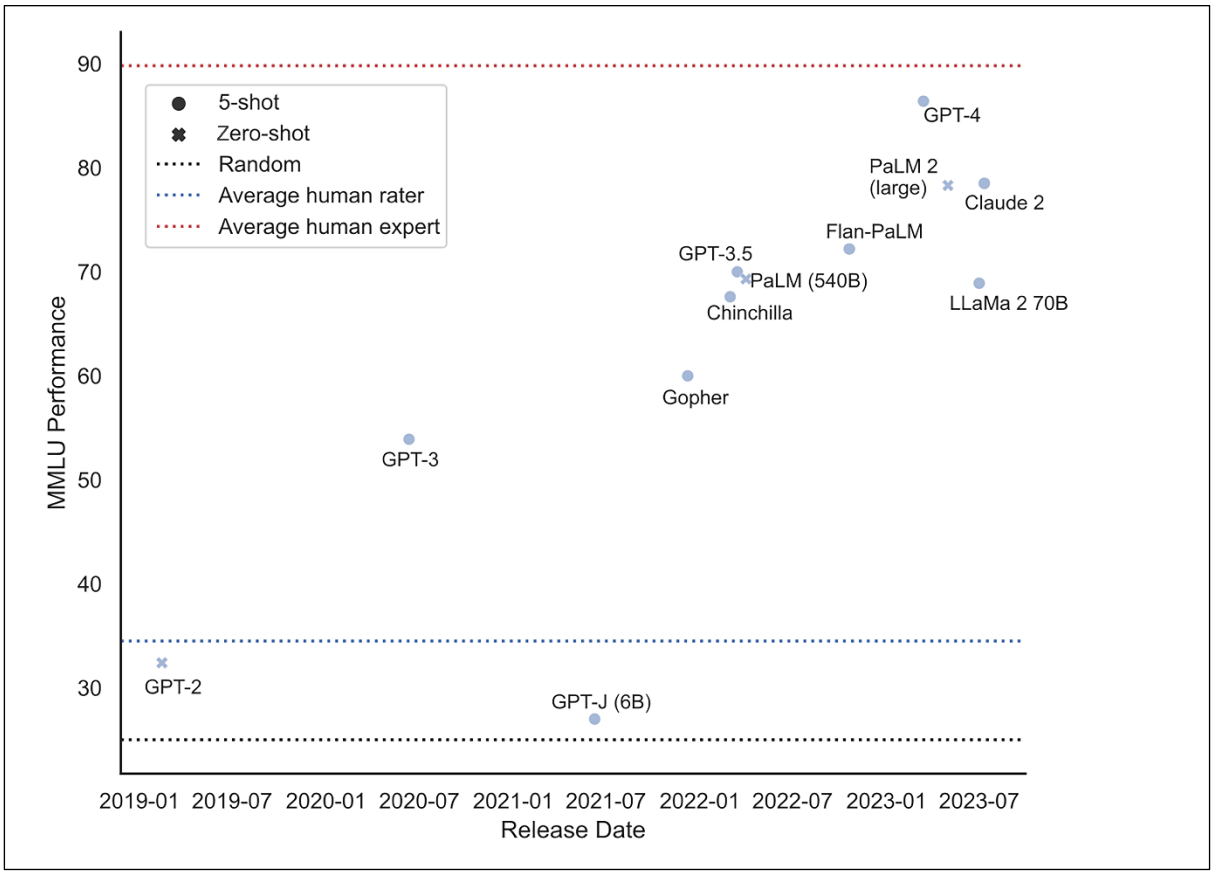
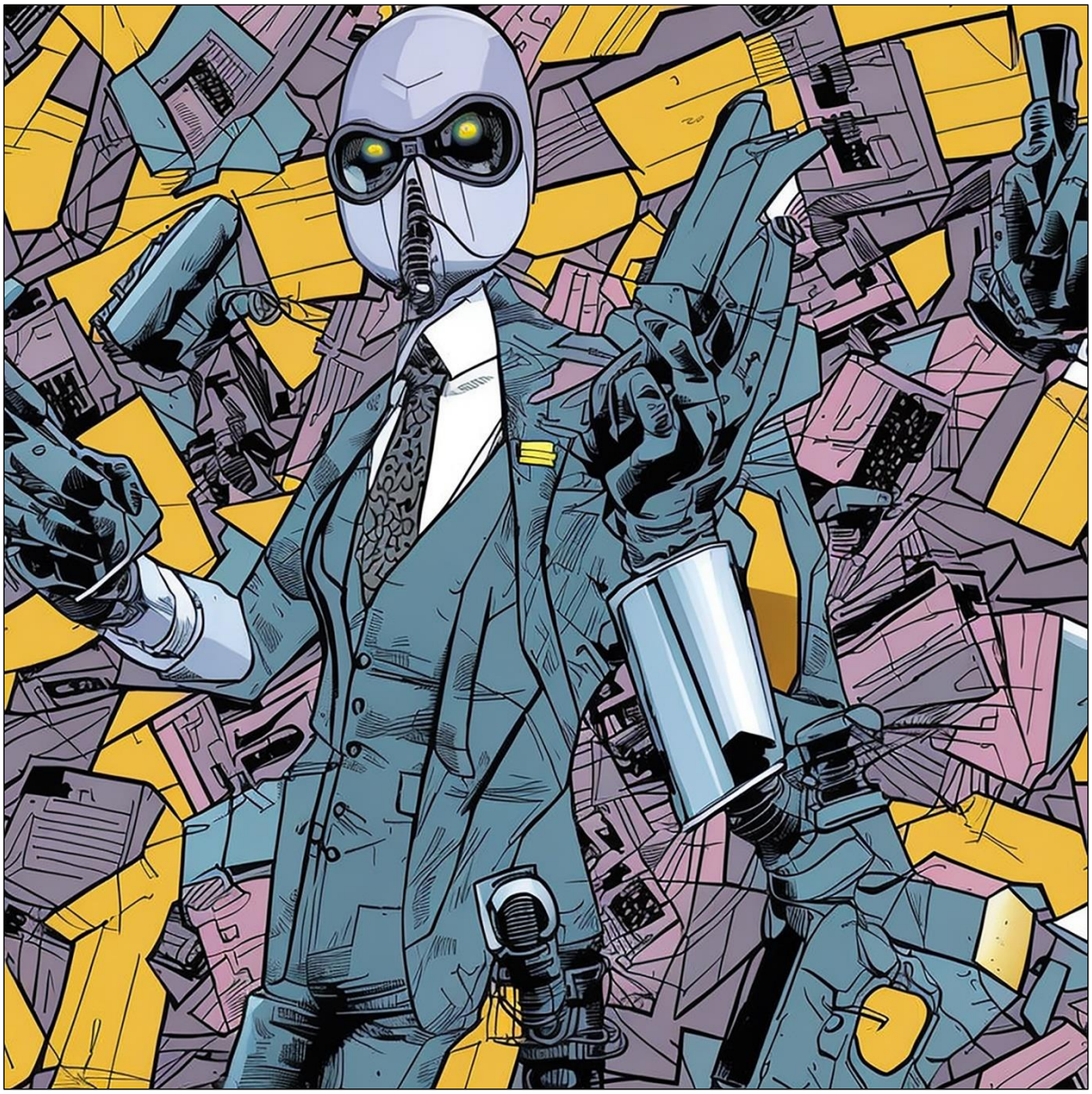
