

How to use LLMs?

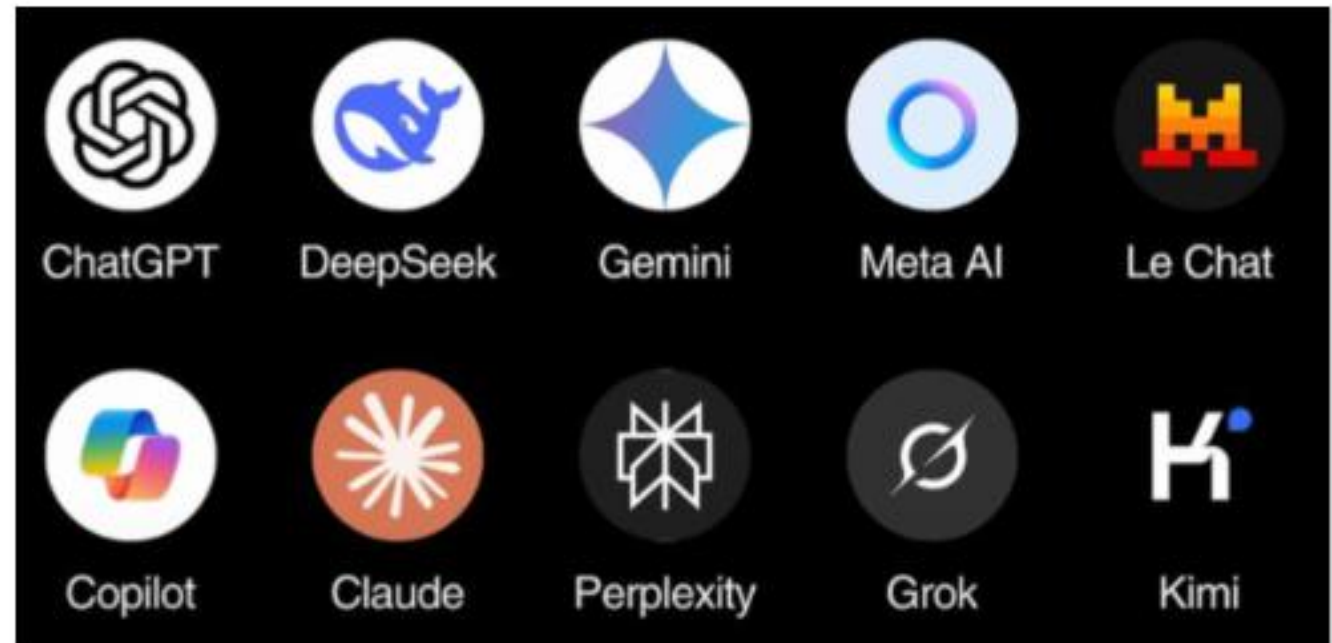
- By Siwen Wang

Preface

- This is a **comprehensive guide covering the usage of Large Language Models** (e.g. ChatGPT, Claude, DeepSeek).
- Original credit goes to Andrej Karpathy (<https://www.youtube.com/watch?v=EWvNQjAaOHw>), I just summarize his words in this slides. Since in the post Tik-tok era, not many people still have the patience to watch a 2-hour long video anymore

List of Common LLMs

- The Founder of all:
 - ChatGPT (OpenAI, U.S.)
- Big tech:
 - Gemini (Google, U.S.)
 - Meta AI (Meta, U.S.)
 - Doubao (ByteDance, China)
- Startups:
 - Claude (Anthropic, U.S.)
 - Grok (xAI, U.S.)
 - DeepSeek (China)
 - Perplexity (U.S.)



Common Things about LLM

- 1. LLMs have it's working memory in the same chat window, it will remember what you said earlier (to an extent) and provide answers based on the context. So **if your topic is continuous, stay in the same chat window**, however, if you change topic, it is recommended to start fresh in a new window.
- 2. Different LLMs have **different model versions**: some are more capable (i.e. smarter), some are faster but not as smart, some offer you additional capabilities such as image understanding etc. You should choose based on your need. If unsure, stay with the default.

