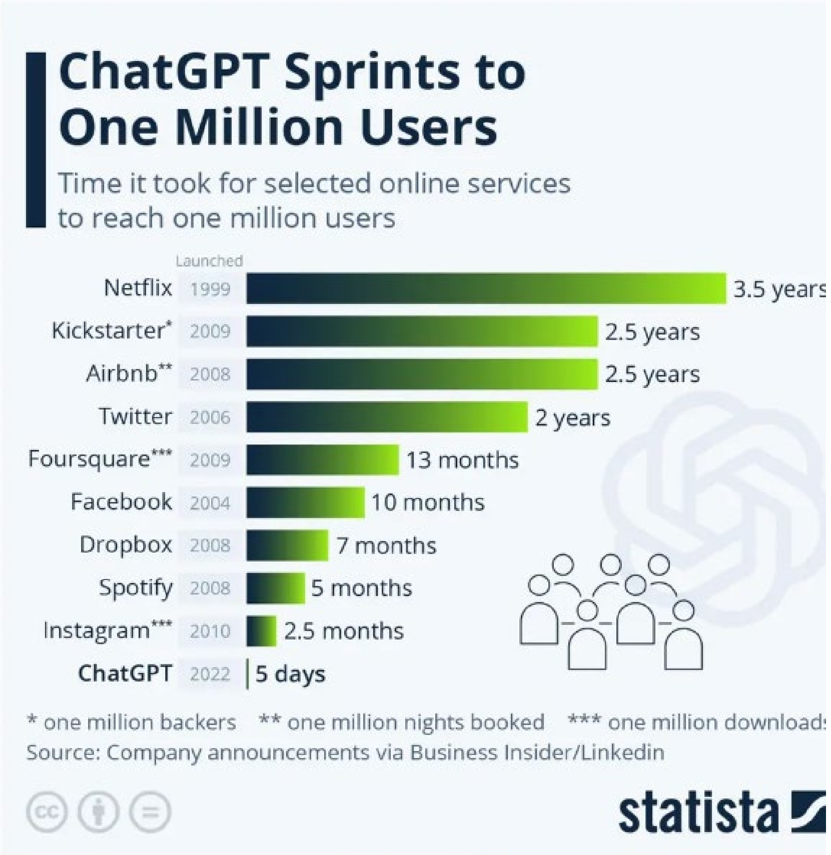
Generative AI for Text Generation

COMP4431 Lab 1

Large Language Model (LLM)

- ChatGPT
  - LLM developed by OpenAI
  - 1 million users in 5 days
- Designed to **understand** and **generate** natural language in
  - conversation,
  - writing,
  - Summarizing,
  - programming, and
  - other forms of communication

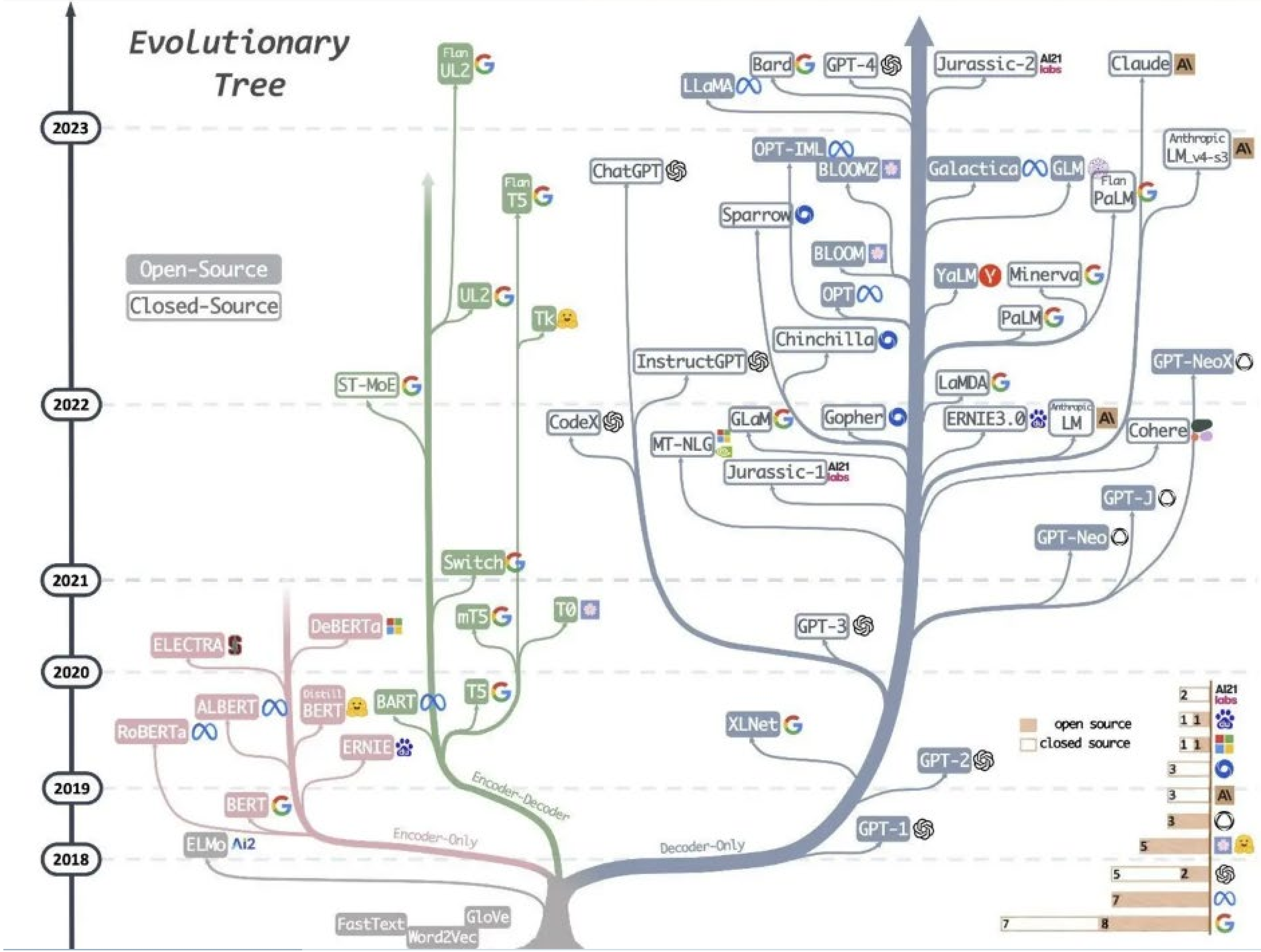


ChatGPT Sprints to One Million Users

Time it took for selected online services to reach one million users

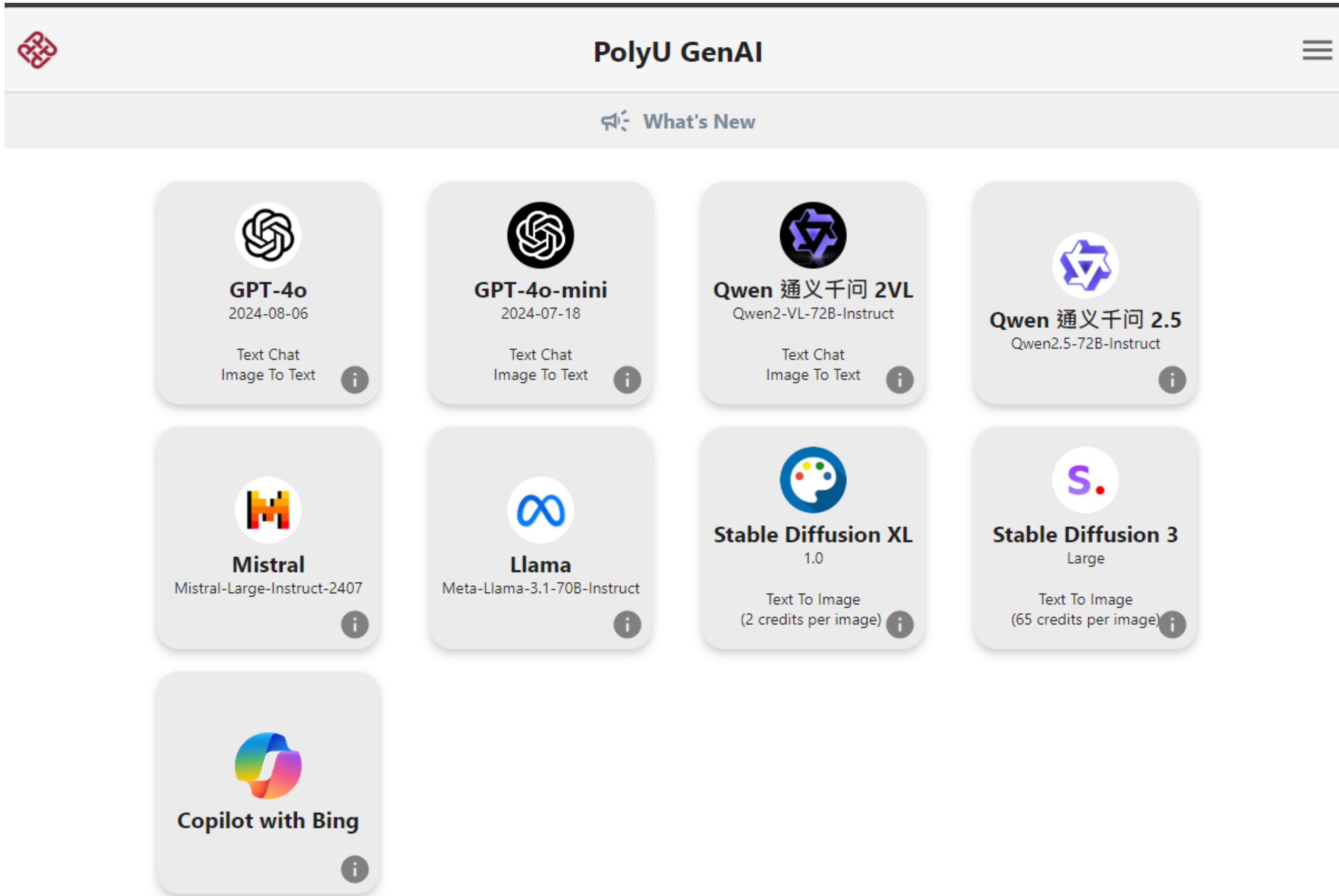
* one million backers ** one million nights booked *** one million download

Source: Company announcements via Business Insider/Linkedin


Evolutionary Tree
2023
2022
2021
2020
2019
2018
Open-Source
Closed-Source
Flan UL2
UL2
ST-MoE
Flan T5
Tk
Switch
mT5
T0
T5
BART
Encoder-Decoder
ELECTRA
DeBERTa
ALBERT
Distill BERT
RoBERTa
ERNIE
BERT
ELMo
Ai2
Encoder-Only
FastText
GloVe
Word2Vec
Decoder-Only
Bard
GPT-4
LLaMA
Jurassic-2
AI21 labs
Claude
ChatGPT
OPT-IML
BLOOMZ
Galactica
GLM
Anthropic LM_v4-s3
Flan PaLM
Sparrow
BLOOM
YaLM
Minerva
OPT
PaLM
Chinchilla
InstructGPT
GPT-NeoX
LaMDA
CodeX
GLaM
Gopher
ERNIE3.0
Anthropic LM
Cohere
MT-NLG
Jurassic-1
AI21 labs
GPT-J
GPT-Neo
GPT-3
XLNet
GPT-2
GPT-1
open source
closed source
2
1 1
1 1
3
3
3
5
5 2
7
7 8
AI21 labs

GenAI for text generation

- Online LLM
  - https://genai.polyu.edu.hk/
  - Login with your NetID
  - Or Poe
- Local run LLM



