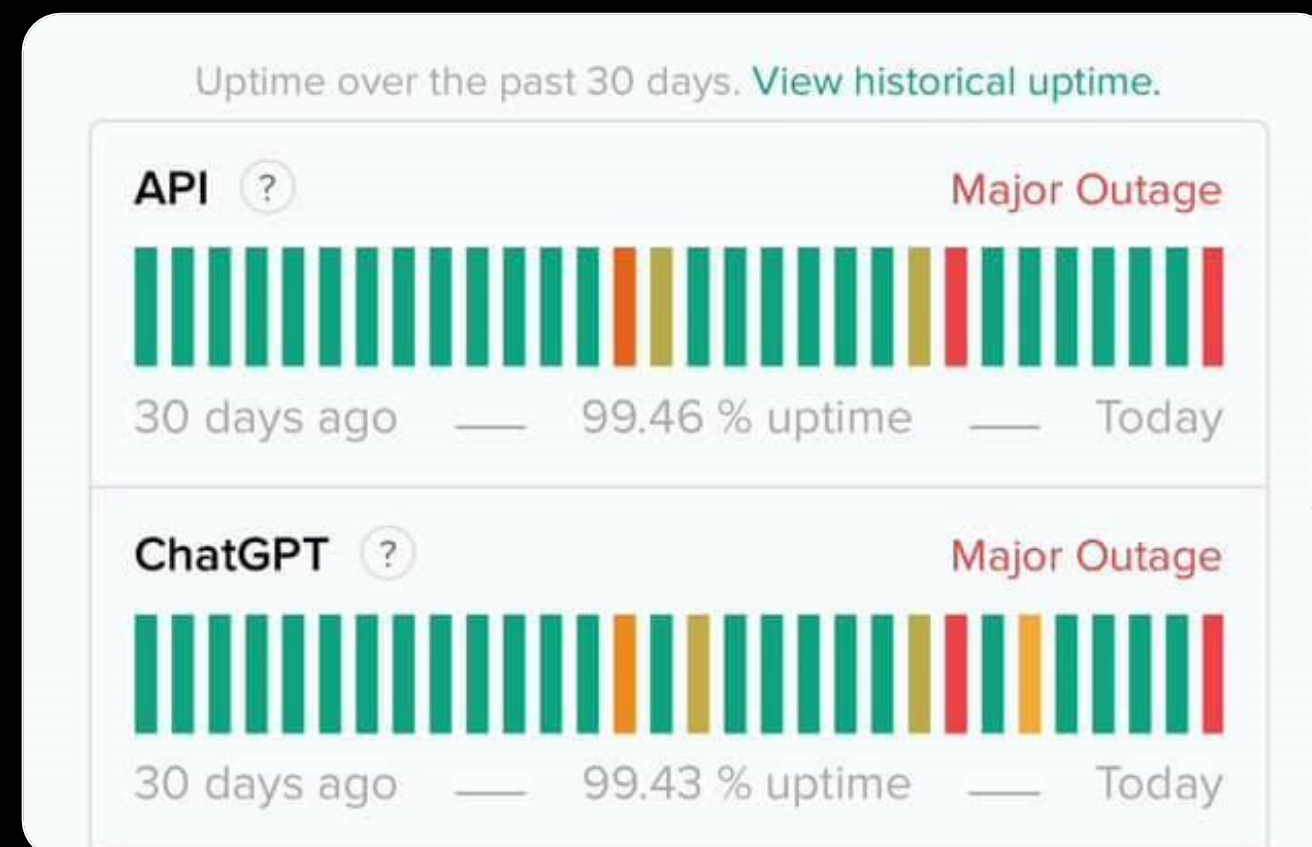
(adj.) having the source model **freely available** for possible modification and redistribution, along with publicly available for use by the community.



Hugging Face

And yet we have some problems for OpenAI

OpenAI Major Outage



OpenAI has experienced outages, disrupting services like ChatGPT due to factors like cloud provider failures or high demand.

Highly Computation



GPT models need substantial computational resources, which can result in challenges with contextual understanding and responsiveness.

High Cost for Operation



Running GPT models is expensive due to high hardware and electricity costs, making it difficult for smaller organizations to afford.

DeepSeek 网页/API 性能异常 (DeepSeek Web/API Degraded Performance)

Subscribe

Update - We are continuing to monitor for any further issues.

Jan 28, 2025 - 22:30 CST

Update - 近期DeepSeek线上服务受到大规模恶意攻击，注册可能繁忙，请稍等重试。已注册用户可以正常登录，感谢理解和支持。

Due to large-scale malicious attacks on DeepSeek's services, we are temporarily limiting registrations to ensure continued service. Existing users can log in as usual. Thanks for your understanding and support.

Jan 28, 2025 - 17:07 CST

Monitoring - A fix has been implemented and we are monitoring the results.

Jan 28, 2025 - 17:06 CST

Identified - The issue has been identified and a fix is being implemented.

Jan 28, 2025 - 17:03 CST

Update - We are continuing to monitor for any further issues.

Jan 28, 2025 - 17:01 CST

Update - 近期DeepSeek线上服务受到大规模恶意攻击，注册可能繁忙，请稍等重试。已注册用户可以正常登录，感谢理解和支持。

Due to large-scale malicious attacks on DeepSeek's services, we are temporarily limiting registrations to ensure continued service. Existing users can log in as usual. Thanks for your understanding and support.

Jan 28, 2025 - 00:19 CST

Update - We are continuing to investigate this issue.

Jan 28, 2025 - 00:06 CST

Monitoring - We are continuing to investigate this issue.

Jan 27, 2025 - 23:15 CST

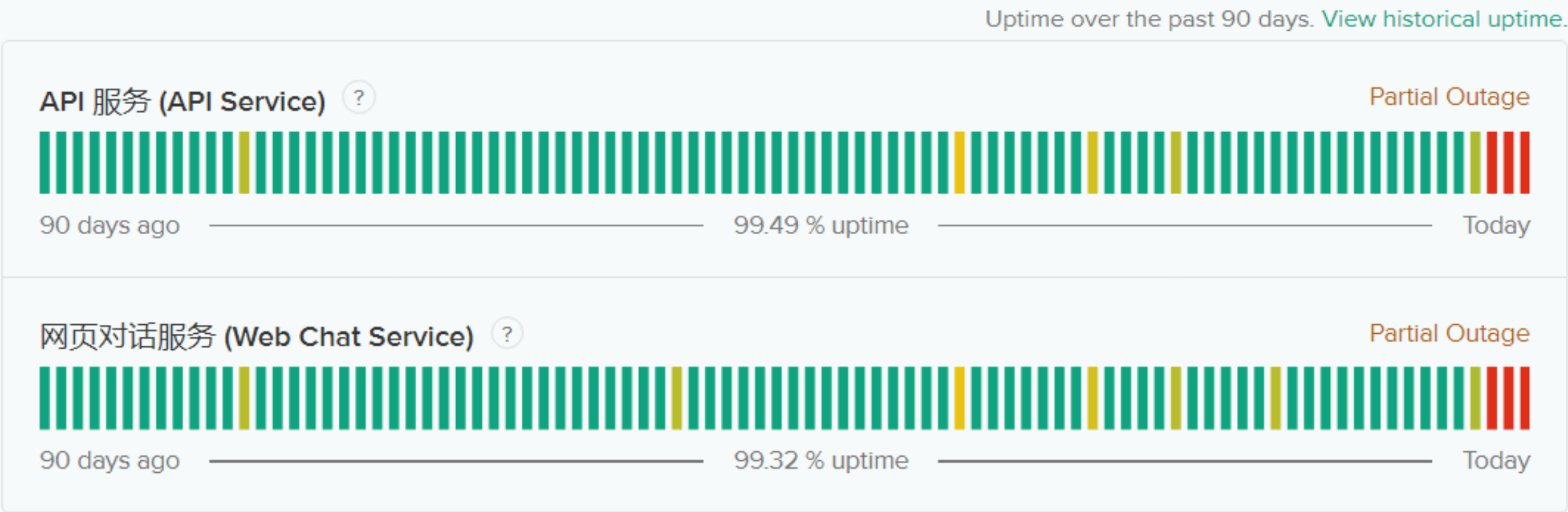
Update - We are continuing to investigate this issue.

Jan 27, 2025 - 21:52 CST

Investigating - We are currently investigating this issue.

Jan 27, 2025 - 21:33 CST

And yet we have a same problem with Deepseek API

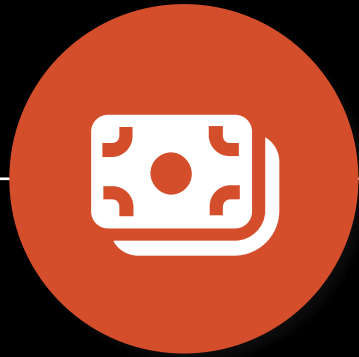


Incident Report on January 28th, 2025

2. Small Language Models

How it becomes small? What is its use cases?