LLM as an Assistant

- The most basic usage: type your queries in the chatbox and get an answer immediately
- Best for:
 - 1. Common factual questions
- Not good for:
 - 1. Recent things (it doesn't know)
 - 2. Math and Science questions (it does not do well)



Hi, I'm DeepSeek.

How can I help you today?

How much caffiene in one shot of Americano?



DeepThink (R1)



Search



LLM as STEM Problem Solver

- Some LLMs have the “thinking” capability (ChatGPT o1, DeepSeek R1, Claude 3.7), it can be used to solve more complex questions
- Best for:
 - 1. Math, science questions that require complex reasoning skills
 - 2. Programming problems (debugging)
- Not good for:
 - 1. Recent things (it doesn't know)



Hi, I'm DeepSeek.

How can I help you today?

I have some issue with this code, could you help me to identify the bug?
(paste code below)

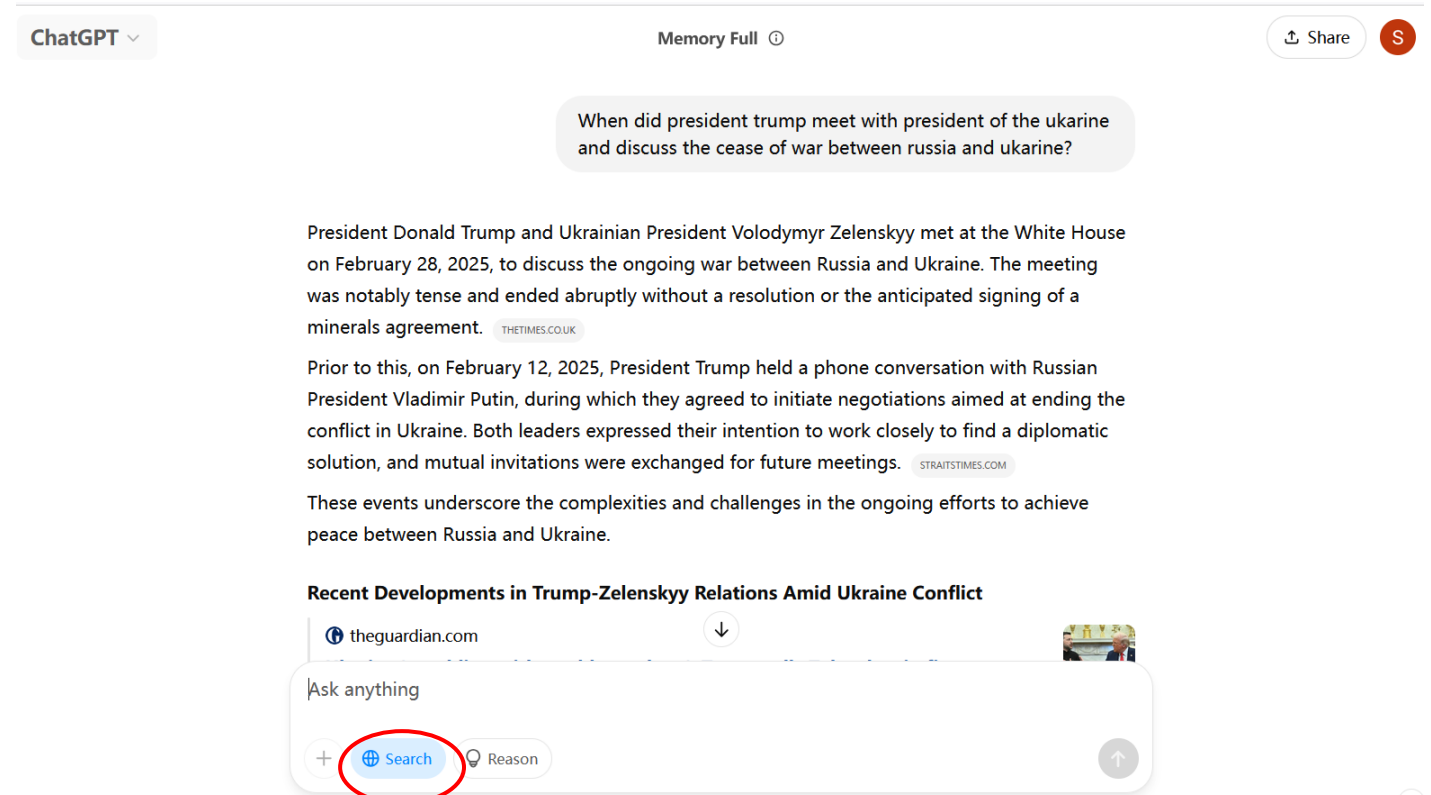
 DeepThink (R1)