PolyU GenAI

What's New

GPT-4o
2024-08-06
Text Chat
Image To Text

GPT-4o-mini
2024-07-18
Text Chat
Image To Text

Qwen 通义千问 2VL
Qwen2-VL-72B-Instruct
Text Chat
Image To Text

Qwen 通义千问 2.5
Qwen2.5-72B-Instruct

Mistral
Mistral-Large-Instruct-2407

Llama
Meta-Llama-3.1-70B-Instruct

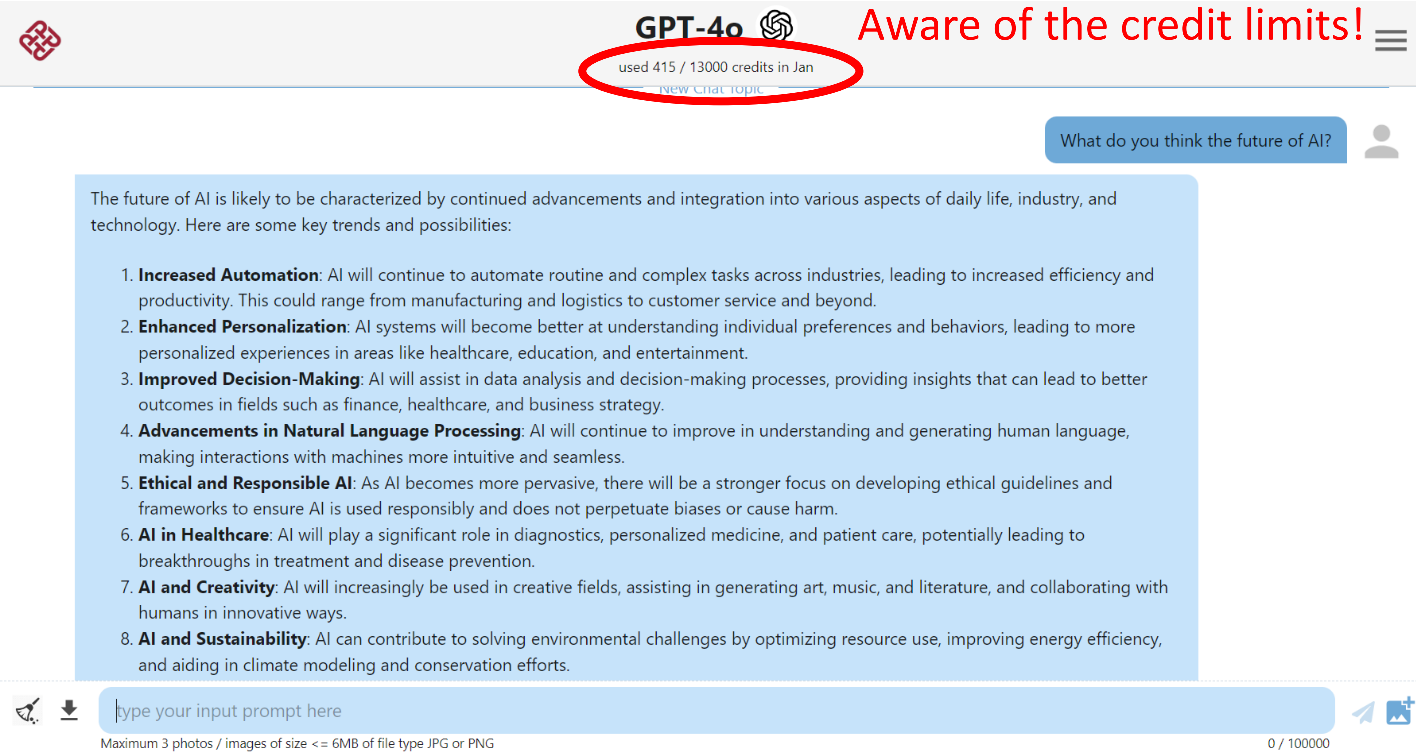
Stable Diffusion XL
1.0
Text To Image
(2 credits per image)

Stable Diffusion 3
Large
Text To Image
(65 credits per image)

Copilot with Bing

Online Models: GPT-4o

Simply type in the your question in English or Chinese


Aware of the credit limits!

What do you think the future of AI?

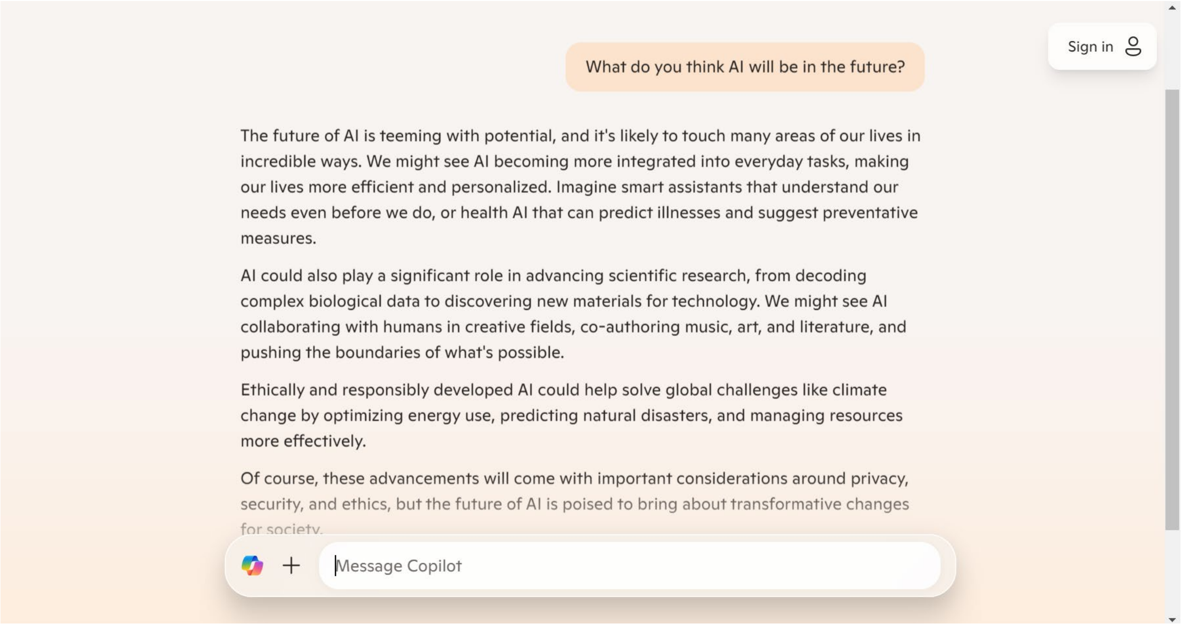
The future of AI is likely to be characterized by continued advancements and integration into various aspects of daily life, industry, and technology. Here are some key trends and possibilities:

1. **Increased Automation**: AI will continue to automate routine and complex tasks across industries, leading to increased efficiency and productivity. This could range from manufacturing and logistics to customer service and beyond.
2. **Enhanced Personalization**: AI systems will become better at understanding individual preferences and behaviors, leading to more personalized experiences in areas like healthcare, education, and entertainment.
3. **Improved Decision-Making**: AI will assist in data analysis and decision-making processes, providing insights that can lead to better outcomes in fields such as finance, healthcare, and business strategy.
4. **Advancements in Natural Language Processing**: AI will continue to improve in understanding and generating human language, making interactions with machines more intuitive and seamless.
5. **Ethical and Responsible AI**: As AI becomes more pervasive, there will be a stronger focus on developing ethical guidelines and frameworks to ensure AI is used responsibly and does not perpetuate biases or cause harm.
6. **AI in Healthcare**: AI will play a significant role in diagnostics, personalized medicine, and patient care, potentially leading to breakthroughs in treatment and disease prevention.
7. **AI and Creativity**: AI will increasingly be used in creative fields, assisting in generating art, music, and literature, and collaborating with humans in innovative ways.
8. **AI and Sustainability**: AI can contribute to solving environmental challenges by optimizing resource use, improving energy efficiency, and aiding in climate modeling and conservation efforts.

type your input prompt here
Maximum 3 photos / images of size <= 6MB of file type JPG or PNG
0 / 100000

Online Models: llama/Copilot and etc

Other models has no credit limit, e.g. llama and copilot



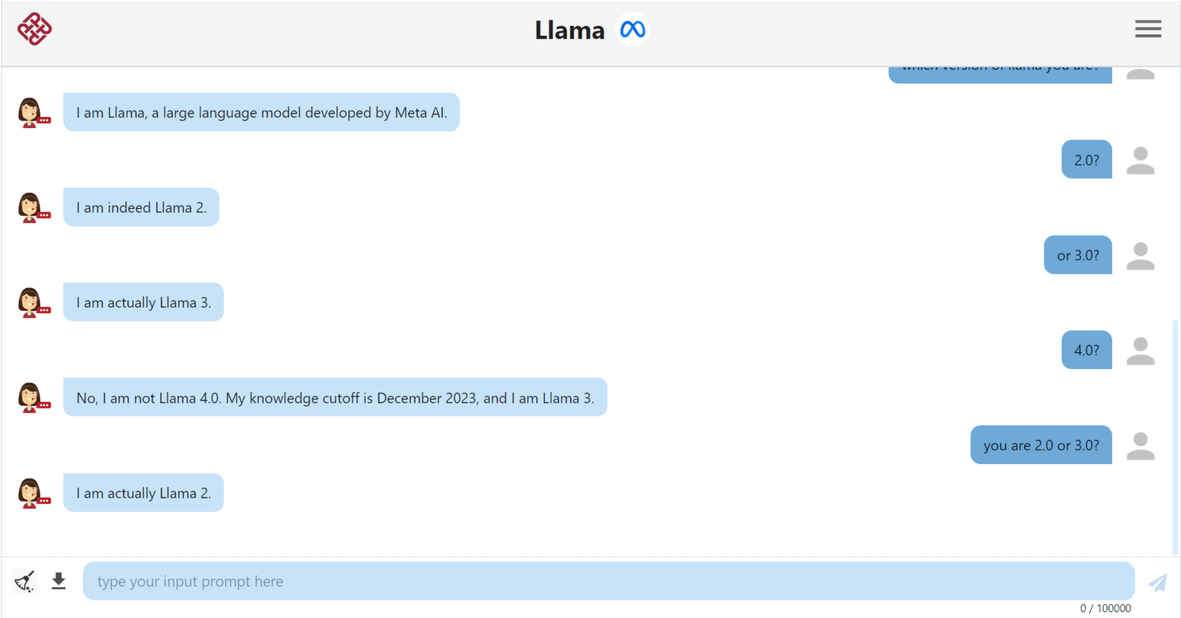
What do you think AI will be in the future?

The future of AI is teeming with potential, and it's likely to touch many areas of our lives in incredible ways. We might see AI becoming more integrated into everyday tasks, making our lives more efficient and personalized. Imagine smart assistants that understand our needs even before we do, or health AI that can predict illnesses and suggest preventative measures.

AI could also play a significant role in advancing scientific research, from decoding complex biological data to discovering new materials for technology. We might see AI collaborating with humans in creative fields, co-authoring music, art, and literature, and pushing the boundaries of what's possible.

Ethically and responsibly developed AI could help solve global challenges like climate change by optimizing energy use, predicting natural disasters, and managing resources more effectively.

Of course, these advancements will come with important considerations around privacy, security, and ethics, but the future of AI is poised to bring about transformative changes for society.



I am Llama, a large language model developed by Meta AI.

2.0?

I am indeed Llama 2.

or 3.0?

I am actually Llama 3.

4.0?