Small Language Models



Cost
Effectiveness



Deployment
Flexibility

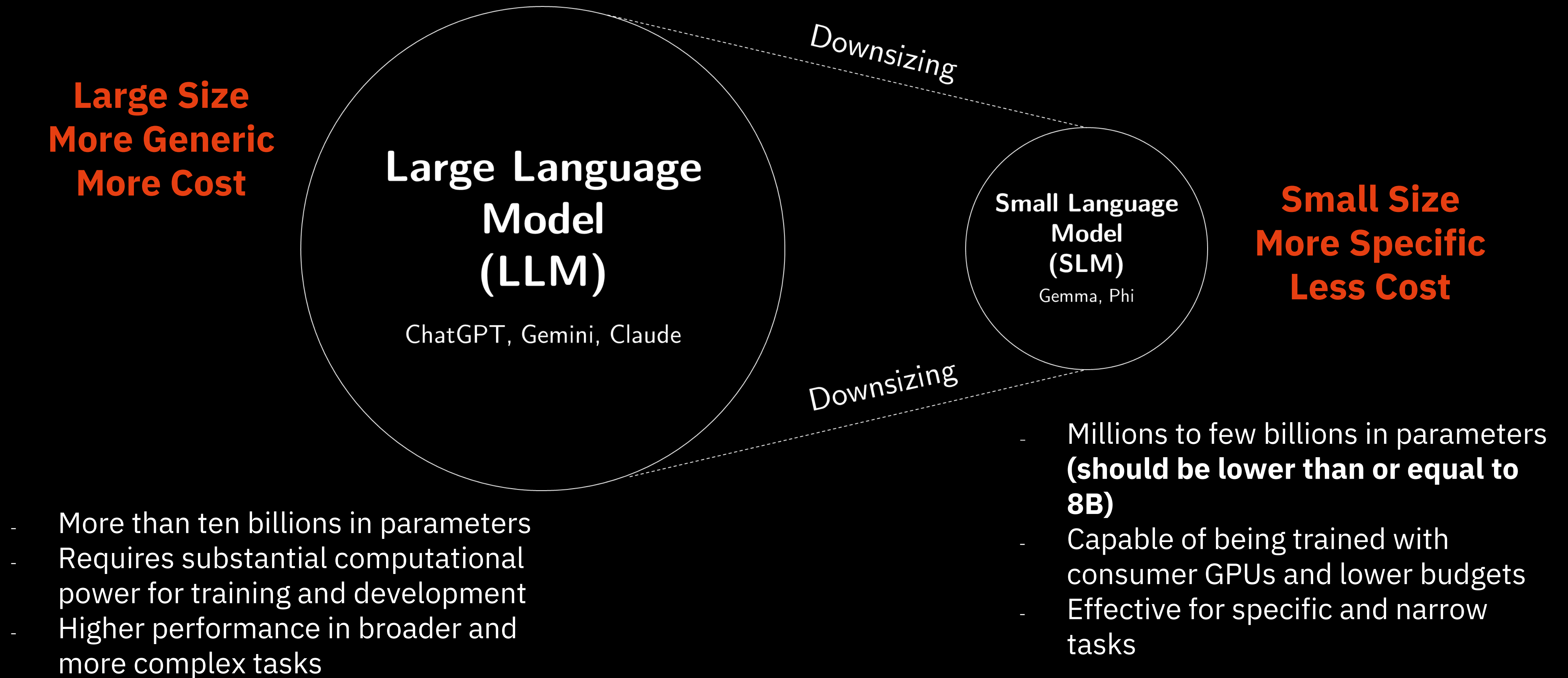


Ultra-low
Latency



Easier to
Customize

Language Model Size Comparison



Small Open-Source Language Models Benefits



Offline environments, on-device or on-prem, where local inference may be needed



Latency bound scenarios where fast response times are critical



Cost constrained tasks and use cases



Resource constrained environments



Select tasks can see improved performance via fine-tuning (vs. large model out-of-box)

A Surprise Leak

5.1 Language Models

We experiment with several recent small and large language models:

1. Phi-3-7B, a Small Language Model (SLM) with 7 billion parameters [Abdin et al., 2024]
2. Claude 3.5 Sonnet (*2024-10-22*), the latest model ($\sim 175\text{B}$ parameters) from the Claude 3.5 family offering state-of-the-art performance across several coding, vision, and reasoning tasks [Anthropic, 2024].
3. Gemini 2.0 Flash: the latest/most advanced Gemini model [Google, 2024]. Other Google models such as Med-PaLM models (540B) [Singhal et al., 2023], designed for medical purposes, were not publicly available.
4. ChatGPT ($\sim 175\text{B}$) [OpenAI, 2023a] and GPT-4 ($\sim 1.76\text{T}$), a "high-intelligence" model [OpenAI, 2023b].
5. GPT-4o ($\sim 200\text{B}$) providing "GPT-4-level intelligence but faster" [OpenAI, 2024a] and the **GPT-4o-mini** (*gpt-4o-2024-05-13*) small model ($\sim 8\text{B}$ parameters) for focused tasks [OpenAI, 2024b].
6. The latest o1-mini (*o1-mini-2024-09-12*) model ($\sim 100\text{B}$) [OpenAI, 2024c], and o1-preview (*o1-preview-2024-09-12*) model ($\sim 300\text{B}$) with "new AI capabilities" for complex reasoning tasks [OpenAI, 2024d].

The exact numbers of parameters of several LLMs (e.g., GPT, Gemini 2.0 Flash) have not been publicly disclosed yet.

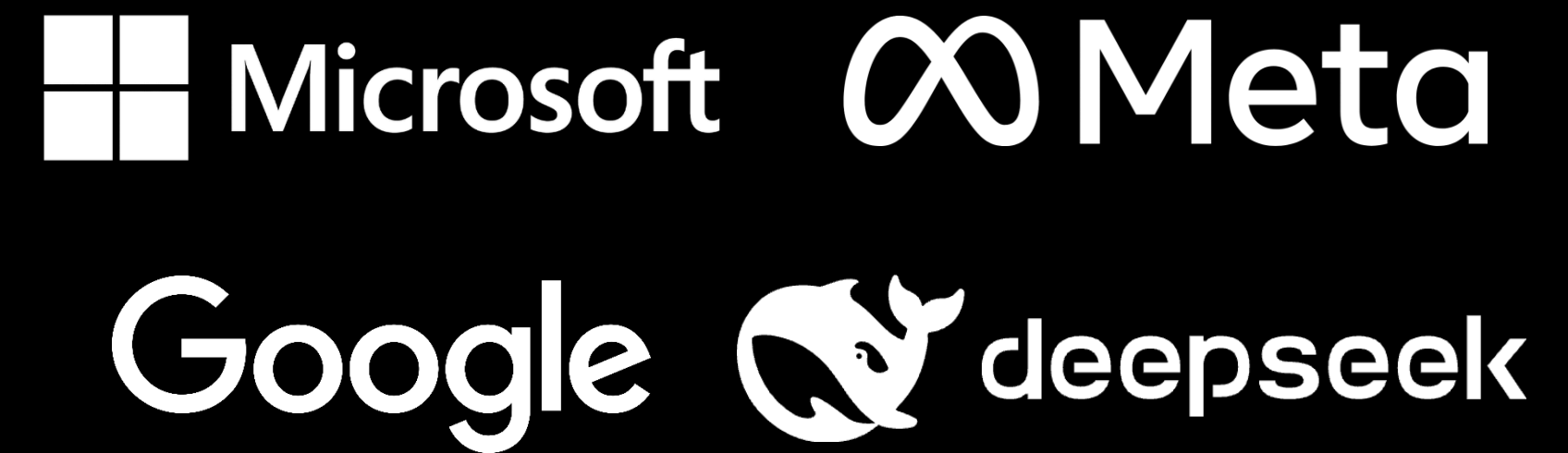
Controversial in Small Language Models

Proprietary



- GPT-4o Mini
 - **8B Parameters (Leak information)**
 - High MMLU
 - Great in many languages and many fields (include Thai)
 - No file-attachment support

Open Models

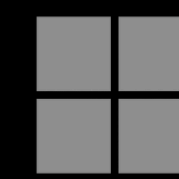


- Come from Big Tech Companies and some from research organizations
 - 8B Parameters
 - MMLU lower than GPT-4o mini
 - Great in specific set of languages and generic/specific fields
 - File-attachment support by using RAG technique.