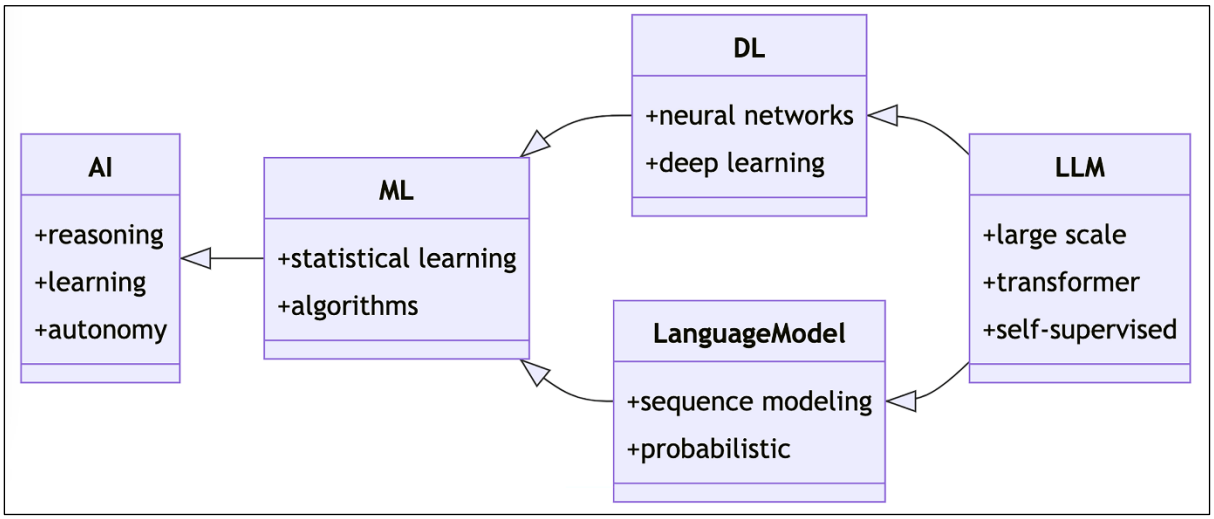
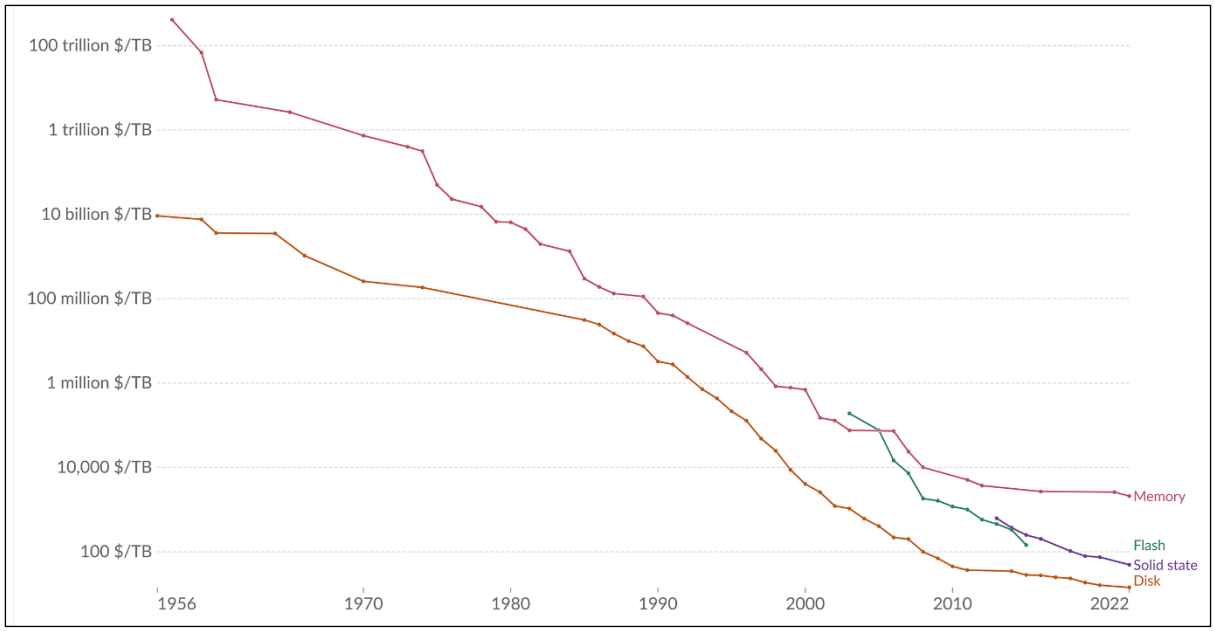