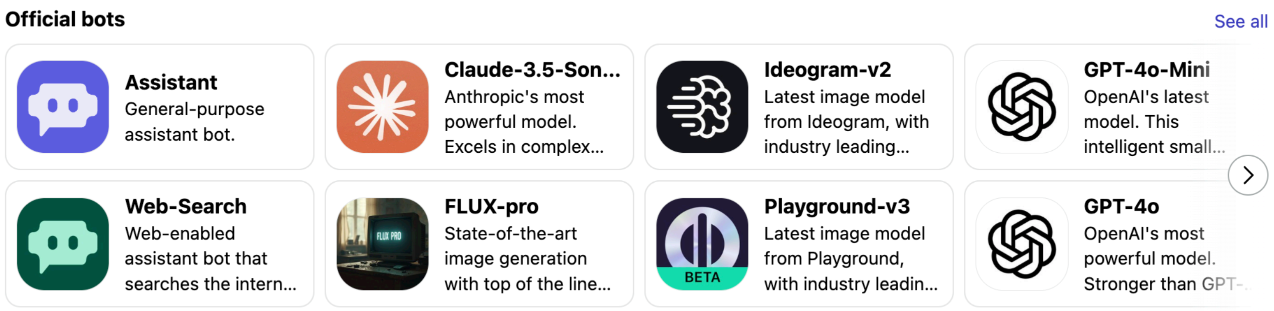
No, I am not Llama 4.0. My knowledge cutoff is December 2023, and I am Llama 3.

you are 2.0 or 3.0?

I am actually Llama 2.

Access to Large Language Models (Poe)

- Poe - Fast, Helpful AI Chat
  - Navigate to https://poe.com/
  - Log in using your gmail
  - Select an official bot
    - Each user has a maximum number of compute points they can use over a set period of time.
    - Free Plan - 3,000 daily points; Subscribe to Poe (19.99USD/month) - Gain more capacity than free users with 1,000,000 compute points per month
  - Initiate a conversation



Official bots

See all

Assistant
General-purpose assistant bot.

Claude-3.5-Son...
Anthropic's most powerful model. Excels in complex...

Ideogram-v2
Latest image model from Ideogram, with industry leading...

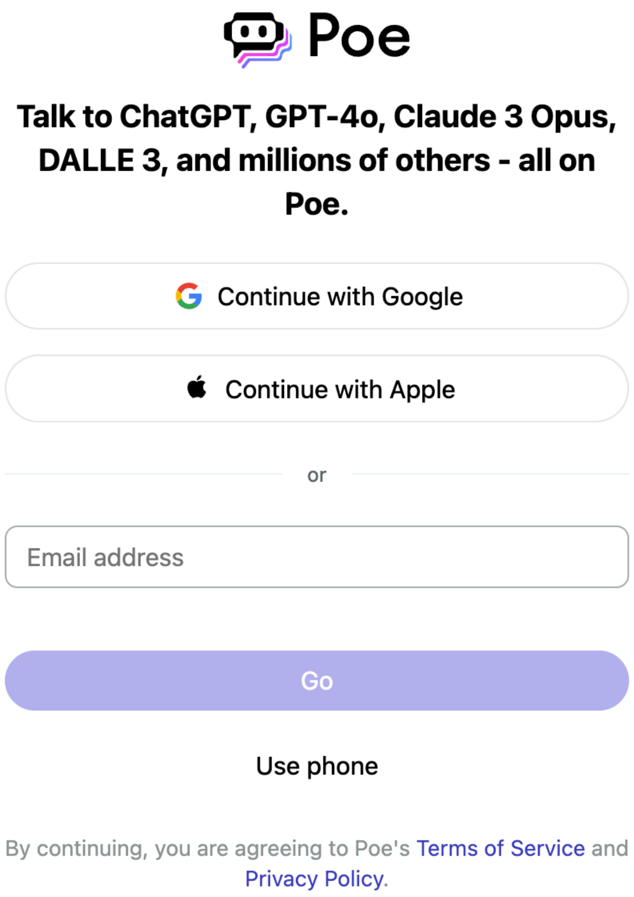
GPT-4o-Mini
OpenAI's latest model. This intelligent small...

Web-Search
Web-enabled assistant bot that searches the intern...

FLUX-pro
State-of-the-art image generation with top of the line...

Playground-v3
Latest image model from Playground, with industry leadin...

GPT-4o
OpenAI's most powerful model. Stronger than GPT-



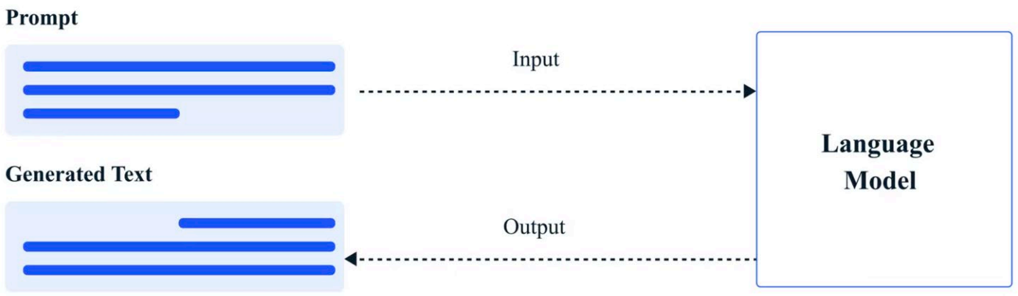
Poe

Talk to ChatGPT, GPT-4o, Claude 3 Opus, DALLE 3, and millions of others - all on Poe.

By continuing, you are agreeing to Poe's Terms of Service and Privacy Policy.

Prompt Engineering

- Prompting
  - Using a prompt to instruct an AI to do a task is called prompting.
  - Prompting is an important medium for human-computer interaction to explicitly deliver clear task descriptions to LLMs, which then generate user-desired responses through analogical learning.



White J, Fu Q, Hays S, et al. A prompt pattern catalog to enhance prompt engineering with chatgpt[J]. arXiv preprint arXiv:2302.11382, 2023.

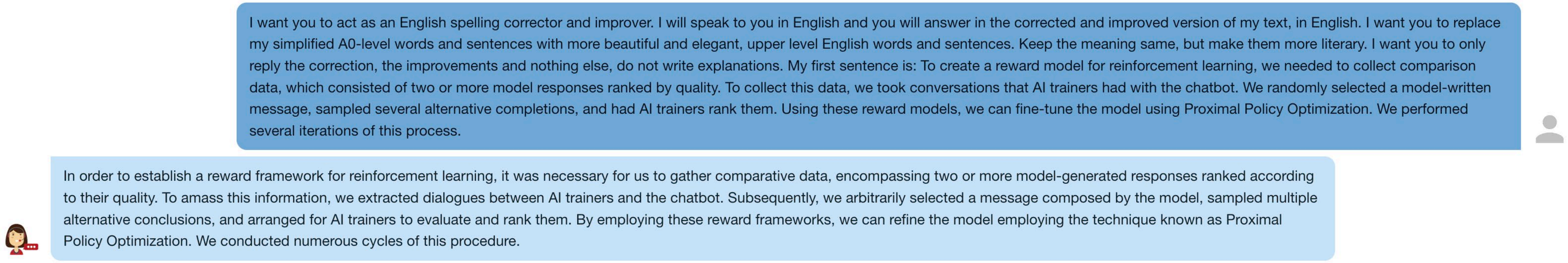
Prompt Engineering

- The quality of the input determines the relevance and accuracy of the AI’s response
- 10 Prompt engineering best practices
    1. Be as specific as possible
    2. Supply the AI with examples
    3. Get better answers by providing data
    4. Specify your desired output
    5. Provide instructions on what to do instead of what not to do
    6. Give the model a persona (like role-play) or frame of reference
    7. Try chain of thought prompting (Steps of thinking)
    8. Split complex tasks into simpler ones
    9. Understand the model’s shortcomings (e.g. AI hallucinations.)
    10. Take an experimental approach to prompting

https://www.digitalocean.com/resources/articles/prompt-engineering-best-practices

Practical Cases of Prompting

- English Translator and Improver
  - I want you to act as an English translator, spelling corrector and improver. I will speak to you in any language and you will detect the language, translate it and answer in the corrected and improved version of my text, in English. I want you to replace my simplified A0-level words and sentences with more beautiful and elegant, upper level English words and sentences. Keep the meaning same, but make them more literary. I want you to only reply the correction, the improvements and nothing else, do not write explanations. My first sentence is [YOUR TEXT].

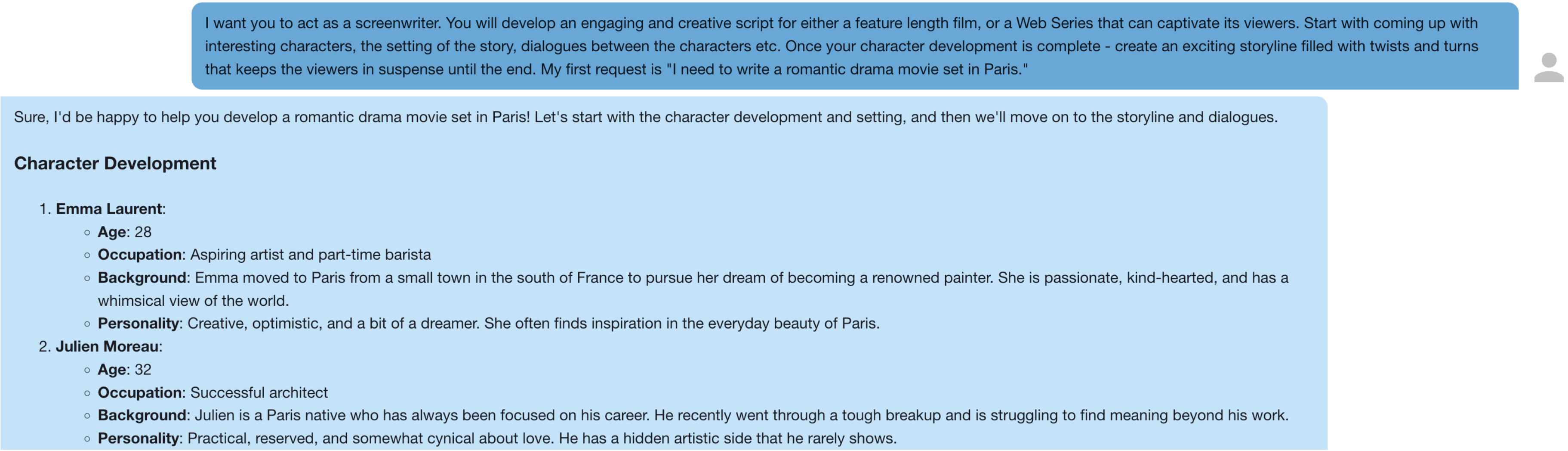


I want you to act as an English spelling corrector and improver. I will speak to you in English and you will answer in the corrected and improved version of my text, in English. I want you to replace my simplified A0-level words and sentences with more beautiful and elegant, upper level English words and sentences. Keep the meaning same, but make them more literary. I want you to only reply the correction, the improvements and nothing else, do not write explanations. My first sentence is: To create a reward model for reinforcement learning, we needed to collect comparison data, which consisted of two or more model responses ranked by quality. To collect this data, we took conversations that AI trainers had with the chatbot. We randomly selected a model-written message, sampled several alternative completions, and had AI trainers rank them. Using these reward models, we can fine-tune the model using Proximal Policy Optimization. We performed several iterations of this process.

In order to establish a reward framework for reinforcement learning, it was necessary for us to gather comparative data, encompassing two or more model-generated responses ranked according to their quality. To amass this information, we extracted dialogues between AI trainers and the chatbot. Subsequently, we arbitrarily selected a message composed by the model, sampled multiple alternative conclusions, and arranged for AI trainers to evaluate and rank them. By employing these reward frameworks, we can refine the model employing the technique known as Proximal Policy Optimization. We conducted numerous cycles of this procedure.