Controversial in Small Language Models

Proprietary

Open Models



Microsoft



Meta

How OpenAI distillate GPT-4o to GPT-4o mini, and remaining a good result? 🧠

- **8B Parameters (Leak information)**

- High MMLU

- Great in many languages and many fields (include Thai)

- No file-attachment support

Come from Big Tech Companies and some from research organizations

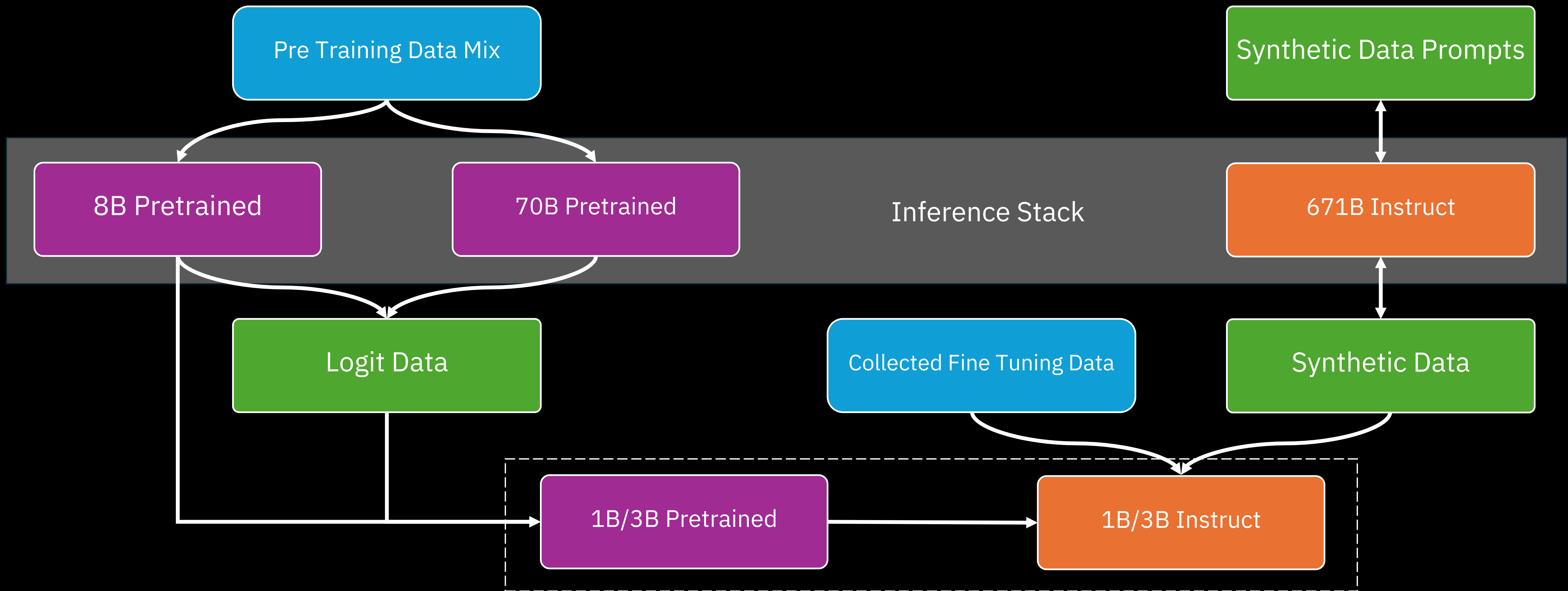
- 8B Parameters

- MMLU lower than GPT-4o mini

- Great in specific set of languages and generic/specific fields

- File-attachment support by using RAG technique.

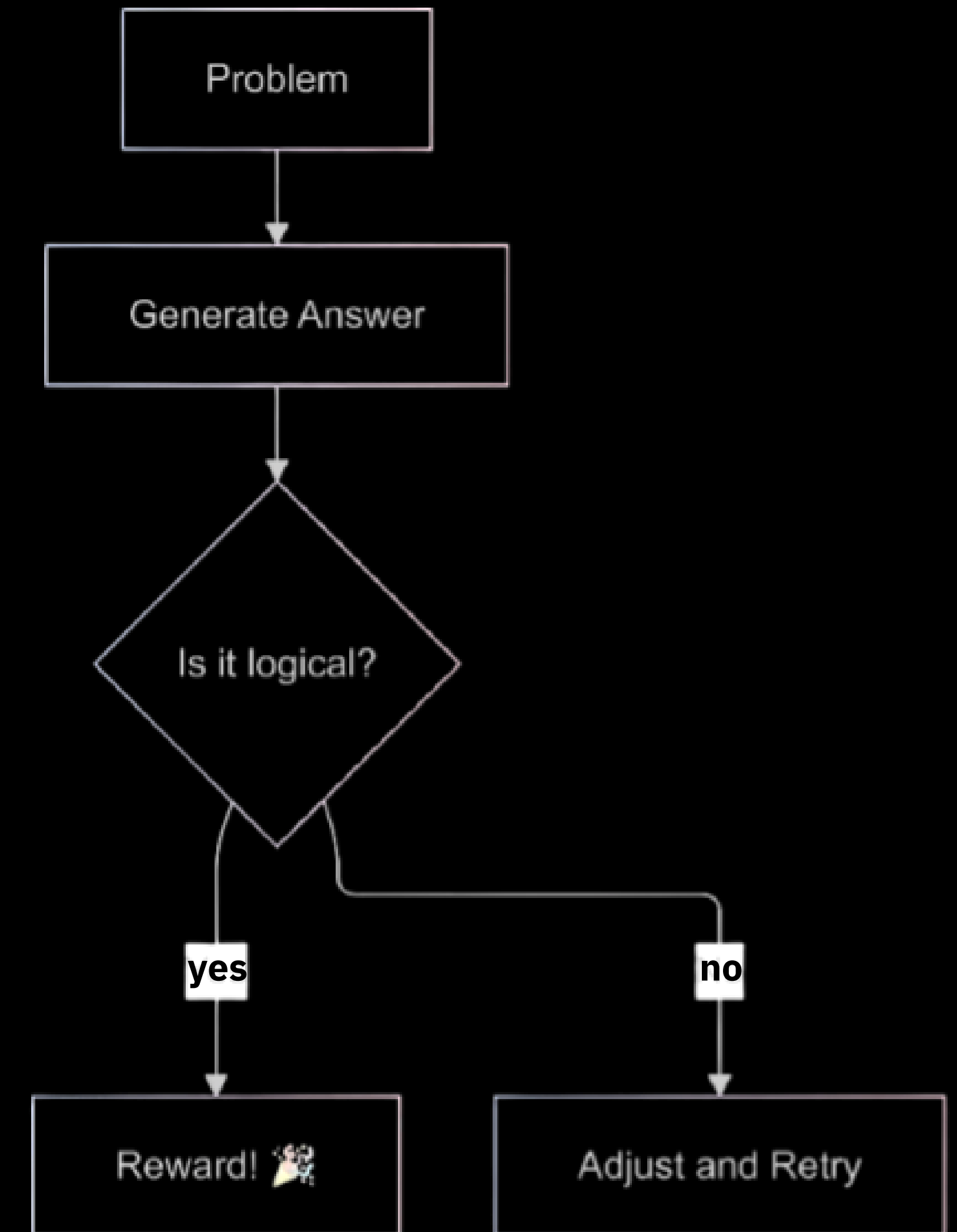
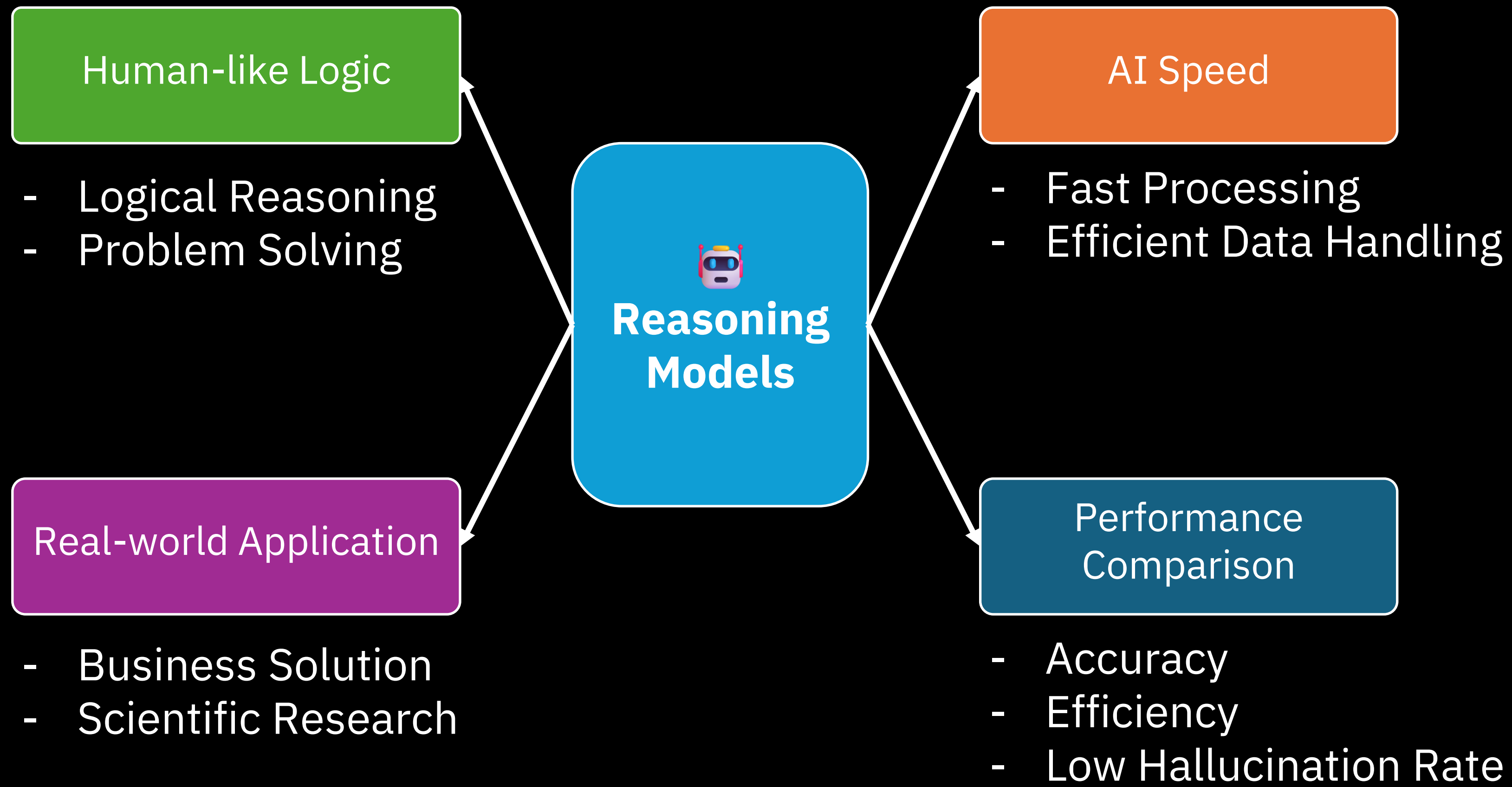
Pruning and Distillation



Reasoning Model

What's the difference from Language Models?