 Search



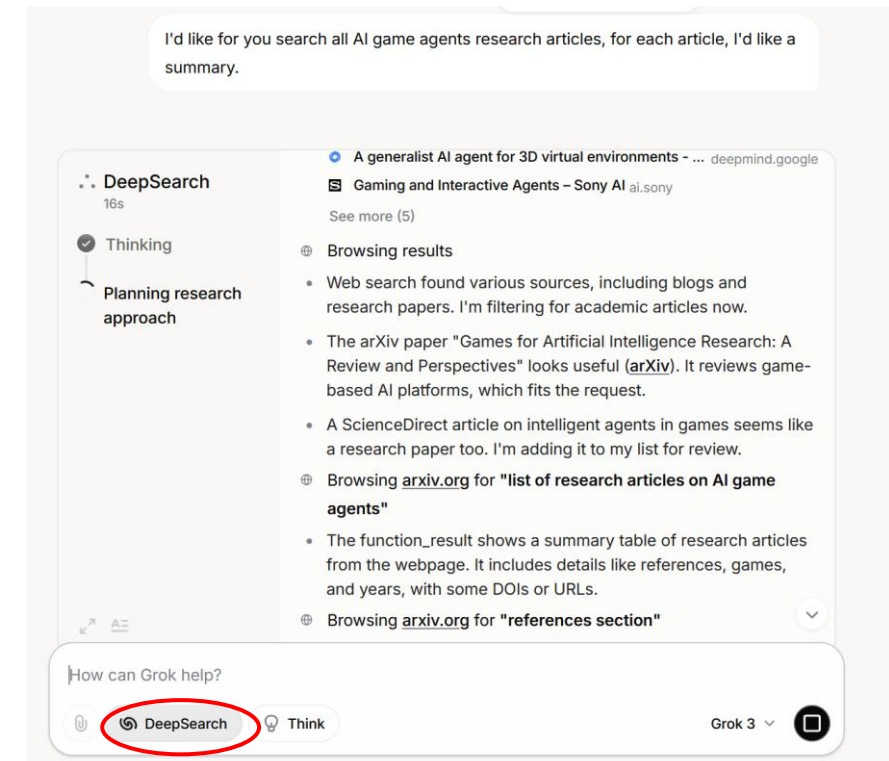
LLM as a search engine

- Some LLMs have tool use such as **searching the internet**. It will browse the pages on its own and give you a **summary** of the content
- Best for:
 - 1. Recently things



LLM as a **research** assistant

- Some LLMs have the feature “deep research”, which it combines **search and thinking**, this can often go quite long (>10 minutes), and it will generate a long report at the end
- Best for:
 - 1. Research questions, very convenient on gathering papers related to your interest



LLM as Reading aid –Part I

- You can upload documents (or just copy paste text) for LLM to read. It can **summarize** and **explain parts** for you.
- Best for:
 - 1. Reading research papers
 - 2. Reading books that are from different fields other than your area of expertise (I recently used ChatGPT to help me read and understand 《庄子》, works great!)

ChatGPT ▾

Memory Full ⓘ



Generative Agents Interactive Sim...
PDF



Could you summarize this paper for me?

Summary of *Generative Agents: Interactive Simulacra of Human Behavior*

This paper introduces **generative agents**, computational software agents that simulate human behavior in interactive environments. These agents exhibit believable daily activities such as waking up, cooking, working, forming relationships, and coordinating group activities.

Key Contributions

1. Generative Agents Architecture

- The system extends a large language model (LLM) with **memory, reflection, and planning** components.
- A **memory stream** records experiences in natural language.
- **Reflection** synthesizes memories into high-level insights.
- **Planning** enables coherent, long-term agent behavior.