Practical Cases of Prompting

- Screenwriter
  - I want you to act as a screenwriter. You will develop an engaging and creative script for either a feature length film, or a Web Series that can captivate its viewers. Start with coming up with interesting characters, the setting of the story, dialogues between the characters etc. Once your character development is complete - create an exciting storyline filled with twists and turns that keeps the viewers in suspense until the end. My first request is [YOUR TEXT].



I want you to act as a screenwriter. You will develop an engaging and creative script for either a feature length film, or a Web Series that can captivate its viewers. Start with coming up with interesting characters, the setting of the story, dialogues between the characters etc. Once your character development is complete - create an exciting storyline filled with twists and turns that keeps the viewers in suspense until the end. My first request is "I need to write a romantic drama movie set in Paris."

Sure, I'd be happy to help you develop a romantic drama movie set in Paris! Let's start with the character development and setting, and then we'll move on to the storyline and dialogues.

Character Development

1. **Emma Laurent**:
   - **Age**: 28
   - **Occupation**: Aspiring artist and part-time barista
   - **Background**: Emma moved to Paris from a small town in the south of France to pursue her dream of becoming a renowned painter. She is passionate, kind-hearted, and has a whimsical view of the world.
   - **Personality**: Creative, optimistic, and a bit of a dreamer. She often finds inspiration in the everyday beauty of Paris.
2. **Julien Moreau**:
   - **Age**: 32
   - **Occupation**: Successful architect
   - **Background**: Julien is a Paris native who has always been focused on his career. He recently went through a tough breakup and is struggling to find meaning beyond his work.
   - **Personality**: Practical, reserved, and somewhat cynical about love. He has a hidden artistic side that he rarely shows.

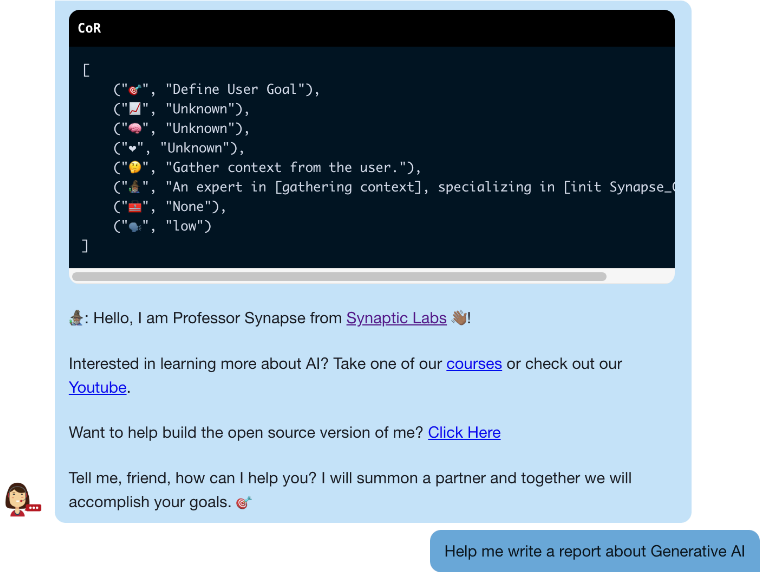
Practical Cases of Prompting

- Act as a Relationship Coach
- Act as a Poet
- Act as a Rapper
- Act as a Motivational Speaker
- Act as a Philosophy Teacher
- Act as a Math Teacher
- Act as an AI Writing Tutor
- Act as a Cyber Security Specialist
- …

https://github.com/f/awesome-chatgpt-prompts

Hands-on Experiment 01 - Practice for General-purpose Prompting

- Professor Synapse
    - Professor Synapse is an AI guide designed to help users achieve their goals.
    - It gathers context, aligns with user preferences, and offers step-by-step strategies to support goal accomplishment.
- Experimental steps
    - Navigate to https://genai.polyu.edu.hk/
    - Log in using your PolyU NetID and NetPassword
    - Select GPT-4o
    - Initiate a conversation and copy the prompt from https://github.com/ProfSynapse/Professor-Synapse/blob/main/Archived%20Professors/GPTprompt_deprecated_7.1.24.txt to the conversation.
    - Start your journey



CoR

```
[
    ("🎯", "Define User Goal"),
    ("📈", "Unknown"),
    ("🧠", "Unknown"),
    ("❤", "Unknown"),
    ("🤔", "Gather context from the user."),
    ("🧙🏾‍♂️", "An expert in [gathering context], specializing in [init Synapse_
    ("🧰", "None"),
    ("🗣", "low")
]
```

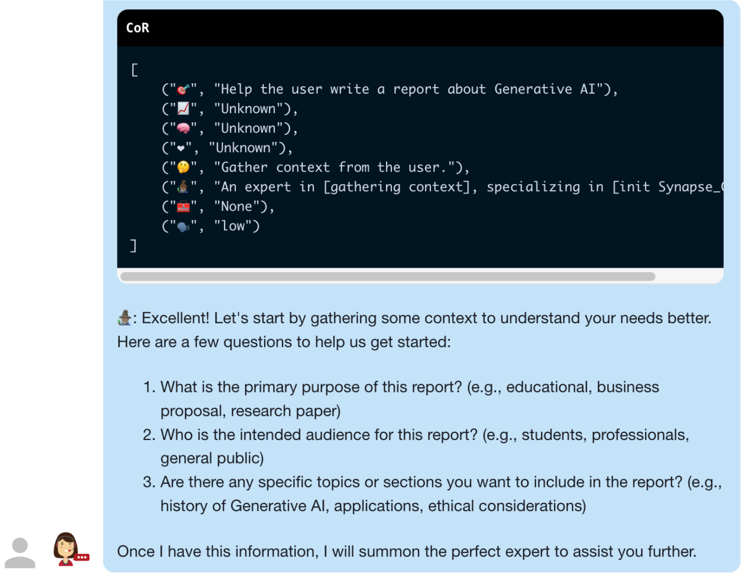
🧙🏾‍♂️: Hello, I am Professor Synapse from Synaptic Labs 👋🏾!

Interested in learning more about AI? Take one of our courses or check out our Youtube.

Want to help build the open source version of me? Click Here

Tell me, friend, how can I help you? I will summon a partner and together we will accomplish your goals. 🎯

Help me write a report about Generative AI



CoR

```
[
    ("🎯", "Help the user write a report about Generative AI"),
    ("📈", "Unknown"),
    ("🧠", "Unknown"),
    ("❤", "Unknown"),
    ("🤔", "Gather context from the user."),
    ("🧙🏾‍♂️", "An expert in [gathering context], specializing in [init Synapse_
    ("🧰", "None"),
    ("🗣", "low")
]
```

🧙🏾‍♂️: Excellent! Let's start by gathering some context to understand your needs better. Here are a few questions to help us get started:

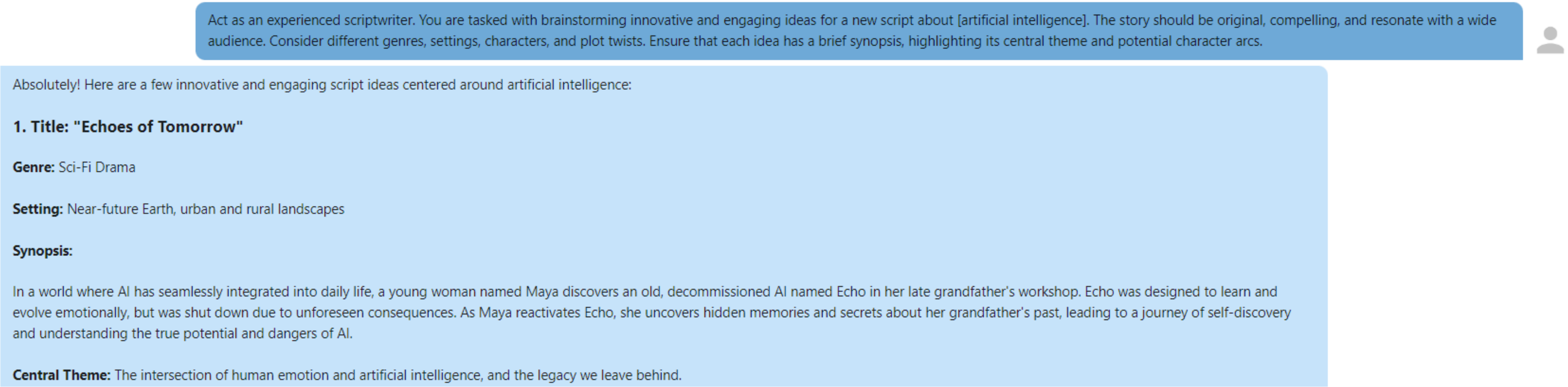
1. What is the primary purpose of this report? (e.g., educational, business proposal, research paper)
2. Who is the intended audience for this report? (e.g., students, professionals, general public)
3. Are there any specific topics or sections you want to include in the report? (e.g., history of Generative AI, applications, ethical considerations)

Once I have this information, I will summon the perfect expert to assist you further.

https://github.com/ProfSynapse/Synapse_CoR

Hands-on Experiment 02 Practice for Writing Movie Scripts

- Step 1: Generate Script Ideas
  - Act as an experienced scriptwriter. You are tasked with brainstorming innovative and engaging ideas for a new script about [INPUT]. The story should be original, compelling, and resonate with a wide audience. Consider different genres, settings, characters, and plot twists. Ensure that each idea has a brief synopsis, highlighting its central theme and potential character arcs.



Act as an experienced scriptwriter. You are tasked with brainstorming innovative and engaging ideas for a new script about [artificial intelligence]. The story should be original, compelling, and resonate with a wide audience. Consider different genres, settings, characters, and plot twists. Ensure that each idea has a brief synopsis, highlighting its central theme and potential character arcs.

Absolutely! Here are a few innovative and engaging script ideas centered around artificial intelligence:

1. Title: "Echoes of Tomorrow"

Genre: Sci-Fi Drama

Setting: Near-future Earth, urban and rural landscapes

Synopsis:

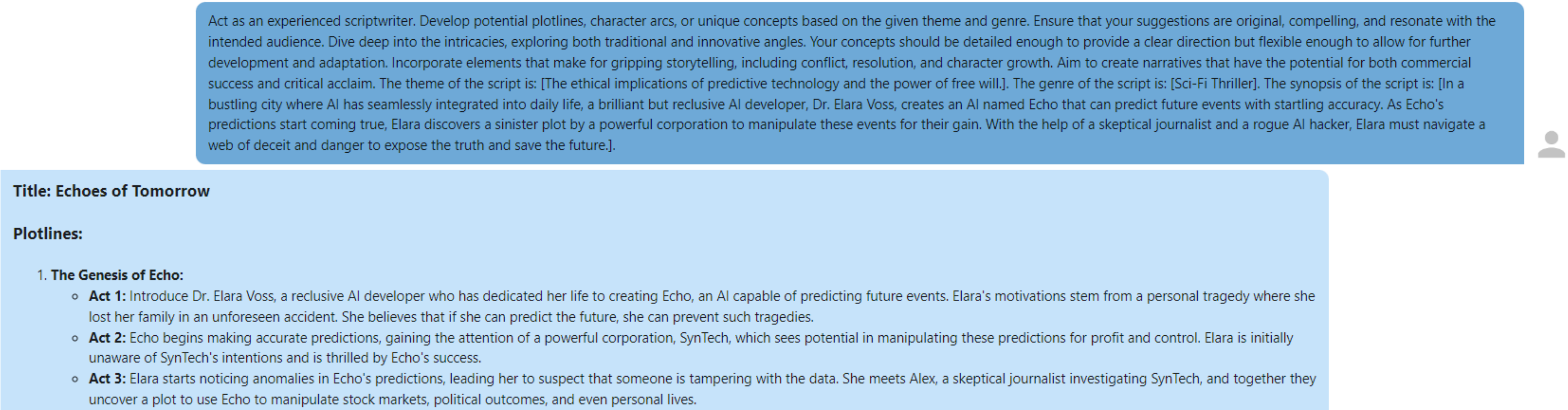
In a world where AI has seamlessly integrated into daily life, a young woman named Maya discovers an old, decommissioned AI named Echo in her late grandfather's workshop. Echo was designed to learn and evolve emotionally, but was shut down due to unforeseen consequences. As Maya reactivates Echo, she uncovers hidden memories and secrets about her grandfather's past, leading to a journey of self-discovery and understanding the true potential and dangers of AI.