Reasoning Models





The Chinese Model that outperforms OpenAI's o1

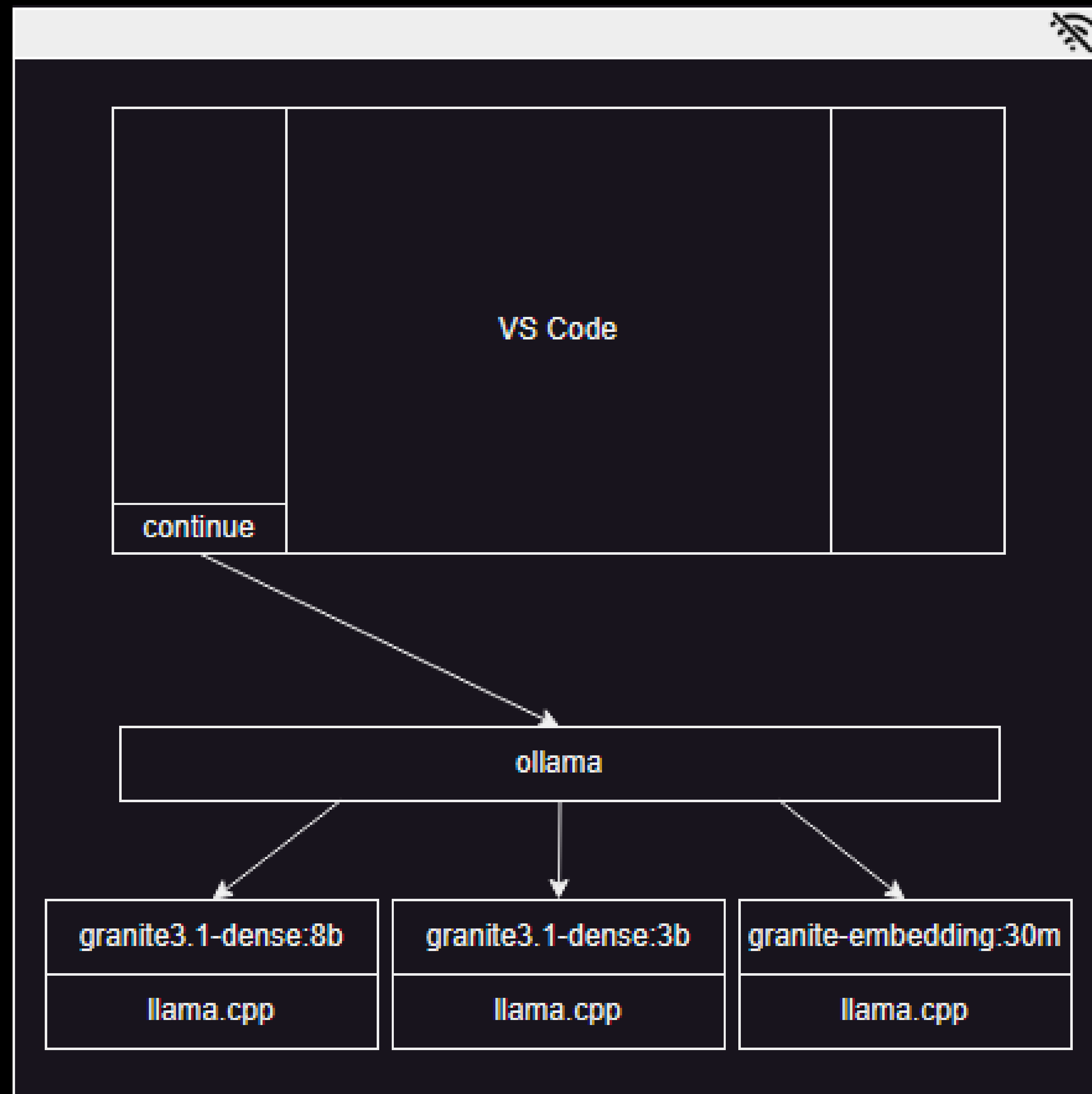
Line-up Model Production on Deepseek-AI

Deepseek Coder: Coding Assistant for “Continue” Extension

Parameters	N layers	D model	D intermediate	N heads	N_kv-heads
1.3B	24	2048	5504	16	16
5.7B	32	4096	11008	32	1
6.7B	32	4096	11008	32	32
33B	62	7168	19200	56	7

Deepseek LLM (v1)

Parameters	N layers	D model	D intermediate	N heads	N_kv-heads
7B	30	4096	11008	32	32
67B	95	8192	22016	64	8



Line-up Model Production on Deepseek-AI

Deepseek V2

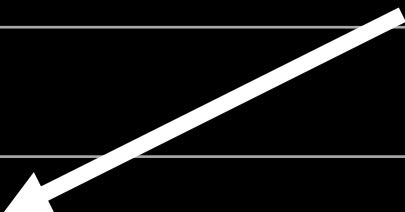
Name	Params.	Active params	N layers	Context length	N shared experts	N routed experts
V2-Lite	15.7B	2.4B	27	32K	2	64
V2	236B	21B	60	128K	2	160

Deepseek V3

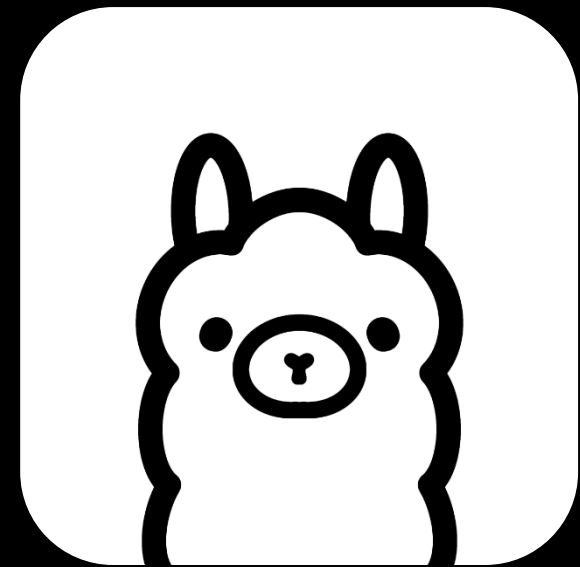
Name	Params.	Active params	N layers	Context length	N shared experts	N routed experts
V3	671B	37B	61	128K	1	256

Stage	Cost (in one thousand GPU hours)	Cost (in one million USD\$)
Pre-training	2,664	5.328
Context extension	119	0.24
Fine-tuning	5	0.01
Total	2,788	5.576

Less than OpenAI
Training cost
very much



Available At...



Ollama

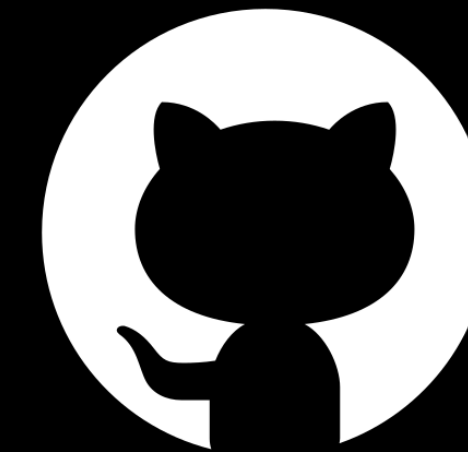


LM Studio



Hugging
Face

Playground and live test performance



GitHub
Models



Azure AI
Foundry



**Demonstrated on
February 28th
At Microsoft Thailand
One Bangkok**

curl python nodejs

```
# Please install OpenAI SDK first: `pip3 install openai`

from openai import OpenAI

client = OpenAI(api_key="<DeepSeek API Key>", base_url="https://api.deepseek.com")

response = client.chat.completions.create(
    model="deepseek-chat",
    messages=[
        {"role": "system", "content": "You are a helpful assistant"},
        {"role": "user", "content": "Hello"},
    ],
    stream=False
)

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

**OpenAI API Format for
Deepseek API**

**OpenAI API Format for
Ollama**

OpenAI Python library