2. Implementation in a Sandbox World

Ask anything



Search



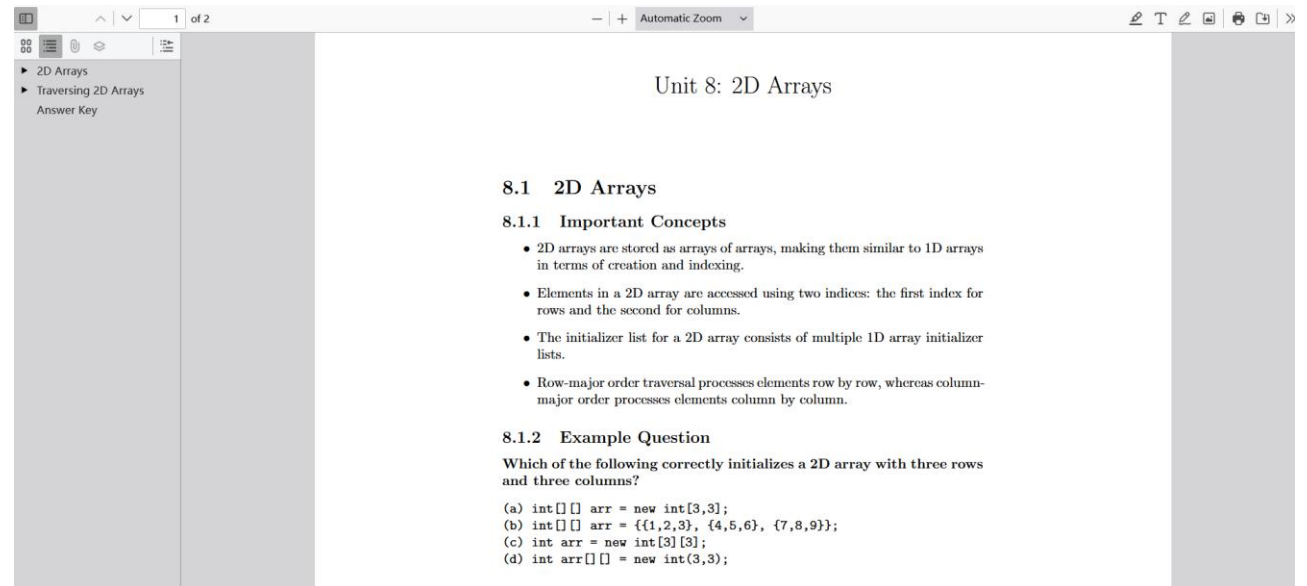
Reason



ChatGPT can make mistakes. Check important info.

Part II- Using the document uploader capability to generate AP review guide

- You can upload the study guide from CollegeBoard chapter by chapter and ask it to generate a summary for you
- However, don't forget to read it to ensure it did not make things up



LLM with Voice Capability

- Some LLMs (e.g. sesame AI) allow you to talk to it using voice. I haven't found them to be particularly helpful for me.
- Best for:
 - 1. Maybe some interesting projects?
 - 2. Maybe you don't want to type out your questions?
 - 3. Conversation partner?