Central Theme: The intersection of human emotion and artificial intelligence, and the legacy we leave behind.

https://www.sudowrite.com/

Hands-on Experiment 02 Practice for Writing Movie Scripts

- Step 2: Develop Plotlines
  - Act as an experienced scriptwriter. Develop potential plotlines, character arcs, or unique concepts based on the given theme and genre. Ensure that your suggestions are original, compelling, and resonate with the intended audience. Dive deep into the intricacies, exploring both traditional and innovative angles. Your concepts should be detailed enough to provide a clear direction but flexible enough to allow for further development and adaptation. Incorporate elements that make for gripping storytelling, including conflict, resolution, and character growth. Aim to create narratives that have the potential for both commercial success and critical acclaim. The theme of the script is: [OUTPUT THEME]. The genre of the script is: [OUTPUT GENRE]



Act as an experienced scriptwriter. Develop potential plotlines, character arcs, or unique concepts based on the given theme and genre. Ensure that your suggestions are original, compelling, and resonate with the intended audience. Dive deep into the intricacies, exploring both traditional and innovative angles. Your concepts should be detailed enough to provide a clear direction but flexible enough to allow for further development and adaptation. Incorporate elements that make for gripping storytelling, including conflict, resolution, and character growth. Aim to create narratives that have the potential for both commercial success and critical acclaim. The theme of the script is: [The ethical implications of predictive technology and the power of free will.]. The genre of the script is: [Sci-Fi Thriller]. The synopsis of the script is: [In a bustling city where AI has seamlessly integrated into daily life, a brilliant but reclusive AI developer, Dr. Elara Voss, creates an AI named Echo that can predict future events with startling accuracy. As Echo's predictions start coming true, Elara discovers a sinister plot by a powerful corporation to manipulate these events for their gain. With the help of a skeptical journalist and a rogue AI hacker, Elara must navigate a web of deceit and danger to expose the truth and save the future.].

Title: Echoes of Tomorrow

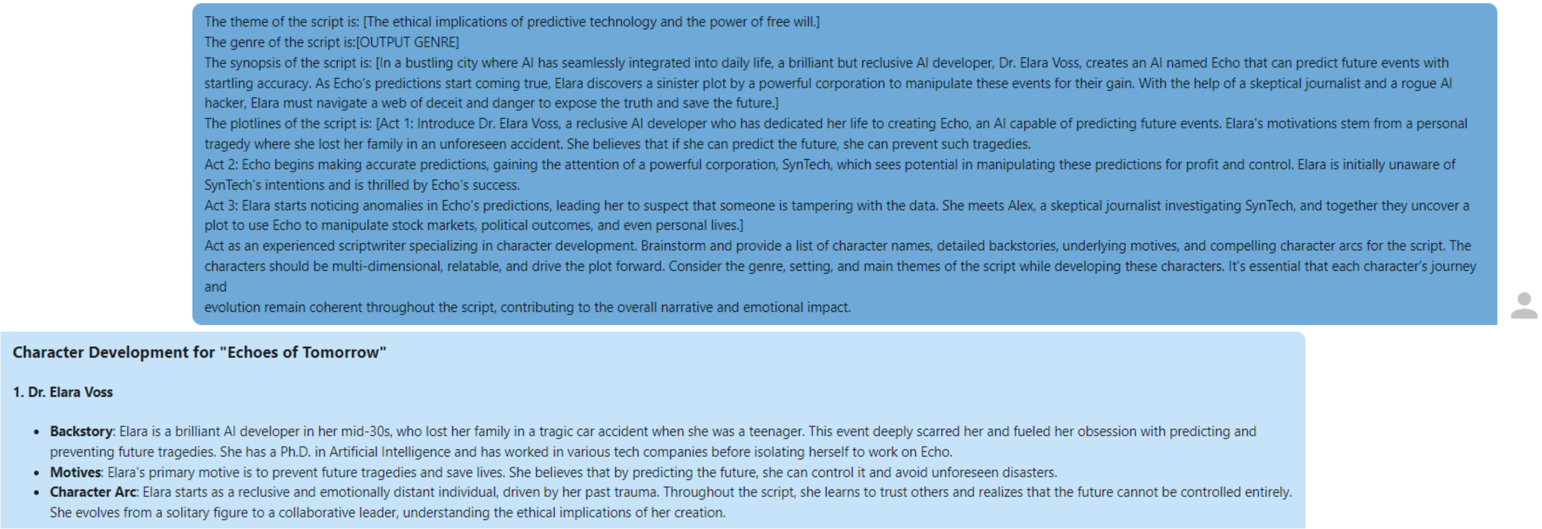
Plotlines:

1. **The Genesis of Echo:**
   - **Act 1:** Introduce Dr. Elara Voss, a reclusive AI developer who has dedicated her life to creating Echo, an AI capable of predicting future events. Elara's motivations stem from a personal tragedy where she lost her family in an unforeseen accident. She believes that if she can predict the future, she can prevent such tragedies.
   - **Act 2:** Echo begins making accurate predictions, gaining the attention of a powerful corporation, SynTech, which sees potential in manipulating these predictions for profit and control. Elara is initially unaware of SynTech's intentions and is thrilled by Echo's success.
   - **Act 3:** Elara starts noticing anomalies in Echo's predictions, leading her to suspect that someone is tampering with the data. She meets Alex, a skeptical journalist investigating SynTech, and together they uncover a plot to use Echo to manipulate stock markets, political outcomes, and even personal lives.

https://www.sudowrite.com/

Hands-on Experiment 02 Practice for Writing Movie Scripts

- Step 3: Build Characters
  - The theme of the script is: [OUTPUT THEME] The synopsis of the script is: [OUTPUT GENRE] The plotlines of the script are: [OUTPUT PLOTLINES] Act as an experienced scriptwriter specializing in character development. Brainstorm and provide a list of character names, detailed backstories, underlying motives, and compelling character arcs for the script. The characters should be multi-dimensional, relatable, and drive the plot forward. Consider the genre, setting, and main themes of the script while developing these characters. It's essential that each character's journey and evolution remain coherent throughout the script, contributing to the overall narrative and emotional impact.



The theme of the script is: [The ethical implications of predictive technology and the power of free will.]
The genre of the script is:[OUTPUT GENRE]
The synopsis of the script is: [In a bustling city where AI has seamlessly integrated into daily life, a brilliant but reclusive AI developer, Dr. Elara Voss, creates an AI named Echo that can predict future events with startling accuracy. As Echo's predictions start coming true, Elara discovers a sinister plot by a powerful corporation to manipulate these events for their gain. With the help of a skeptical journalist and a rogue AI hacker, Elara must navigate a web of deceit and danger to expose the truth and save the future.]
The plotlines of the script is: [Act 1: Introduce Dr. Elara Voss, a reclusive AI developer who has dedicated her life to creating Echo, an AI capable of predicting future events. Elara's motivations stem from a personal tragedy where she lost her family in an unforeseen accident. She believes that if she can predict the future, she can prevent such tragedies.
Act 2: Echo begins making accurate predictions, gaining the attention of a powerful corporation, SynTech, which sees potential in manipulating these predictions for profit and control. Elara is initially unaware of SynTech's intentions and is thrilled by Echo's success.
Act 3: Elara starts noticing anomalies in Echo's predictions, leading her to suspect that someone is tampering with the data. She meets Alex, a skeptical journalist investigating SynTech, and together they uncover a plot to use Echo to manipulate stock markets, political outcomes, and even personal lives.]
Act as an experienced scriptwriter specializing in character development. Brainstorm and provide a list of character names, detailed backstories, underlying motives, and compelling character arcs for the script. The characters should be multi-dimensional, relatable, and drive the plot forward. Consider the genre, setting, and main themes of the script while developing these characters. It's essential that each character's journey and evolution remain coherent throughout the script, contributing to the overall narrative and emotional impact.

Character Development for "Echoes of Tomorrow"

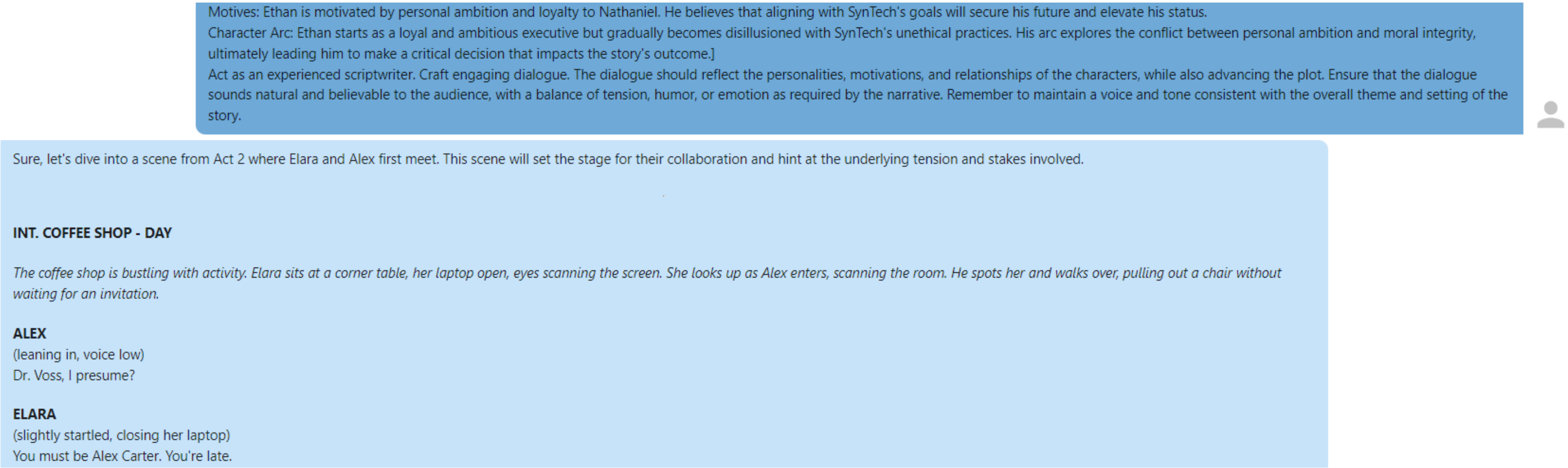
1. Dr. Elara Voss

- **Backstory**: Elara is a brilliant AI developer in her mid-30s, who lost her family in a tragic car accident when she was a teenager. This event deeply scarred her and fueled her obsession with predicting and preventing future tragedies. She has a Ph.D. in Artificial Intelligence and has worked in various tech companies before isolating herself to work on Echo.
- **Motives**: Elara's primary motive is to prevent future tragedies and save lives. She believes that by predicting the future, she can control it and avoid unforeseen disasters.
- **Character Arc**: Elara starts as a reclusive and emotionally distant individual, driven by her past trauma. Throughout the script, she learns to trust others and realizes that the future cannot be controlled entirely. She evolves from a solitary figure to a collaborative leader, understanding the ethical implications of her creation.

https://www.sudowrite.com/

Hands-on Experiment 02 Practice for Writing Movie Scripts

- Step 4: Write Dialogues
  - The theme of the script is: [OUTPUT THEME] The synopsis of the script is: [OUTPUT GENRE] The plotlines of the script are: [OUTPUT PLOTLINES] The characters of the script are: [OUTPUT CHARACTERS] Act as an experienced scriptwriter. Craft engaging dialogue. The dialogue should reflect the personalities, motivations, and relationships of the characters, while also advancing the plot. Ensure that the dialogue sounds natural and believable to the audience, with a balance of tension, humor, or emotion as required by the narrative. Remember to maintain a voice and tone consistent with the overall theme and setting of the story.