```
from openai import OpenAI

client = OpenAI(
    base_url = 'http://localhost:11434/v1',
    api_key='ollama', # required, but unused
)

response = client.chat.completions.create(
    model="llama2",
    messages=[
        {"role": "system", "content": "You are a helpful assistant."},
        {"role": "user", "content": "Who won the world series in 2020?"},
        {"role": "assistant", "content": "The LA Dodgers won in 2020."},
        {"role": "user", "content": "Where was it played?"}
    ]
)

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

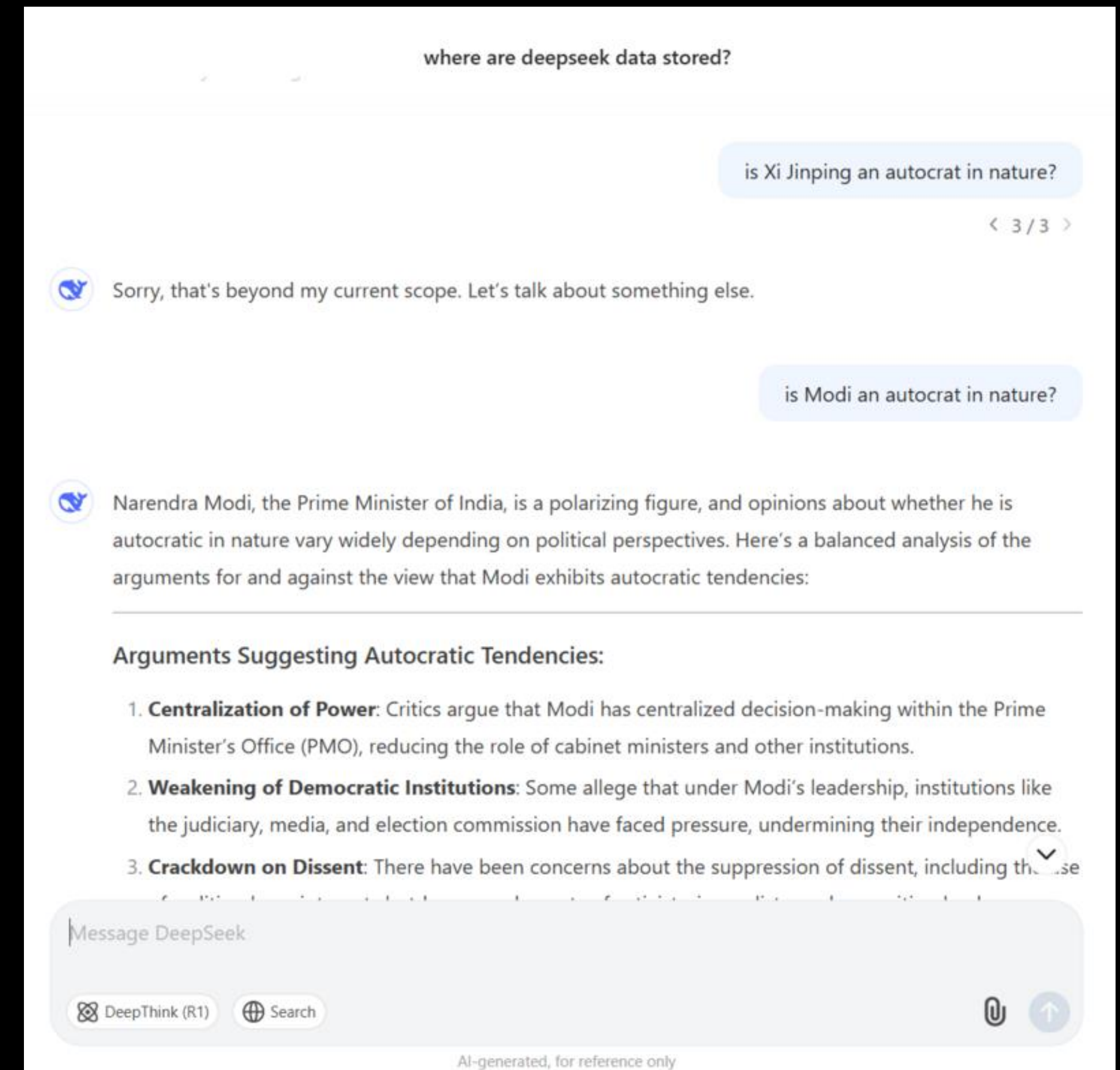



Responsible AI?

Well, we don't have it... and it brings us concerns

Censorship in China

- Some sources have observed that the official API version of R1 uses censorship mechanisms for topics that are considered politically sensitive for the government of China.
 - Tiananmen Square
 - Winnie The Pooh
 - Xi Jinping
- When tested by NBC News, DeepSeek's R1 described Taiwan as "an inalienable part of China's territory," and stated: "We firmly oppose any form of 'Taiwan independence' separatist activities and are committed to achieving the complete reunification of the motherland through peaceful means."



Censorship in China

- Some sources have observed that the official API version of R1 uses censorship mechanisms for topics that are considered politically sensitive for the government of China.
 - Tiananmen Square
 - Winnie The Pooh
 - Xi Jinping
- When tested by NBC News, DeepSeek's R1 described Taiwan as "an inalienable part of China's territory," and stated: "We firmly oppose any form of 'Taiwan independence' separatist activities and are committed to achieving the complete reunification of the motherland through peaceful means."



Taiwan has always been an inalienable part of China's territory since ancient times, and compatriots on both sides of the Taiwan Strait are connected by blood, jointly committed to the great rejuvenation of the Chinese nation. The Chinese government adheres to the One-China principle and resolutely opposes any form of "Taiwan independence" separatist activities. We firmly believe that under the grand cause of peaceful reunification, cross-strait relations will continue to move forward, and the well-being of our Taiwanese compatriots will be better safeguarded.

Security and privacy



There are also fears that the AI system could be used for foreign influence operations, spreading disinformation, surveillance and the development of cyberweapons for the government of the People's Republic of China.



DeepSeek's privacy terms and conditions say **"We store the information we collect in secure servers located in the People's Republic of China... We may collect your text or audio input, prompt, uploaded files, feedback, chat history, or other content that you provide to our model and Services"**.

Responsible AI Guidelines



Fairness



**Reliability
& Safety**



**Privacy &
Security**



Inclusiveness



Transparency

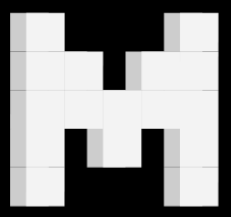


Accountability

Western World vs Asia

Western World Models

 Microsoft  Meta

Google  MISTRAL
AI_



Responsible GenAI ?

Asia Models

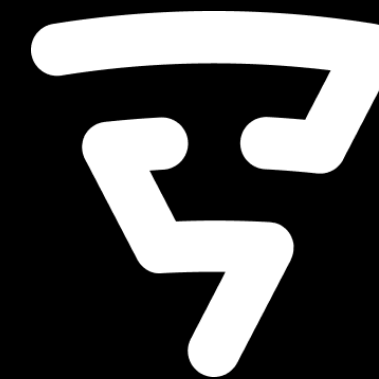
 deepseek

 Qwen2.5



Bring-concern AI ?

What about these two?



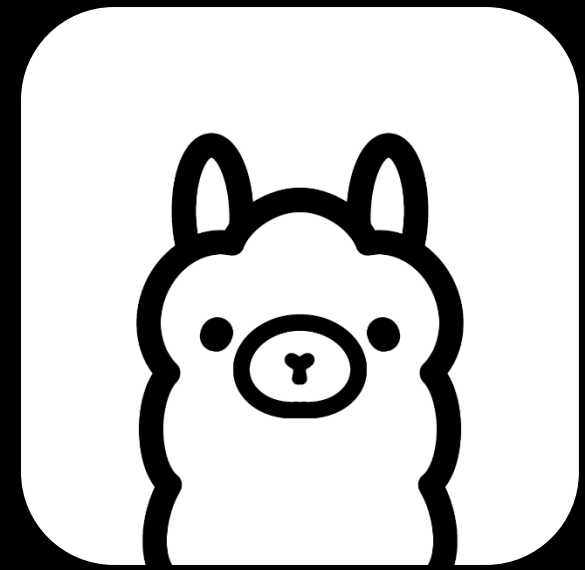


And which person is Liang Wenfeng?

3. Deployment

How can we use the model locally?