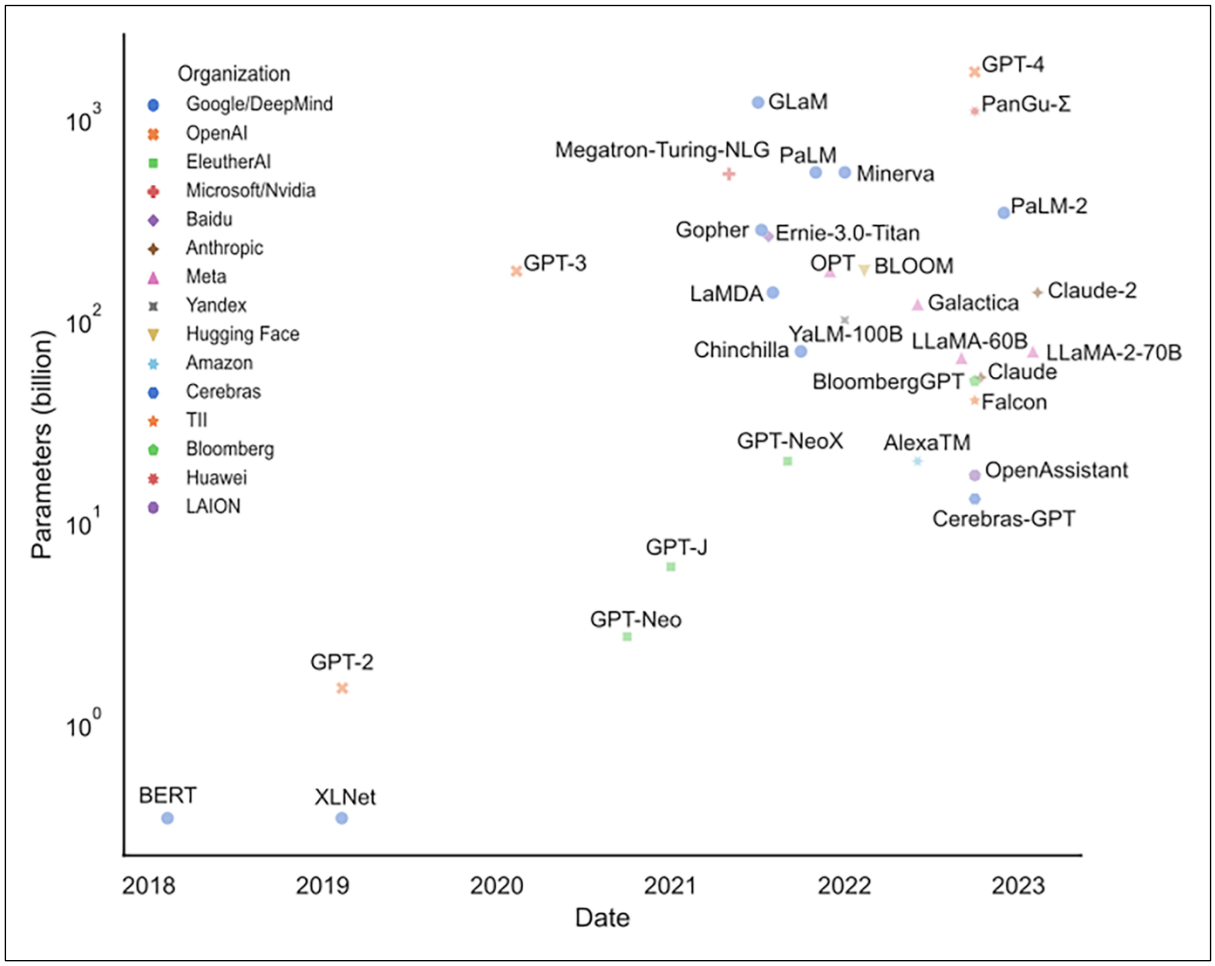
Chapter 1: What Is Generative AI?