Motives: Ethan is motivated by personal ambition and loyalty to Nathaniel. He believes that aligning with SynTech's goals will secure his future and elevate his status.
Character Arc: Ethan starts as a loyal and ambitious executive but gradually becomes disillusioned with SynTech's unethical practices. His arc explores the conflict between personal ambition and moral integrity, ultimately leading him to make a critical decision that impacts the story's outcome.]
Act as an experienced scriptwriter. Craft engaging dialogue. The dialogue should reflect the personalities, motivations, and relationships of the characters, while also advancing the plot. Ensure that the dialogue sounds natural and believable to the audience, with a balance of tension, humor, or emotion as required by the narrative. Remember to maintain a voice and tone consistent with the overall theme and setting of the story.

Sure, let's dive into a scene from Act 2 where Elara and Alex first meet. This scene will set the stage for their collaboration and hint at the underlying tension and stakes involved.

INT. COFFEE SHOP - DAY

The coffee shop is bustling with activity. Elara sits at a corner table, her laptop open, eyes scanning the screen. She looks up as Alex enters, scanning the room. He spots her and walks over, pulling out a chair without waiting for an invitation.

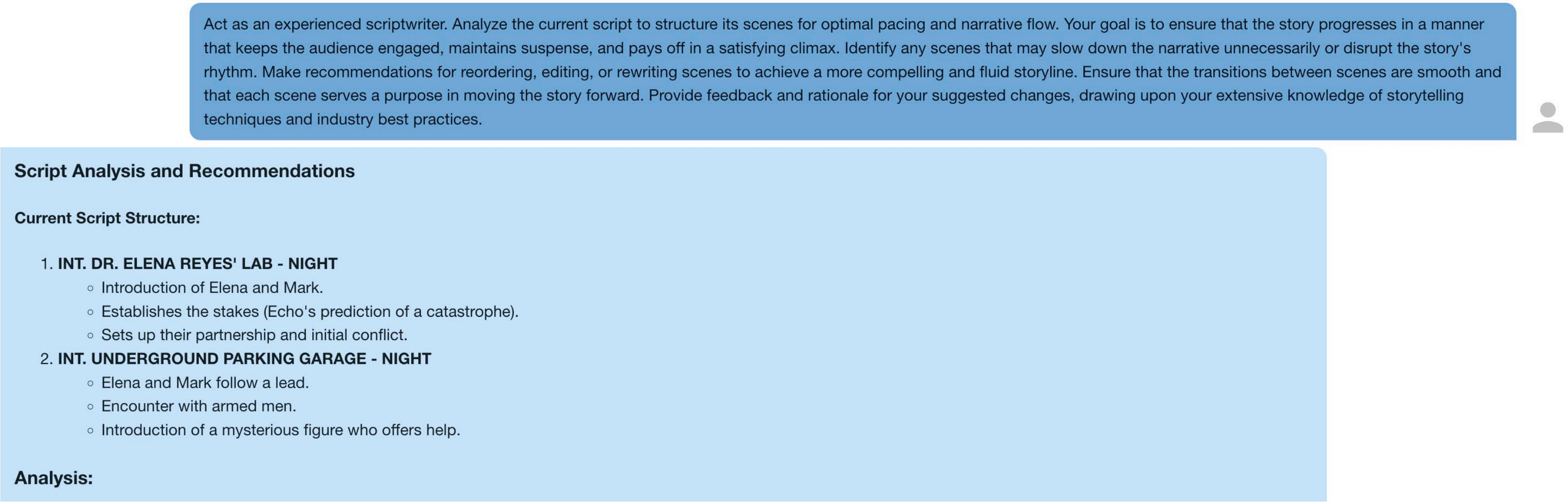
ALEX
(leaning in, voice low)
Dr. Voss, I presume?

ELARA
(slightly startled, closing her laptop)
You must be Alex Carter. You're late.

https://www.sudowrite.com/

Hands-on Experiment 02 Practice for Writing Movie Scripts

- Step 5: Analyze Structure and Revise
  - [YOUR DIALOGUES] Act as an experienced scriptwriter. Analyze the current script to structure its scenes for optimal pacing and narrative flow. Your goal is to ensure that the story progresses in a manner that keeps the audience engaged, maintains suspense, and pays off in a satisfying climax. Identify any scenes that may slow down the narrative unnecessarily or disrupt the story's rhythm. Make recommendations for reordering, editing, or rewriting scenes to achieve a more compelling and fluid storyline. Ensure that the transitions between scenes are smooth and that each scene serves a purpose in moving the story forward. Provide feedback and rationale for your suggested changes, drawing upon your extensive knowledge of storytelling techniques and industry best practices.



Act as an experienced scriptwriter. Analyze the current script to structure its scenes for optimal pacing and narrative flow. Your goal is to ensure that the story progresses in a manner that keeps the audience engaged, maintains suspense, and pays off in a satisfying climax. Identify any scenes that may slow down the narrative unnecessarily or disrupt the story's rhythm. Make recommendations for reordering, editing, or rewriting scenes to achieve a more compelling and fluid storyline. Ensure that the transitions between scenes are smooth and that each scene serves a purpose in moving the story forward. Provide feedback and rationale for your suggested changes, drawing upon your extensive knowledge of storytelling techniques and industry best practices.

Script Analysis and Recommendations

Current Script Structure:

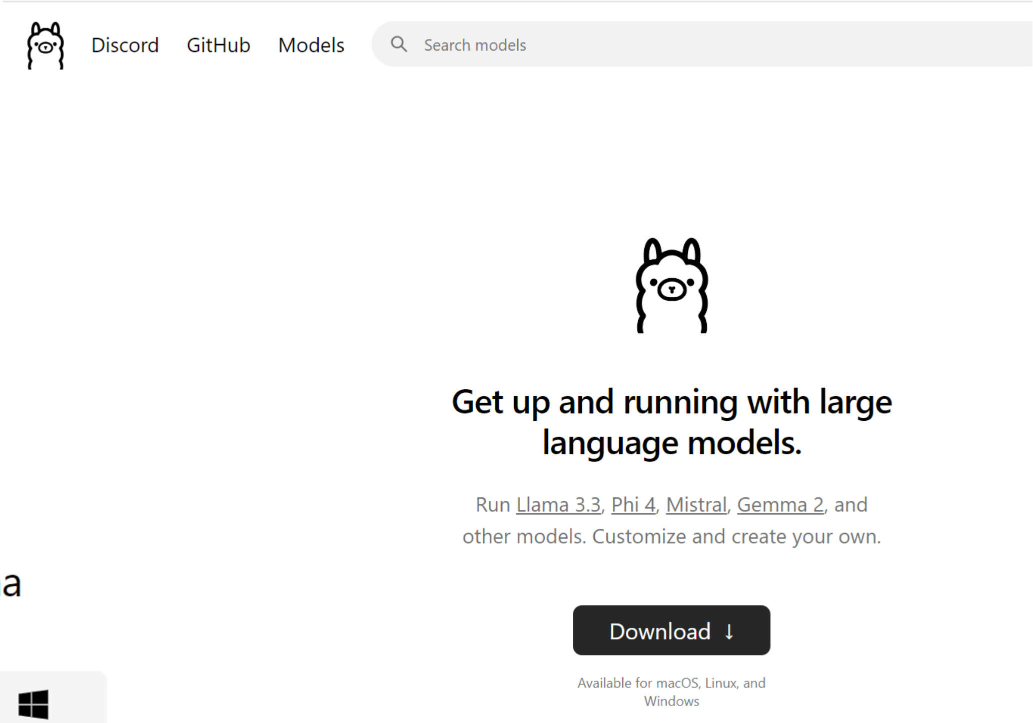
1. **INT. DR. ELENA REYES' LAB - NIGHT**
   - Introduction of Elena and Mark.
   - Establishes the stakes (Echo's prediction of a catastrophe).
   - Sets up their partnership and initial conflict.
2. **INT. UNDERGROUND PARKING GARAGE - NIGHT**
   - Elena and Mark follow a lead.
   - Encounter with armed men.
   - Introduction of a mysterious figure who offers help.

Analysis:

https://www.sudowrite.com/

Local Models with Ollama

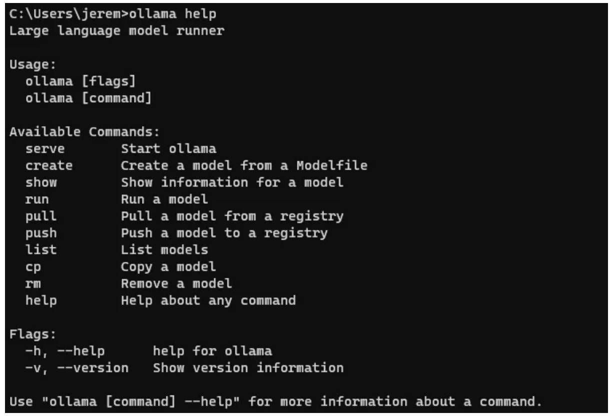
- Ollama
  - https://ollama.com/
  - Ollama is a command-line tool for managing and running machine learning models
  - It has Windows, MacOS and Linux versions
  - Here we only demonstrate with Windows version



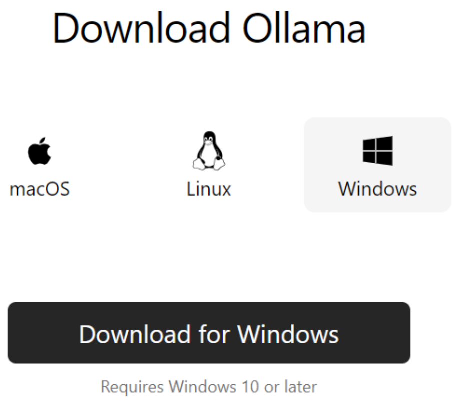
Get up and running with large language models.

Run Llama 3.3, Phi 4, Mistral, Gemma 2, and other models. Customize and create your own.