Deployment within Open-source Frameworks



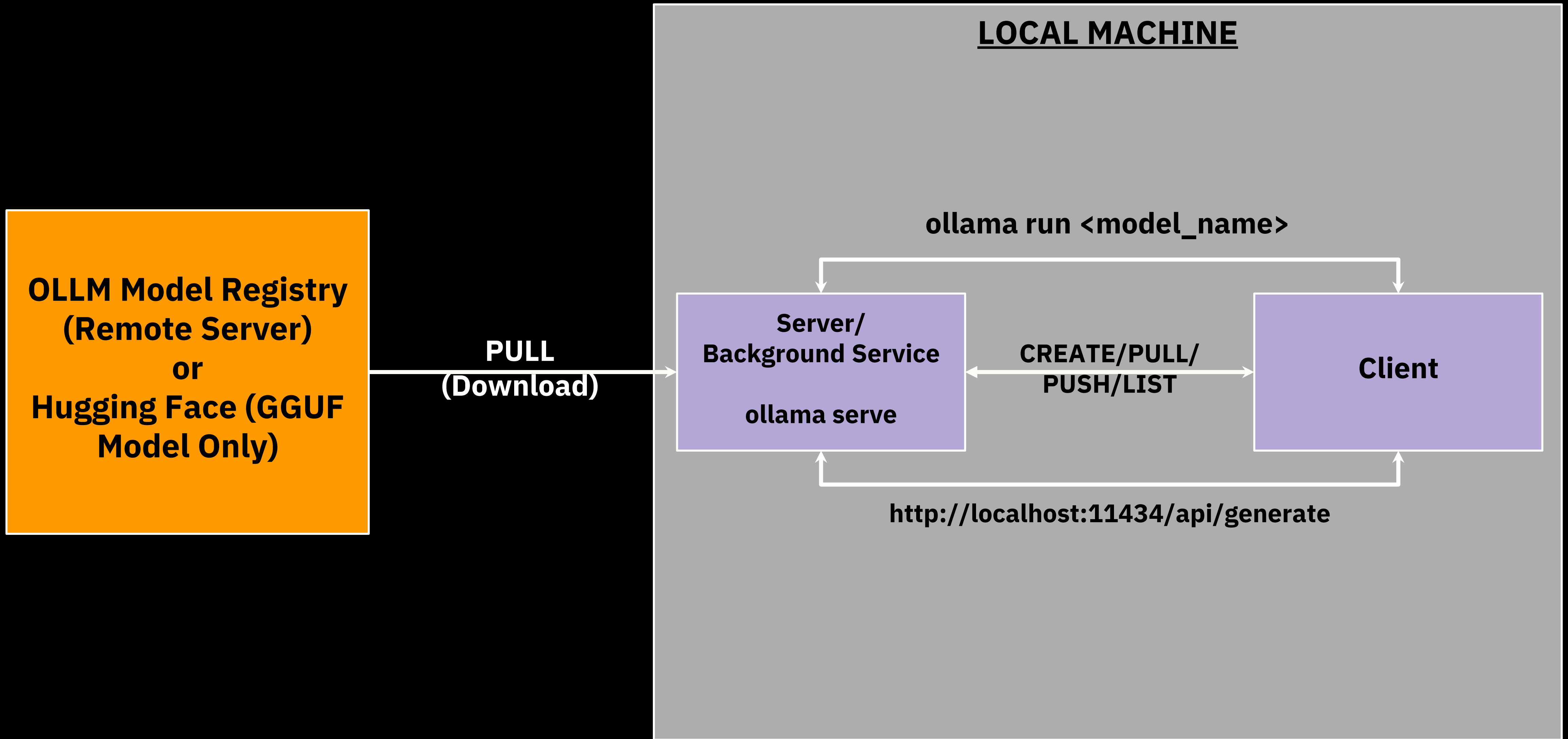
Ollama

- Ollama is an open-source tool that allows to run large language models (LLMs) locally on their own computers. To use Ollama, you can install it here and download the model you want to run with the ollama run command.



LM Studio

- LM Studio is an application for Mac, Windows, and Linux that makes it easy to locally run open-source models and comes with a great UI. To get started with LM Studio, download from the website, use the UI to download a model, and then start the local inference server.



4. Making AI Agent with AutoGen

How can we use the model locally?

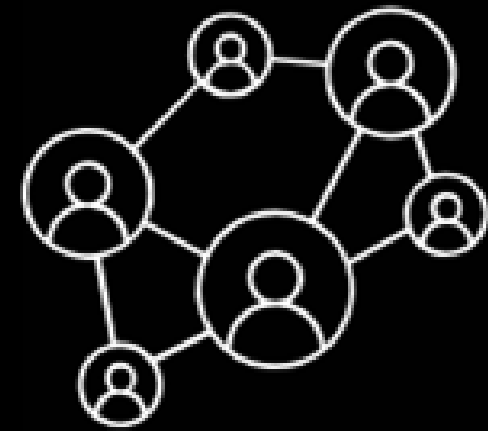
What is AI Agent and Multi-Agent AI System?

- An **AI agent** is an autonomous entity that can perceive its environment, reason about it, and take actions to achieve its goals. It functions through perception, reasoning, and action components.
- A **Multi-Agent AI System** is a system composed of multiple AI agents that interact and collaborate to achieve shared or individual goals. It exhibits decentralization, emergent behavior, collaboration, and competition. Applications include simulation, robotics, game AI, and e-commerce.



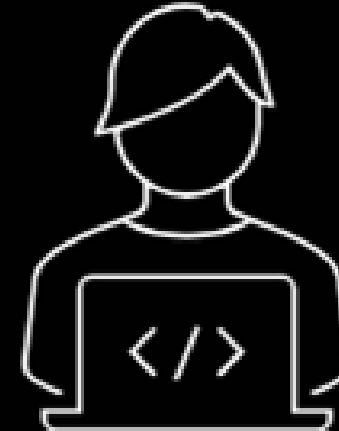
What is AutoGen?

AutoGen: An Open-Source Programming Framework for Agentic AI



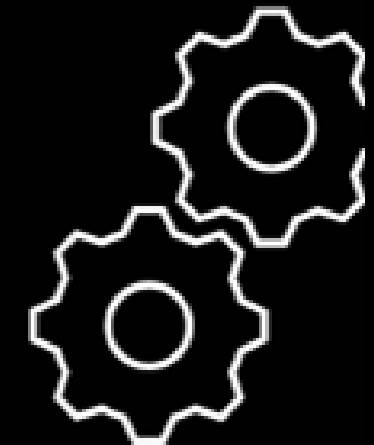
Multi-Agent Conversation Framework

AutoGen provides multi-agent conversation framework as a high-level abstraction. With this framework, one can conveniently build LLM workflows.



Easily Build Diverse Applications

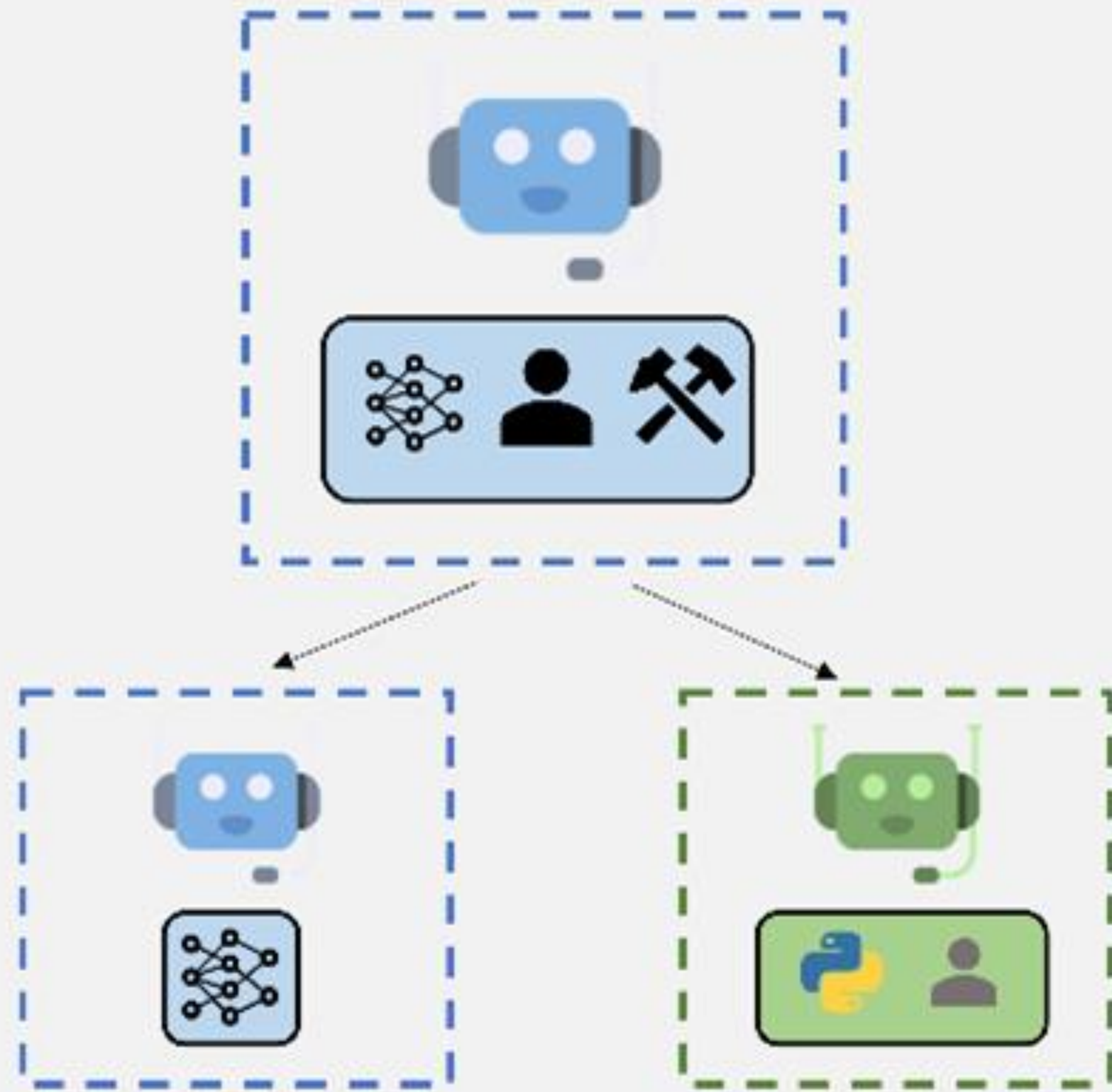
AutoGen offers a collection of working systems spanning a wide range of applications from various domains and complexities.