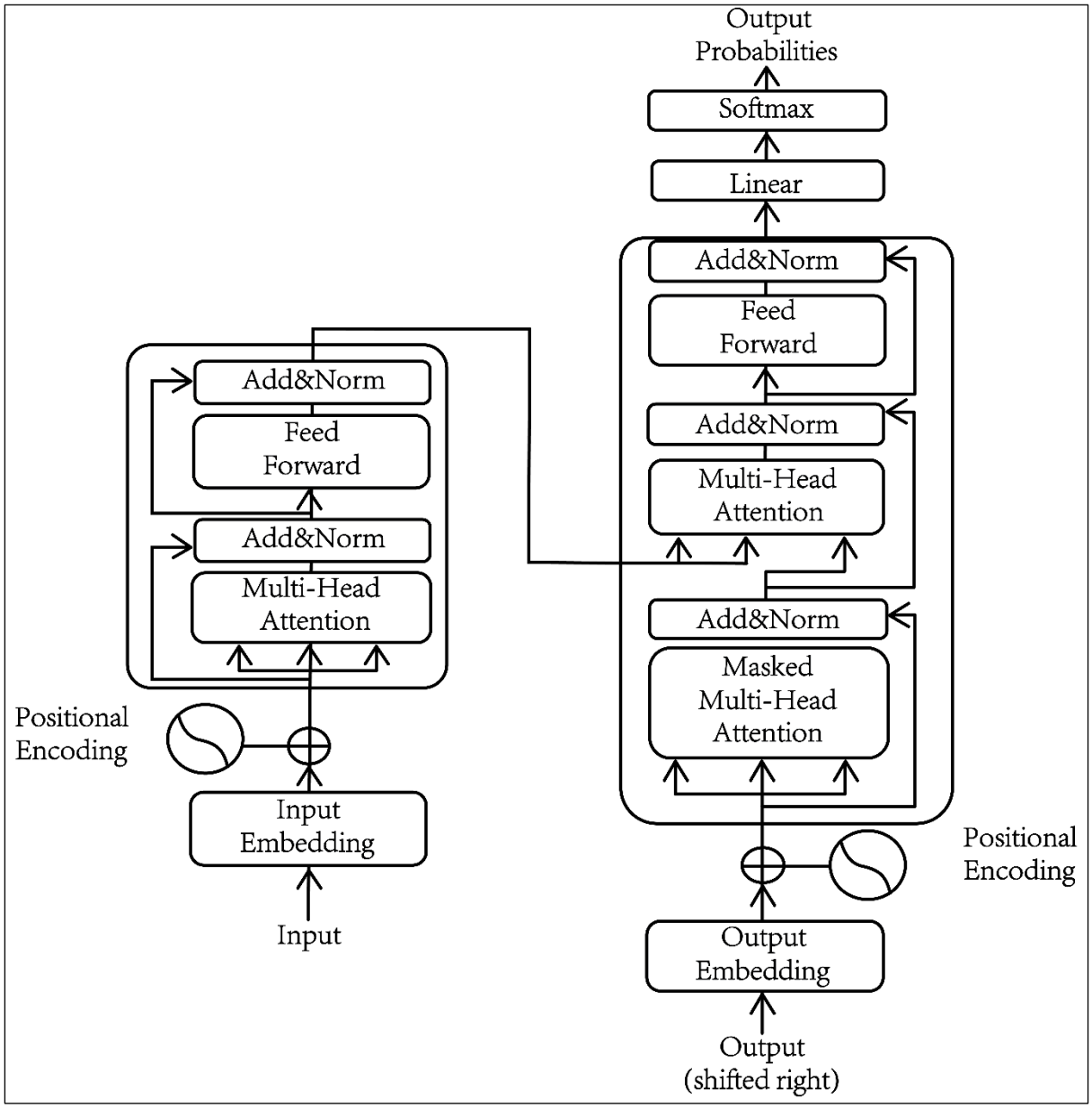


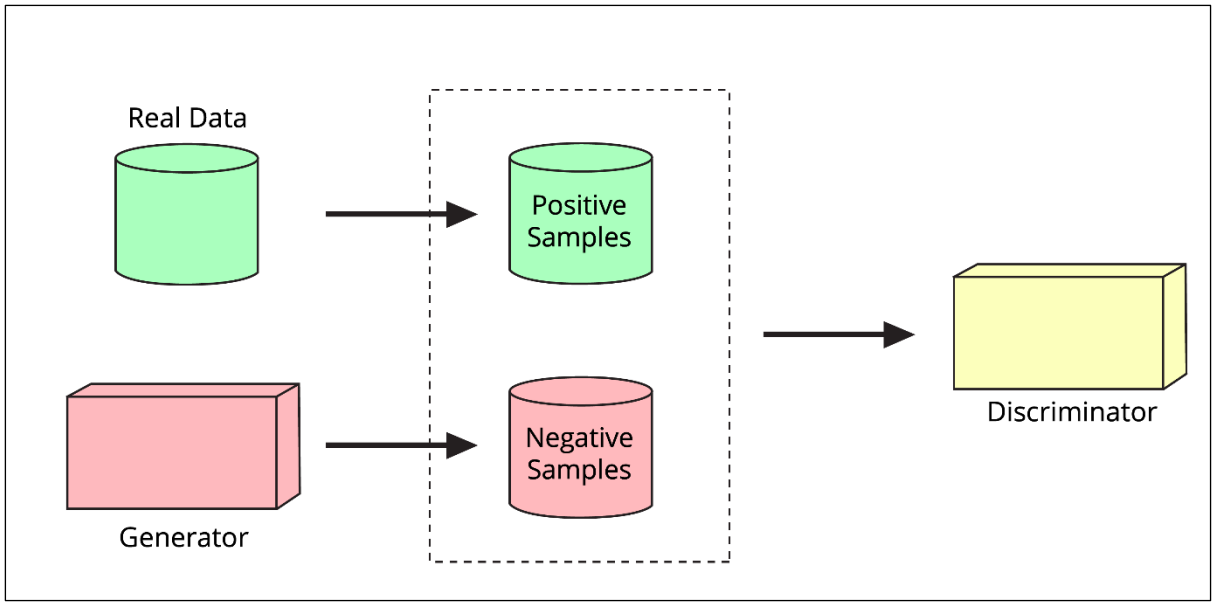


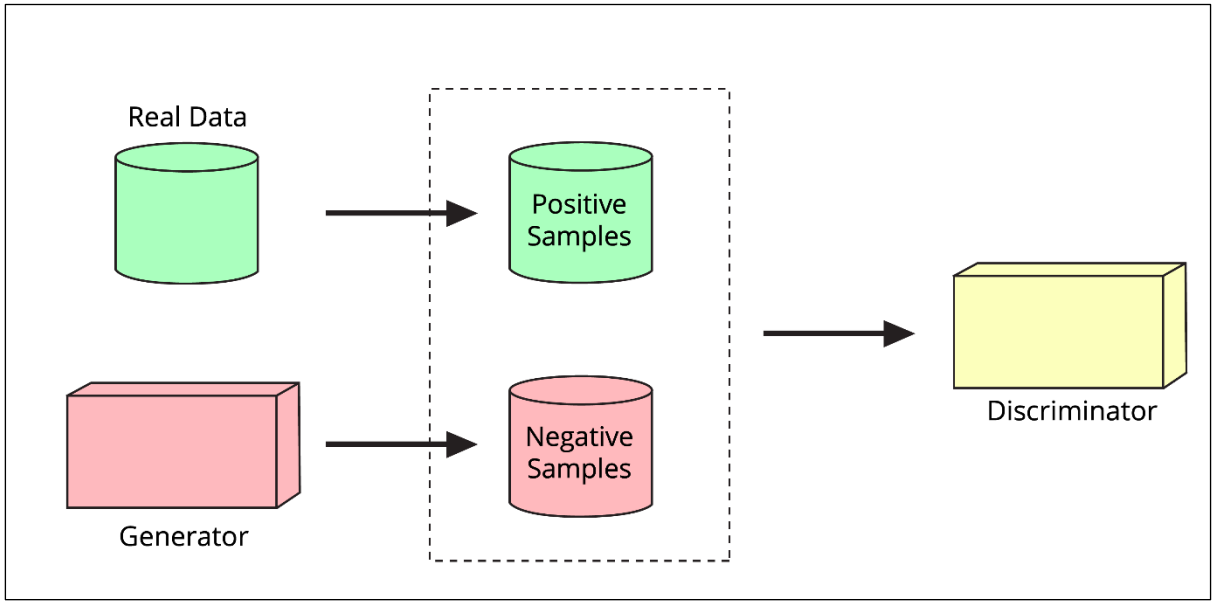

DL
+neural networks
+deep learning
AI
+reasoning
+learning
+autonomy
ML
+statistical learning
+algorithms
LLM
+large scale
+transformer
+self-supervised
LanguageModel
+sequence modeling
+probabilistic