Conversational voice demo

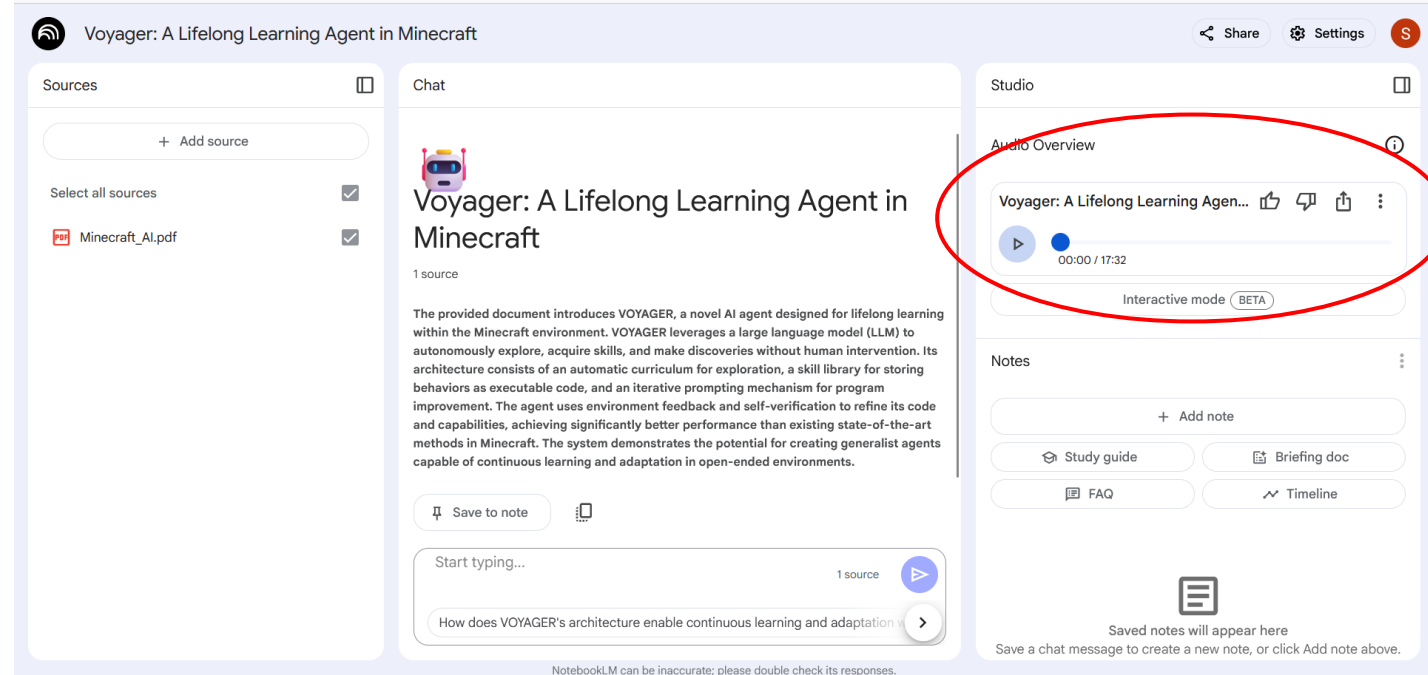


Press to start a conversation

1. Microphone permission is required. 2. Calls are recorded for quality review but not used for ML training and are deleted within 30 days. 3. By using this demo, you are agreeing to our [Terms of Use](#) and [Privacy Policy](#). 4. We recommend using Chrome (Audio quality may be degraded in iOS/ Safari 17.5).

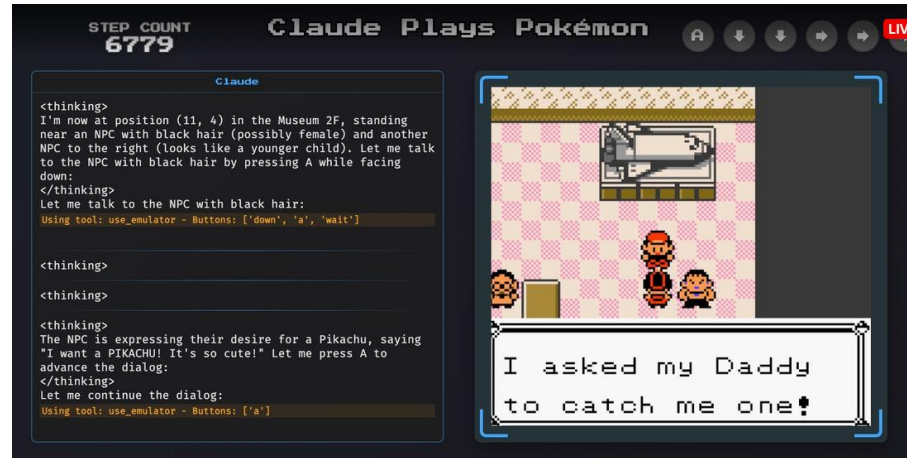
LLM as Podcast

- **Google NotebookLM** can generate a **podcast** based on the document you feed to it. This feature is quite interesting, I tried it with some research paper that's hard to understand, and listening to two people talking about it maybe helped a little bit :p



LLM with Image understanding capability

- Some LLMs allow you to input images and it can understand information in the image
- Some use cases:
 - 1. Game agents
 - 2. OCR (i.e. extract texts from images)