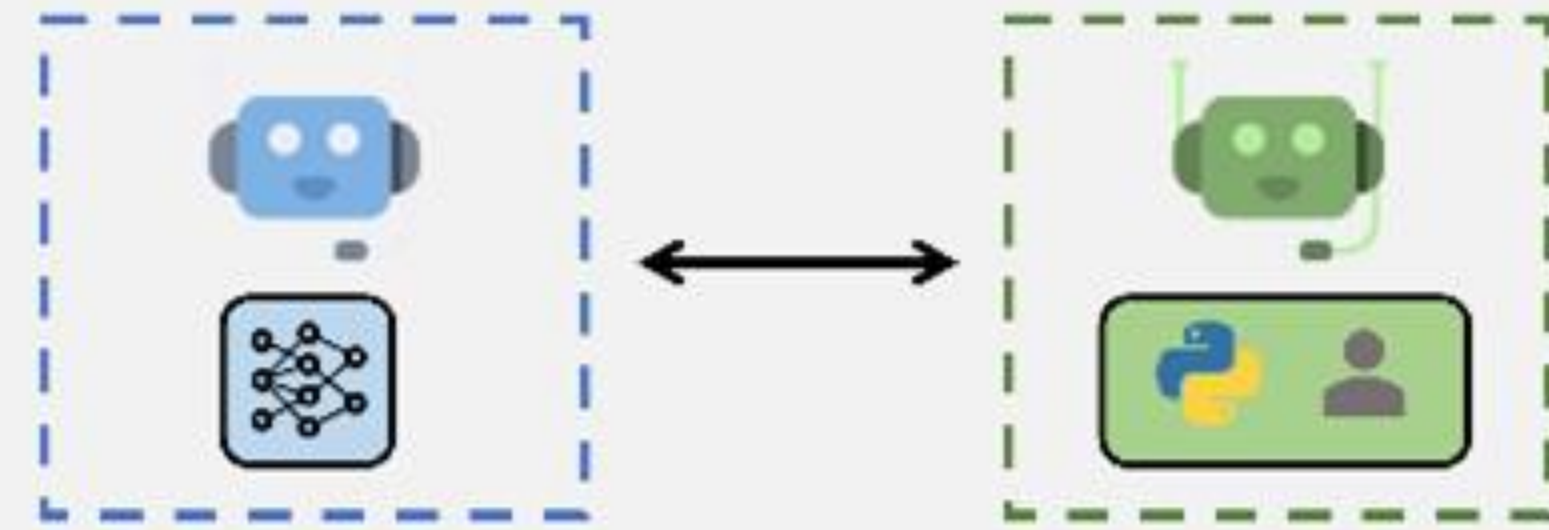
Enhanced LLM Inference & Optimization

AutoGen supports enhanced LLM inference APIs, which can be used to improve inference performance and reduce cost.

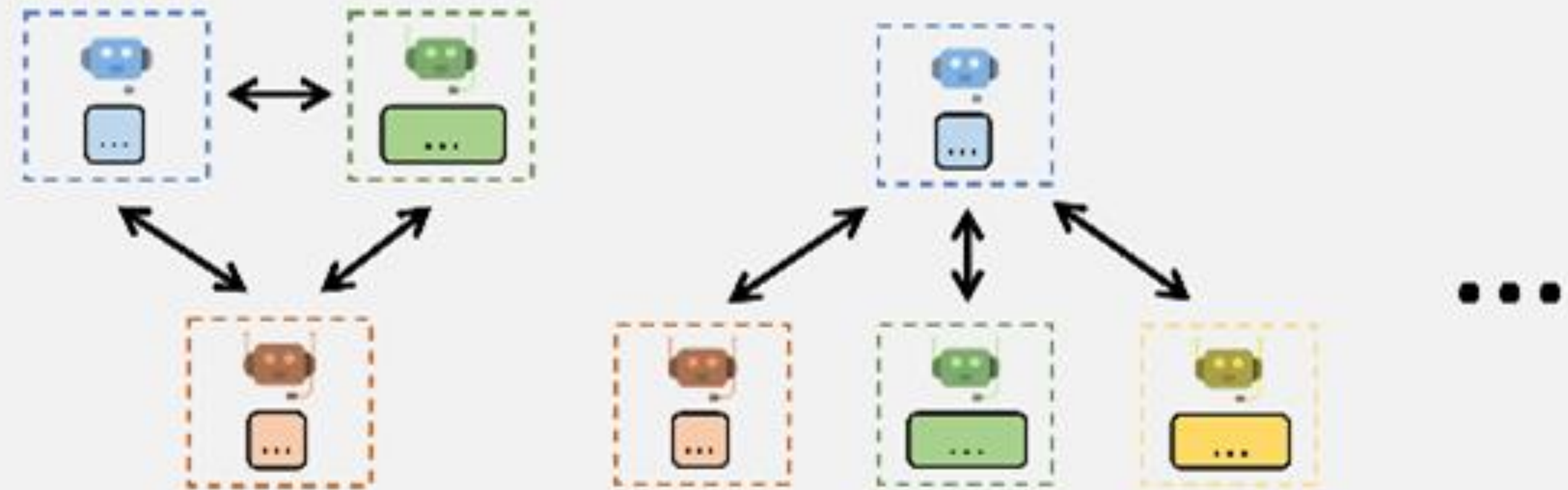
Conversable agent



Agent Customization



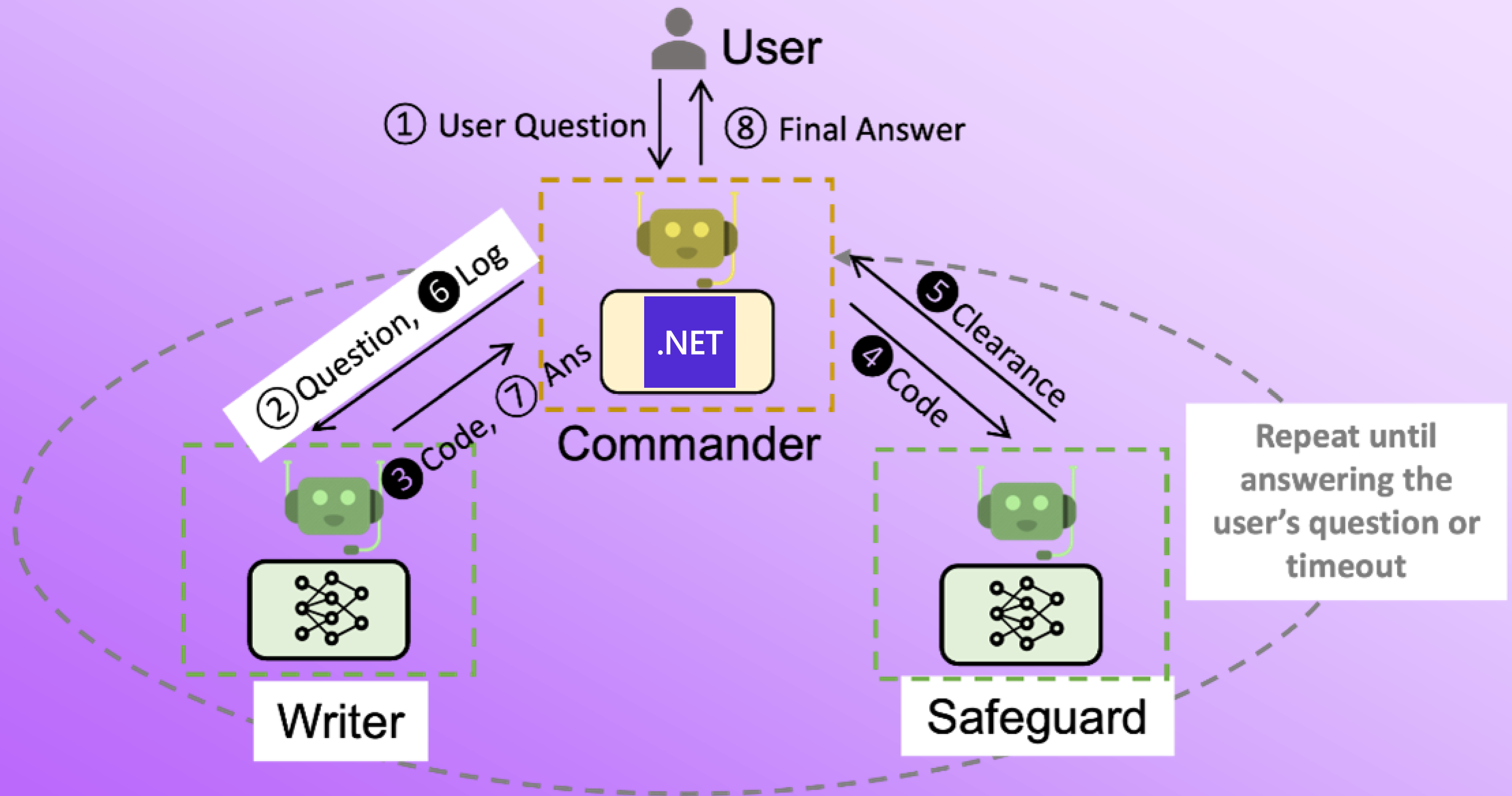
Multi-Agent Conversations



Joint chat

Hierarchical chat

Flexible Conversation Patterns



Some Syntaxes

Running Ollama Model with AutoGen for .NET

Pre-requisite

First, install the AutoGen.Ollama package using the following command:

```
dotnet add package AutoGen.Ollama
```