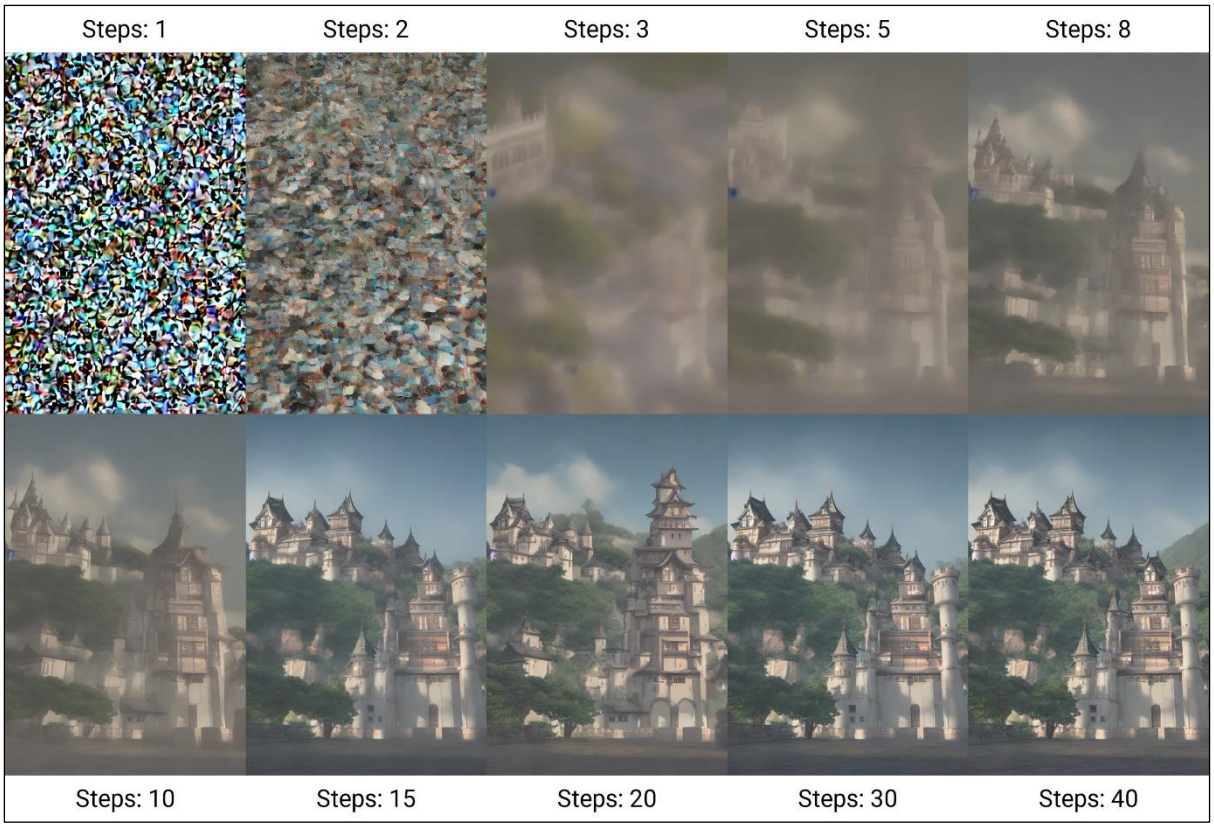

100 trillion $/TB
1 trillion $/TB
10 billion $/TB
100 million $/TB
1 million $/TB
10,000 $/TB
100 $/TB
1956
1970
1980
1990
2000
2010
2022
Memory
Flash
Solid state
Disk


Organization
Google/DeepMind
OpenAI
EleutherAI
Microsoft/Nvidia
Baidu
Anthropic
Meta
Yandex
Hugging Face
Amazon
Cerebras
TII
Bloomberg
Huawei
LAION
Parameters (billion)
10^3
10^2
10^1
10^0
GPT-4
GLaM
PanGu-Σ
Megatron-Turing-NLG
PaLM
Minerva
PaLM-2
Gopher
Ernie-3.0-Titan
GPT-3
OPT
BLOOM
LaMDA
Galactica
Claude-2
YaLM-100B
LLaMA-60B
LLaMA-2-70B
Chinchilla
BloombergGPT
Claude
Falcon
GPT-NeoX
AlexaTM
OpenAssistant
Cerebras-GPT
GPT-J
GPT-Neo
GPT-2
BERT
XLNet
2018
2019
2020
2021
2022
2023
Date


Output
Probabilities
Softmax
Linear
Add&Norm
Feed
Forward
Add&Norm
Multi-Head
Attention
Add&Norm
Masked
Multi-Head
Attention
Positional
Encoding
Output
Embedding
Output
(shifted right)
Add&Norm
Feed
Forward
Add&Norm
Multi-Head
Attention
Positional
Encoding
Input
Embedding
Input


Real Data
Positive Samples
Negative Samples
Generator
Discriminator


Real Data
Positive Samples
Negative Samples
Generator
Discriminator


Steps: 1
Steps: 2
Steps: 3
Steps: 5
Steps: 8
Steps: 10
Steps: 15
Steps: 20
Steps: 30
Steps: 40


vibrant portrait painting of Salvador Dalí with a robotic half face
a shiba inu wearing a beret and black turtleneck