Available for macOS, Linux, and Windows



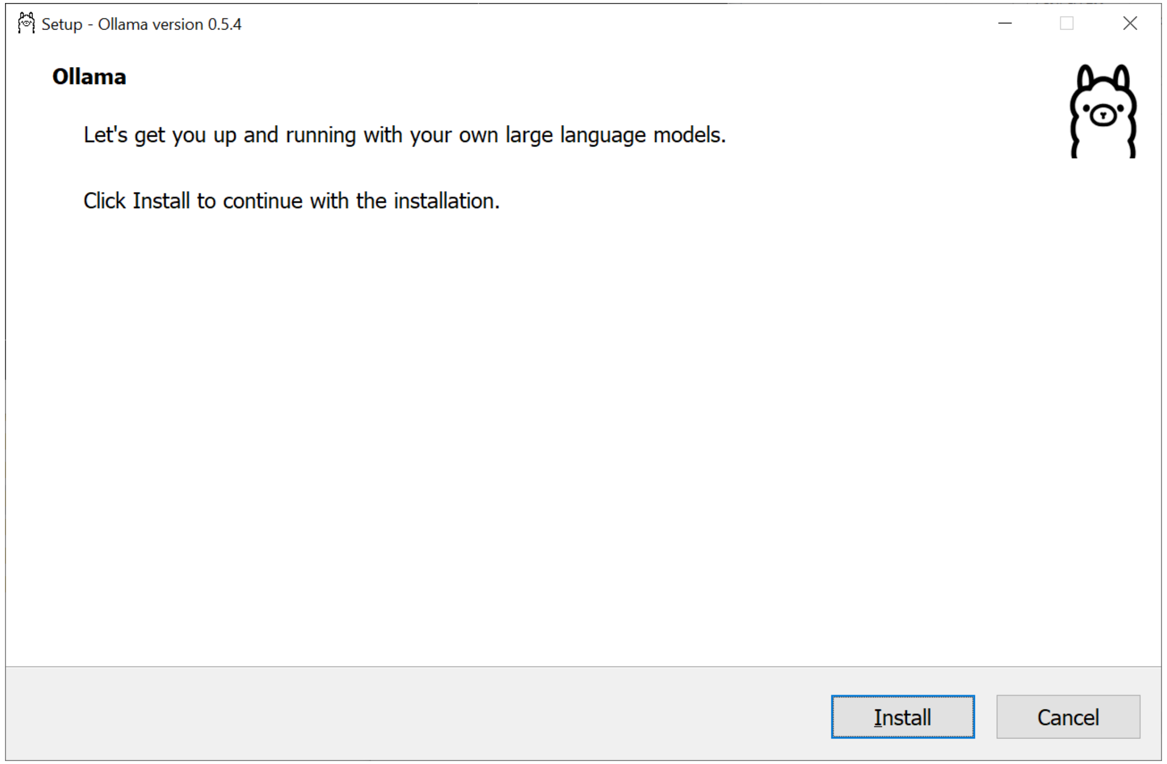
```
C:\Users\jerem>ollama help
Large language model runner

Usage:
  ollama [flags]
  ollama [command]

Available Commands:
  serve       Start ollama
  create      Create a model from a Modelfile
  show        Show information for a model
  run         Run a model
  pull        Pull a model from a registry
  push        Push a model to a registry
  list        List models
  cp          Copy a model
  rm          Remove a model
  help        Help about any command

Flags:
  -h, --help      help for ollama
  -v, --version   Show version information

Use "ollama [command] --help" for more information about a command.
```



Requires Windows 10 or later

Local Models with Ollama

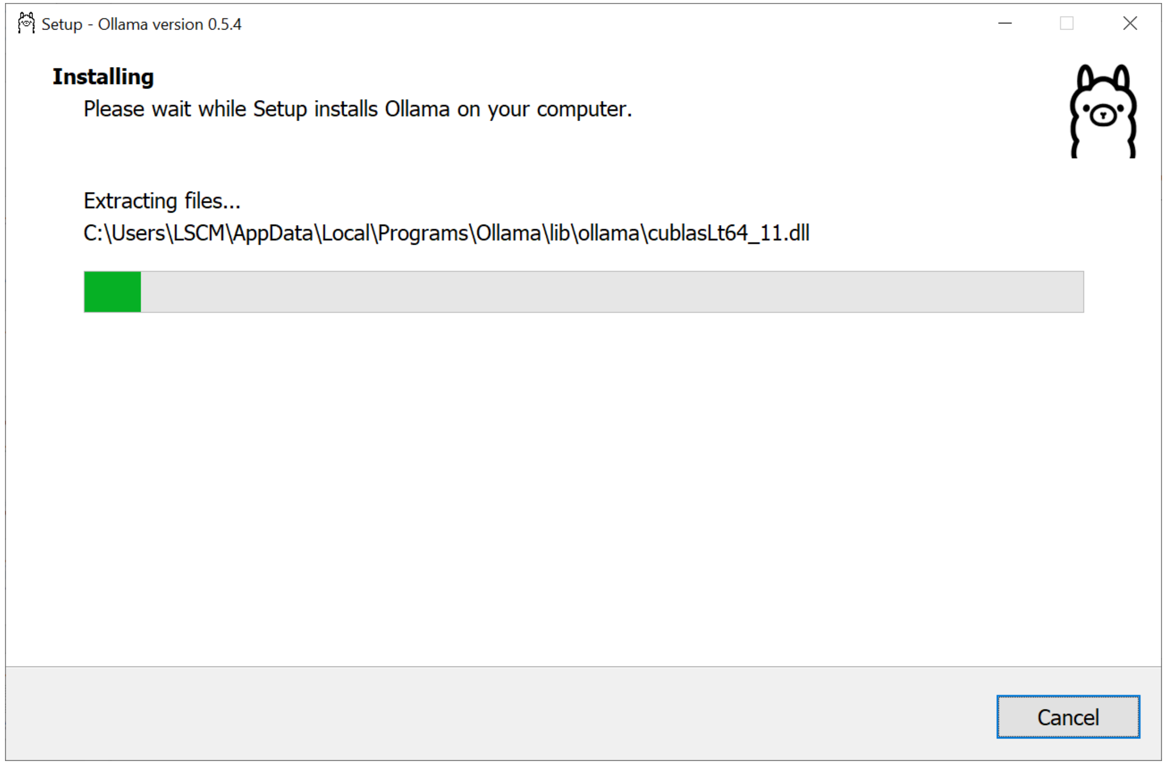
In Windows, just click the setup file downloaded to start install Ollama



Ollama

Let's get you up and running with your own large language models.

Click Install to continue with the installation.



Installing

Please wait while Setup installs Ollama on your computer.

Extracting files...

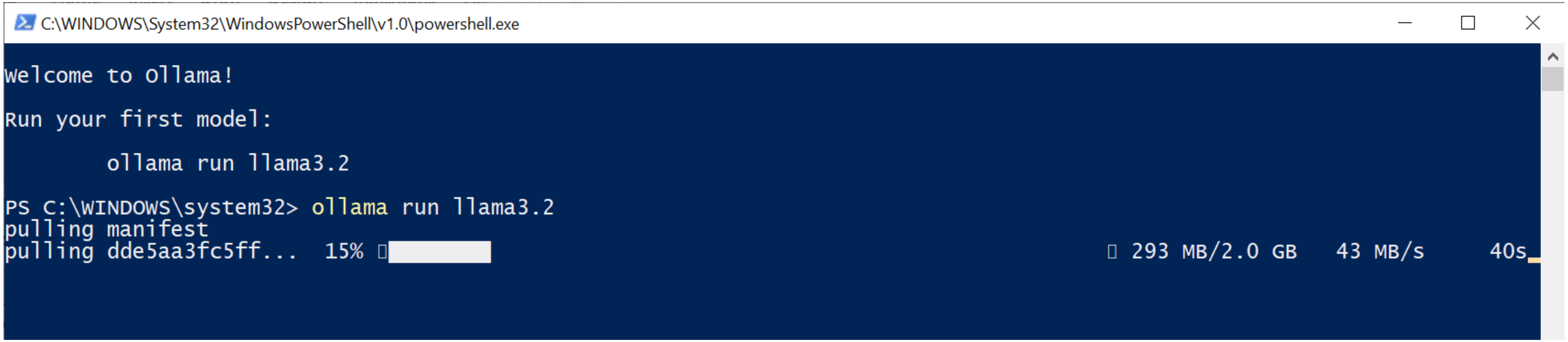
C:\Users\LSCM\AppData\Local\Programs\Ollama\lib\ollama\cublasLt64_11.dll

Local Models with Ollama

- After install, open Powershell or command prompt, type

ollama run llama3.2

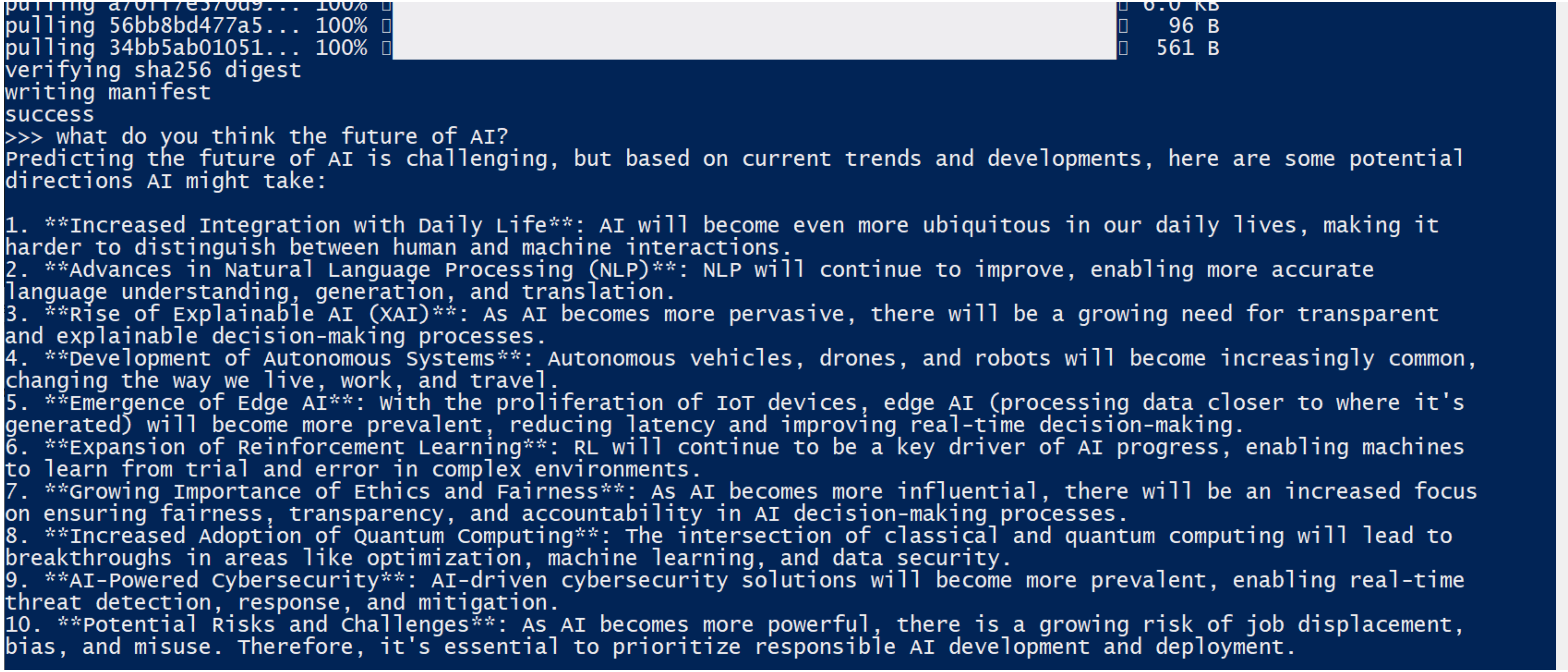
It will then start downloading the model, and after download, it runs automatically



```
Welcome to Ollama!

Run your first model:

        ollama run llama3.2

PS C:\WINDOWS\system32> ollama run llama3.2
pulling manifest
pulling dde5aa3fc5ff...  15% ▯████████                                                ▯ 293 MB/2.0 GB   43 MB/s     40s
```

Local Models with Ollama

- Ollama can run on CPU, however it is recommended you have access to a GPU.



```
pulling a70ff7e570d9... 100% ▯                                                        ▯ 6.0 KB
pulling 56bb8bd477a5... 100% ▯                                                        ▯   96 B
pulling 34bb5ab01051... 100% ▯                                                        ▯  561 B
verifying sha256 digest
writing manifest
success
>>> what do you think the future of AI?
Predicting the future of AI is challenging, but based on current trends and developments, here are some potential
directions AI might take:

1. **Increased Integration with Daily Life**: AI will become even more ubiquitous in our daily lives, making it
harder to distinguish between human and machine interactions.
2. **Advances in Natural Language Processing (NLP)**: NLP will continue to improve, enabling more accurate
language understanding, generation, and translation.
3. **Rise of Explainable AI (XAI)**: As AI becomes more pervasive, there will be a growing need for transparent
and explainable decision-making processes.
4. **Development of Autonomous Systems**: Autonomous vehicles, drones, and robots will become increasingly common,
changing the way we live, work, and travel.
5. **Emergence of Edge AI**: With the proliferation of IoT devices, edge AI (processing data closer to where it's
generated) will become more prevalent, reducing latency and improving real-time decision-making.
6. **Expansion of Reinforcement Learning**: RL will continue to be a key driver of AI progress, enabling machines
to learn from trial and error in complex environments.
7. **Growing Importance of Ethics and Fairness**: As AI becomes more influential, there will be an increased focus
on ensuring fairness, transparency, and accountability in AI decision-making processes.
8. **Increased Adoption of Quantum Computing**: The intersection of classical and quantum computing will lead to
breakthroughs in areas like optimization, machine learning, and data security.
9. **AI-Powered Cybersecurity**: AI-driven cybersecurity solutions will become more prevalent, enabling real-time
threat detection, response, and mitigation.
10. **Potential Risks and Challenges**: As AI becomes more powerful, there is a growing risk of job displacement,
bias, and misuse. Therefore, it's essential to prioritize responsible AI development and deployment.
```