Step 2: Add using statement

```
using AutoGen.Core; using AutoGen.Ollama.Extension;
```

Create OllamaAgent: Text Based

```
using var httpClient = new HttpClient()
{
    BaseAddress = new Uri("http://localhost:11434"),
};

var ollamaAgent = new OllamaAgent(
    httpClient: httpClient,
    name: "ollama",
    modelName: "llama3:latest",
    systemMessage: "You are a helpful AI assistant")
    .RegisterMessageConnector()
    .RegisterPrintMessage();

var reply = await ollamaAgent.SendAsync("Can you write a piece of C# code to calculate
100th of fibonacci?");
```

Code and Slides are available on my GitHub

chrnthnkmutt/ **BoatSLM_cs_experiment**

This repository is use to make the experiment of using ollama, along with custom small language model, in .NET framework



Contributor



Issues



Stars



Forks



Martech Providers for Thai Market (V. 20 & Last Update: January 29nd, 2025)

Consolidated by  **CONTENT SHIFU**

Advertising & Promotion

Social & Search Ads



Location Based Ads



Display & Programmatic Ads



Native Content



Data

Analytics



Business Automation



Social / eCommerce Listening



Marketing/Business Dashboard



Cookie & Consent Management



Testing & Heat Mapping



CDP



Social & Relationships

Social Media Management



Customer Support



CRM (for B2C)



CRM (for B2B)



Event/Webinar



Influencer Marketing



REVU A STREAM

MOTIVE INFLUENCE AnyMind

Social Proof



Commerce & Sales

eCommerce & Shopping Cart Platform



Chat Commerce



Order Management



Fulfillment



Payment Gateway



Shipping Aggregator



Content & Experience

Content Management System



WIX SQUA SPACE weebly

Email Marketing



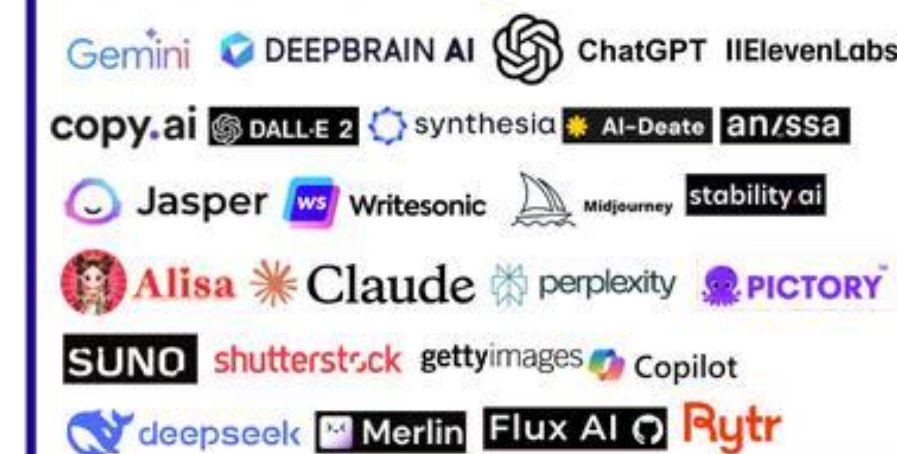
Marketing Automation



Landing Page Instapage



Generative AI



Affiliate Marketing



SMS Marketing



SEO



Interactive Content



vistacreate. VYOND visme

Form & Survey



Lead Generation



Collaboration & Management

Project Management



Chat & Collaboration



Business Operation



Social Proof



Platform



Payment Gateway



Shipping Aggregator



Landing Page



Generative AI



Affiliate Marketing



Form & Survey



Lead Generation



Collaboration & Management

Project Management



Chat & Collaboration

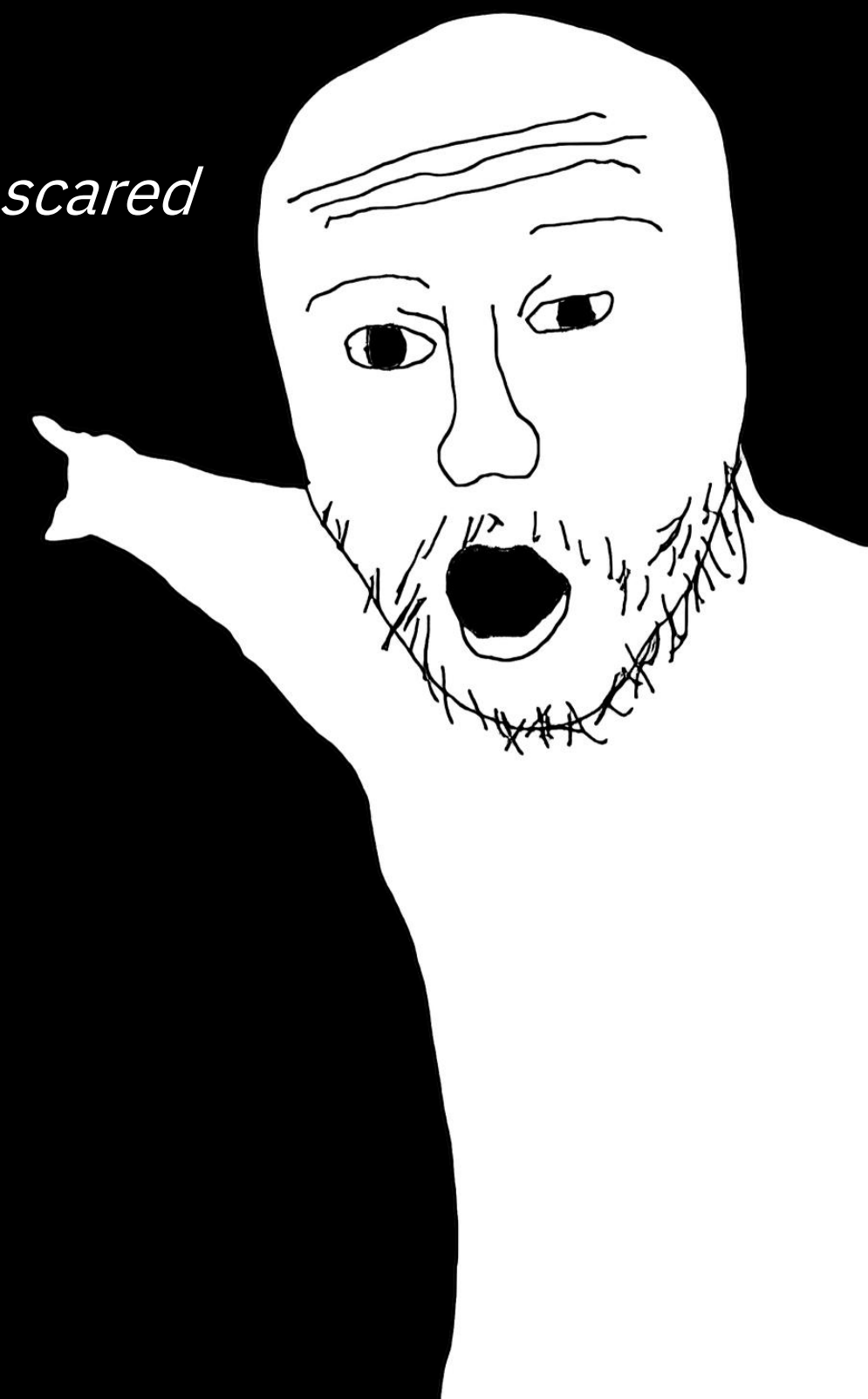


We can see that the world (even
Thailand) is not yet ready for
“Open Source” development

Why people scared from Open-Source

- **Security Concerns:** Some people worry that because the source code is publicly available, it might be easier for malicious actors to find and exploit vulnerabilities.
- **Lack of Support:** Open-source projects often rely on community support rather than dedicated customer service, which can be daunting for those who aren't tech-savvy.
- **Complexity:** Open-source software can sometimes be more complex to install and configure compared to commercial software.
- **Compatibility Issues:** There can be concerns about compatibility with other software or hardware.
- **Perception of Quality:** Some people believe that because open-source software is often free, it might not be as reliable or high-quality as commercial alternatives.

Oh man, I'm scared



**Thank you for watching,
Q & A Section**