Chapter 2: LangChain for LLM Apps

You

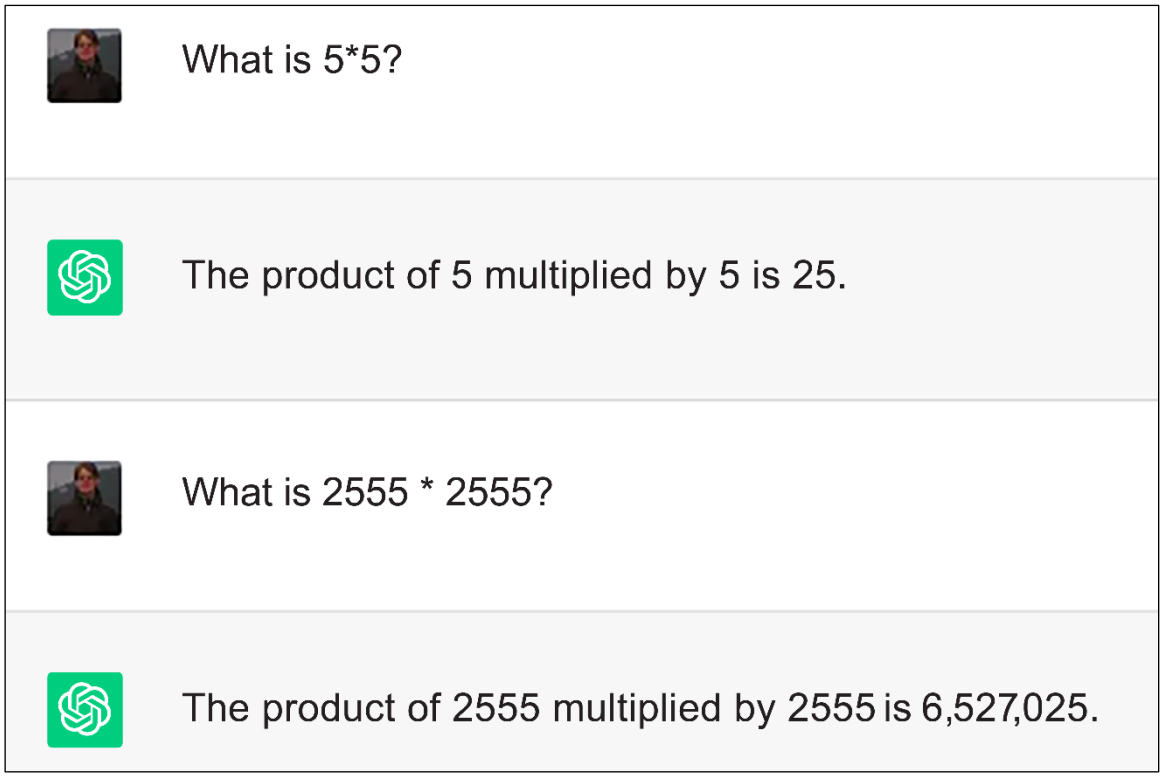
What's LangChain?

ChatGPT

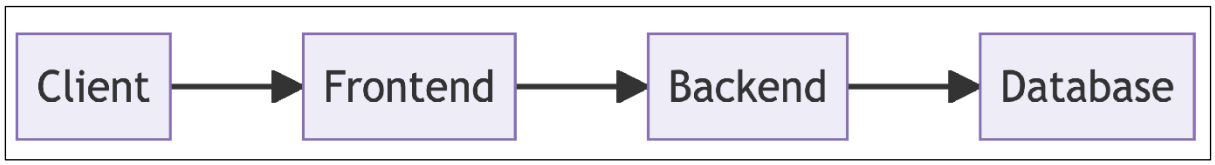
As of my last knowledge update in January 2022, I don't have specific information about "LangChain." It's possible that there have been developments or new entities with that name since then.

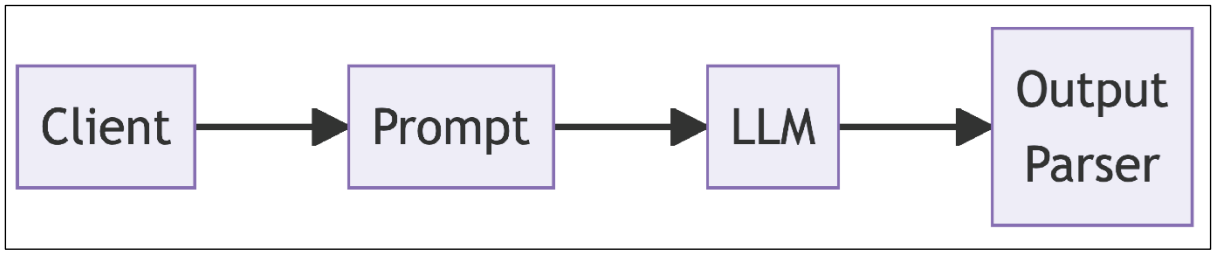
If LangChain is a term or entity that emerged or gained significance after my last update, I recommend checking the latest online sources, such as news articles, official websites, or other reliable platforms for the most recent information. If it's a relatively new or niche topic, it might be best to look for specialized sources or announcements related to the field or industry associated with LangChain.

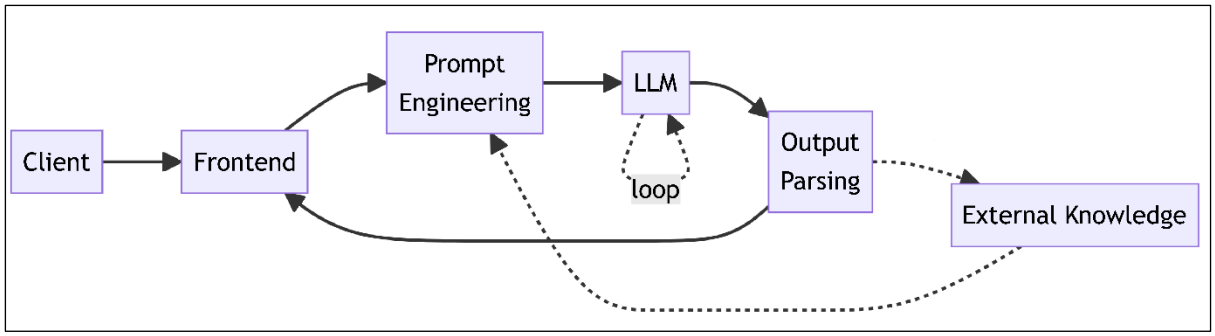
USER	What is LangChain?
ASSISTANT	LangChain is a decentralized platform that aims to revolutionize the global language learning industry. It leverages blockchain technology to provide a secure and transparent learning environment, connecting language learners directly with native speakers and qualified tutors. The platform offers various language learning tools, including interactive lessons, practice exercises, and real-time conversations, all facilitated by its decentralized network. LangChain also incorporates a reward system using its native cryptocurrency, enabling learners to earn tokens by achieving language learning milestones and facilitating peer-to-peer transactions within the platform. Overall, LangChain aims to provide an efficient, affordable, and engaging language learning experience for users worldwide.