Local Models with Ollama

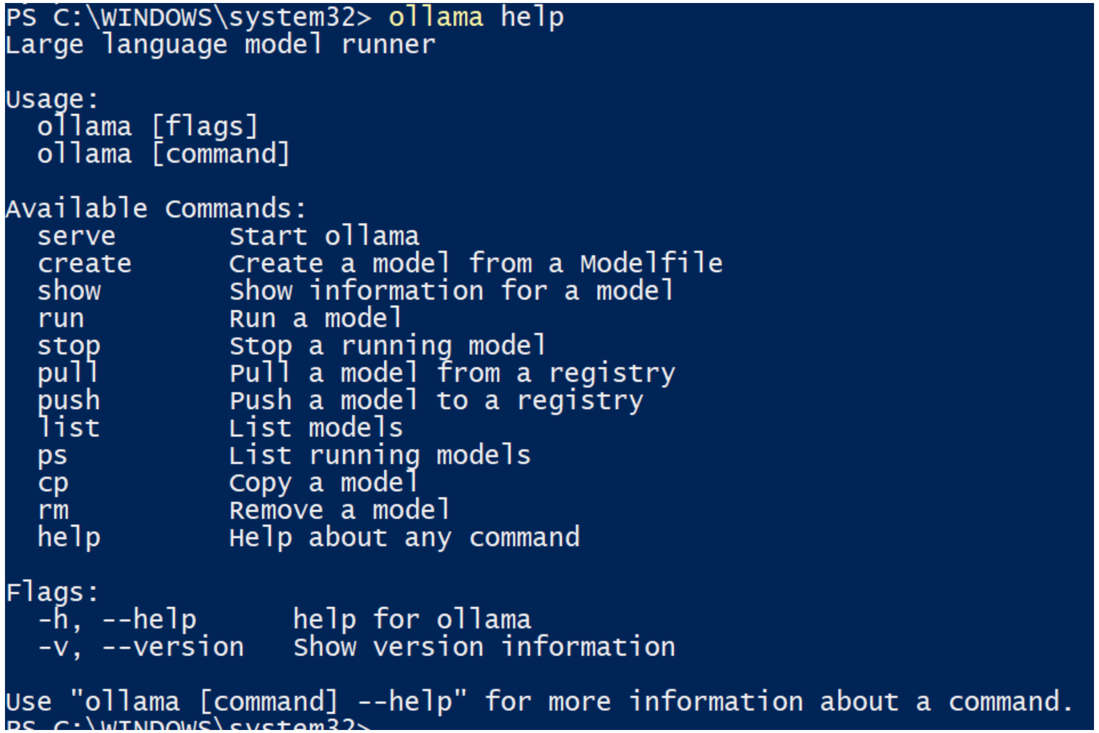
- A bunch of free models available

Model	Parameters	Size	Download
Llama 3.3	70B	43GB	ollama run llama3.3
Llama 3.2	3B	2.0GB	ollama run llama3.2
Llama 3.2	1B	1.3GB	ollama run llama3.2:1b
Llama 3.2 Vision	11B	7.9GB	ollama run llama3.2-vision
Llama 3.2 Vision	90B	55GB	ollama run llama3.2-vision:90b
Llama 3.1	8B	4.7GB	ollama run llama3.1
Llama 3.1	405B	231GB	ollama run llama3.1:405b
Phi 4	14B	9.1GB	ollama run phi4
Phi 3 Mini	3.8B	2.3GB	ollama run phi3

- Some are even multimodal models (e.g. Llava) (can accept image as input)
- Check out the model search page for more : https://ollama.com/search
- However, depends on your computer/hardware/GPU capabilities, it may not run properly, or it may run but **extremely SLOW** !!!

Local Models with Ollama

- You can check out more commands of Ollama from “ollama help”



```
PS C:\WINDOWS\system32> ollama help
Large language model runner

Usage:
  ollama [flags]
  ollama [command]

Available Commands:
  serve       Start ollama
  create      Create a model from a Modelfile
  show        Show information for a model
  run         Run a model
  stop        Stop a running model
  pull        Pull a model from a registry
  push        Push a model to a registry
  list        List models
  ps          List running models
  cp          Copy a model
  rm          Remove a model
  help        Help about any command

Flags:
  -h, --help      help for ollama
  -v, --version   Show version information

Use "ollama [command] --help" for more information about a command.
PS C:\WINDOWS\system32>
```

Or more details:

- [Ollama documentation](https://github.com/ollama/ollama/tree/main/docs) (https://github.com/ollama/ollama/tree/main/docs)

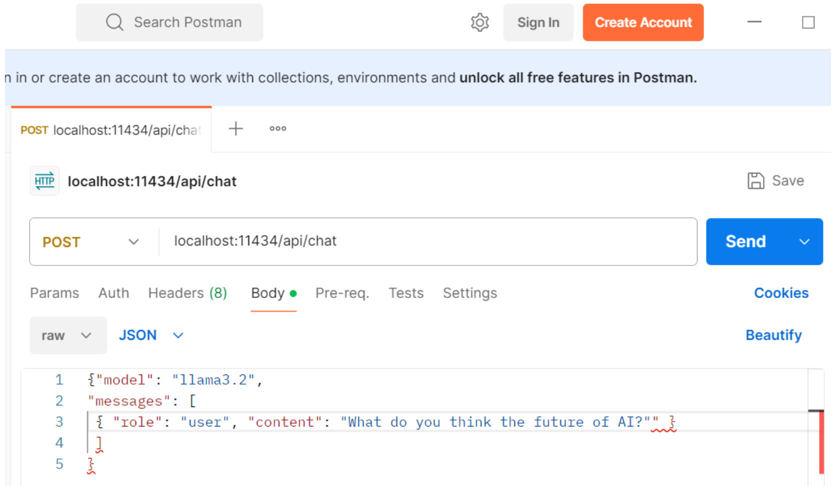
Run Ollama as Server

```
ollama serve
```

- Will start the Ollama server
- Access it using REST API
  - endpoints details: (https://github.com/ollama/ollama/blob/main/docs/api.md)

```
curl –X POST http://localhost:11434/api/generate -d '{
  "model": "llama3.2",
  "prompt":"What do you think the future of AI?"
}'
```

OR



n in or create an account to work with collections, environments and **unlock all free features in Postman.**

POST localhost:11434/api/chat

localhost:11434/api/chat

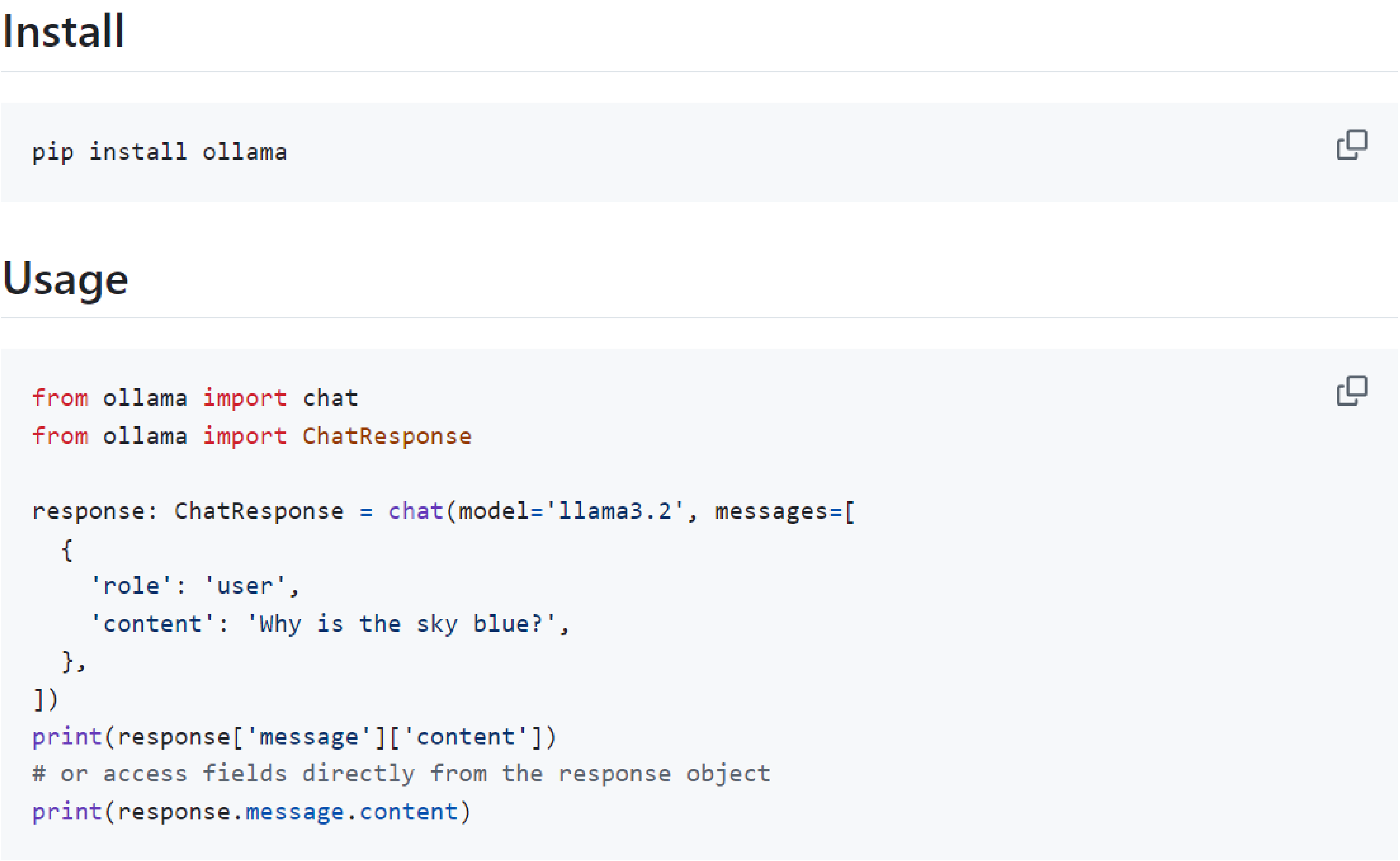
POST localhost:11434/api/chat

```
{"model": "llama3.2",
"messages": [
 { "role": "user", "content": "What do you think the future of AI?"" }
 ]
}
```

Ollama Python Libary

- You can also access to ollama from a Python program!
- Install the library with pip install ollama
- Import the library from your python program
- More details:

https://github.com/ollama/ollama-python



Install

```
pip install ollama
```

Usage

```
from ollama import chat
from ollama import ChatResponse

response: ChatResponse = chat(model='llama3.2', messages=[
  {
    'role': 'user',
    'content': 'Why is the sky blue?',
  },
])
print(response['message']['content'])
# or access fields directly from the response object
print(response.message.content)
```