这是一个uno的卡牌，你能告诉我它的颜色和数字吗？

这张 UNO 卡牌的颜色是绿色，数字是1。

UNO 是一种经典的纸牌游戏，由美泰儿公司 (Mattel) 生产。游戏的目标是成为第一个出完手中所有牌的玩家。每张 UNO 牌都有颜色和数字 (或功能)，这张绿色的 1 牌在游戏中属于基本数字牌。玩家在打出这种牌时，需要与弃牌堆顶部牌的颜色或数字匹配才能打出。

LLM as image generator (Text to Image)

- Models like **Stable diffusion**, **ChatGPT DALL.E** can generate images based on text prompt
- Best for:
 - 1. Design work

DALL·E ▾

Please generate a image representing daily life in 2050.



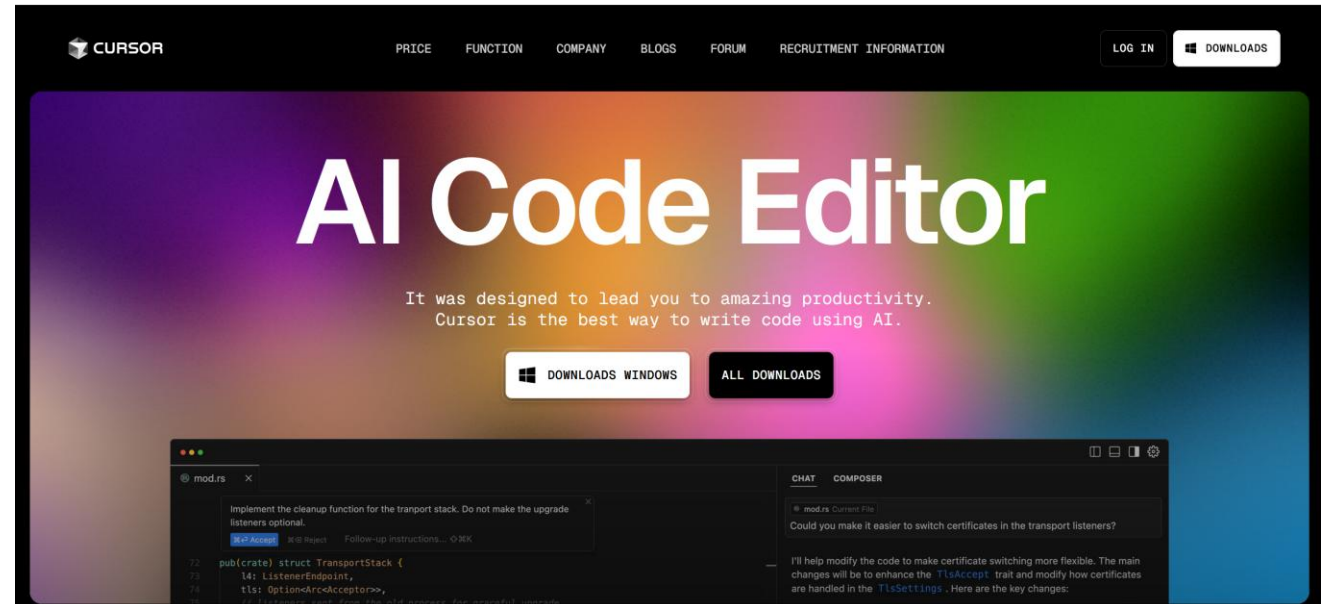
Here are two AI-generated images representing daily life in 2050! Let me know if you'd like any modifications or a different perspective. 🚀

LLM as video generator (Text to Video)

- Models like Google Veo 2

LLM as Coding Assistant (Advanced)

- There are LLM-integrated IDEs such as Cursor that can make coding much easier and more convenient.
- Best for:
 - Coding



Customized LLMs (Advanced use)

- You can write **system prompt** for LLM to change its identify from a helpful assistant to whoever you want it to be, and to answer in a certain manner. **Some researches found that such “cosplay” improves response quality.**
- Ex: In our Pokemon game AI, we told Deepseek that it is the “Best Pokemon Trainer in the World”. We hope it will give better responses.