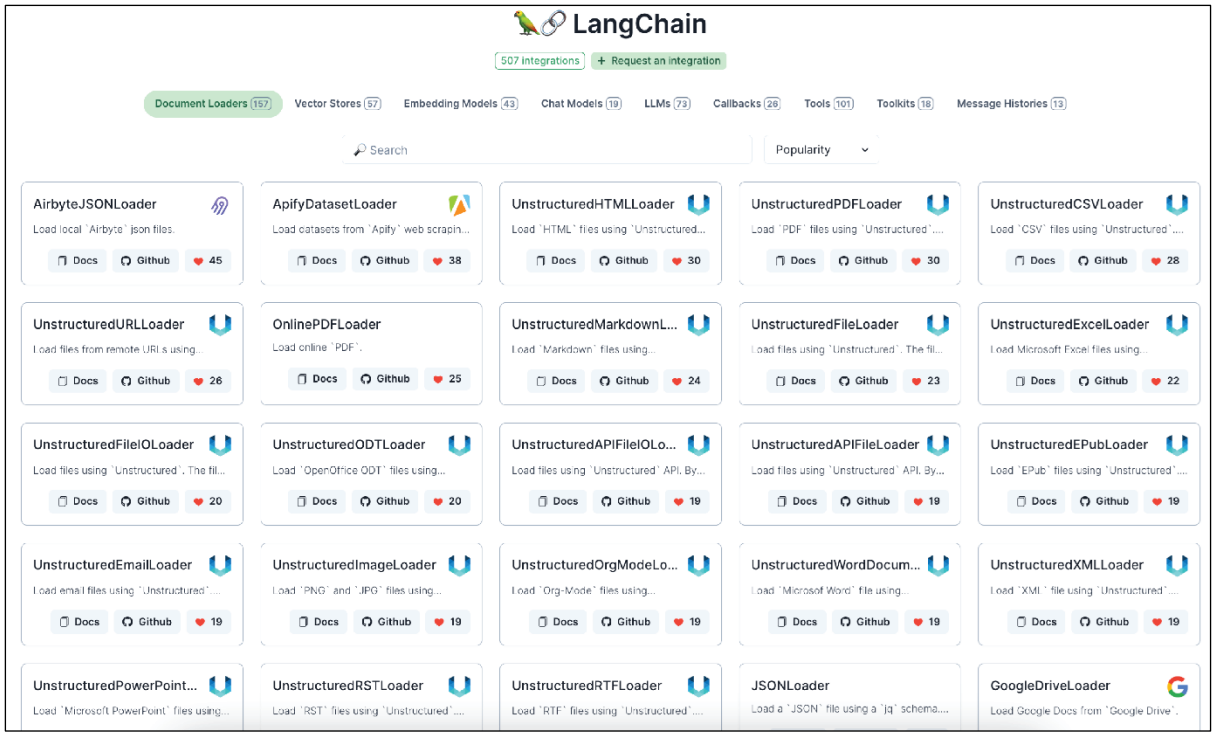


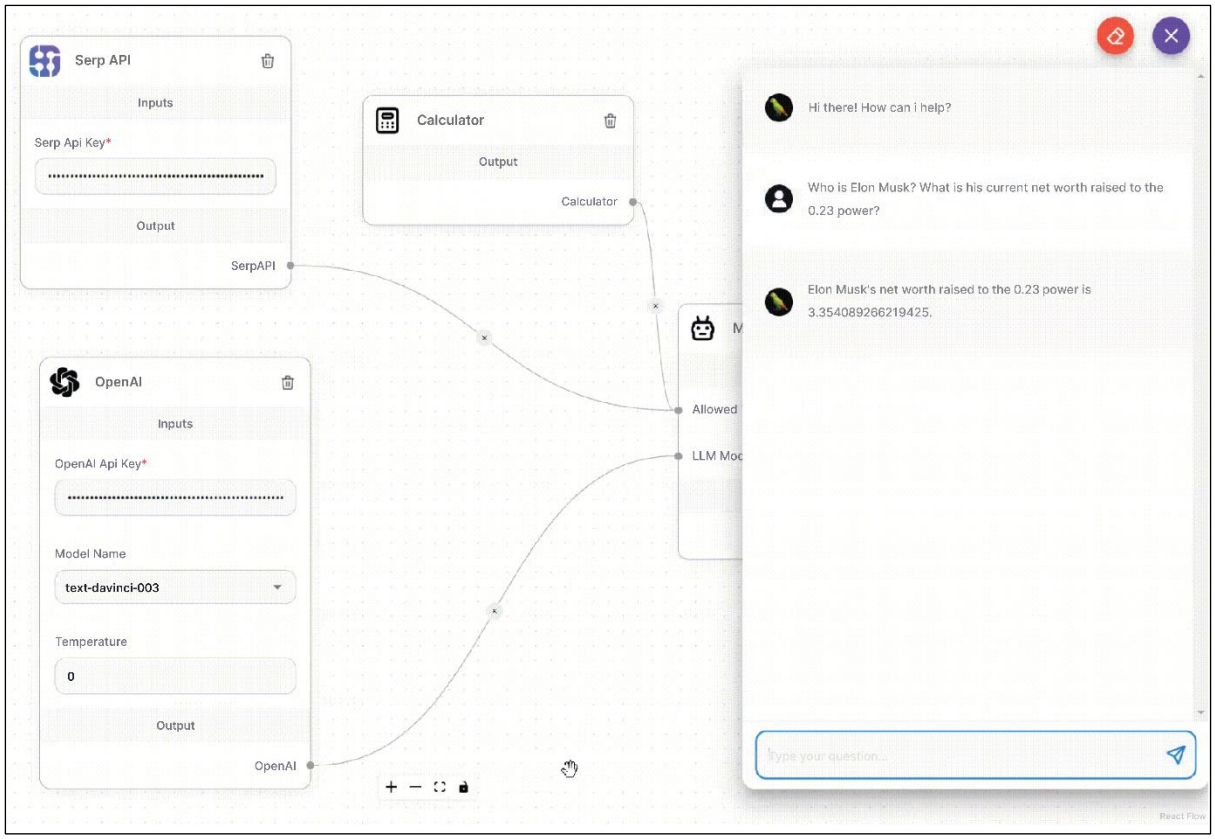

```
(base) ~ % bc -l
bc 1.06
Copyright 1991-1994, 1997, 1998, 2000 Free Software Foundation, Inc.
This is free software with ABSOLUTELY NO WARRANTY.
For details type `warranty'.
2555 * 2555
6528025
```

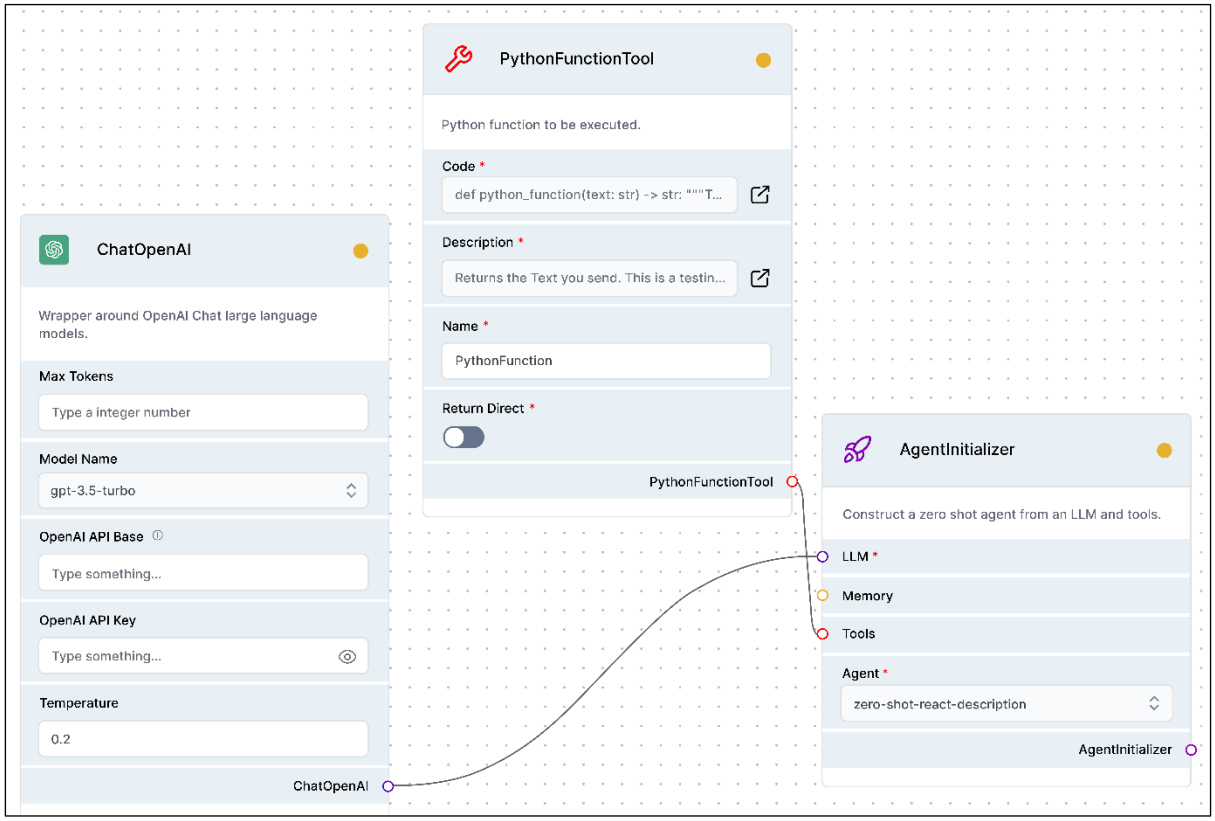


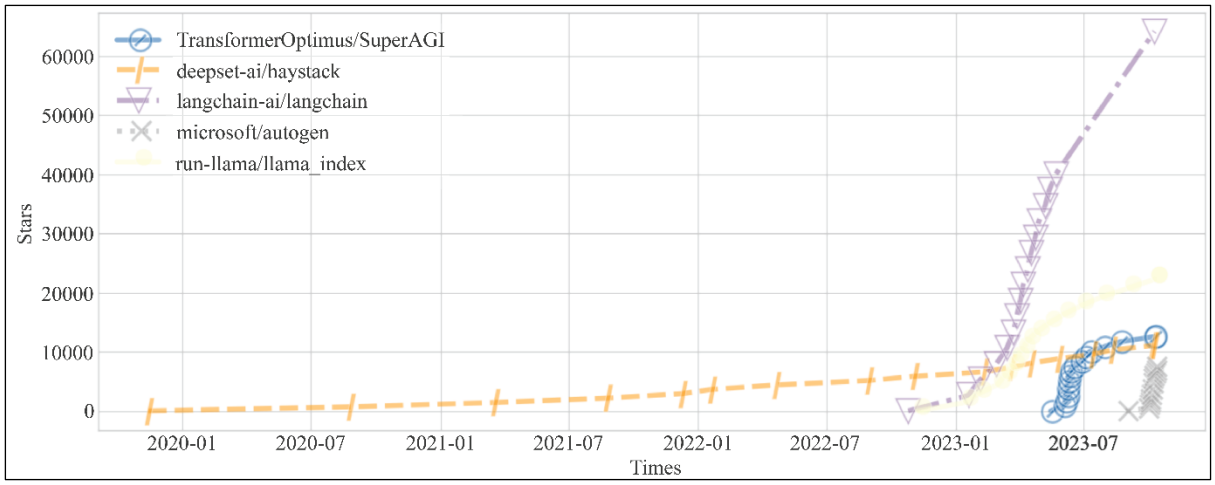



Client
Frontend
Prompt Engineering
LLM
loop
Output Parsing
External Knowledge

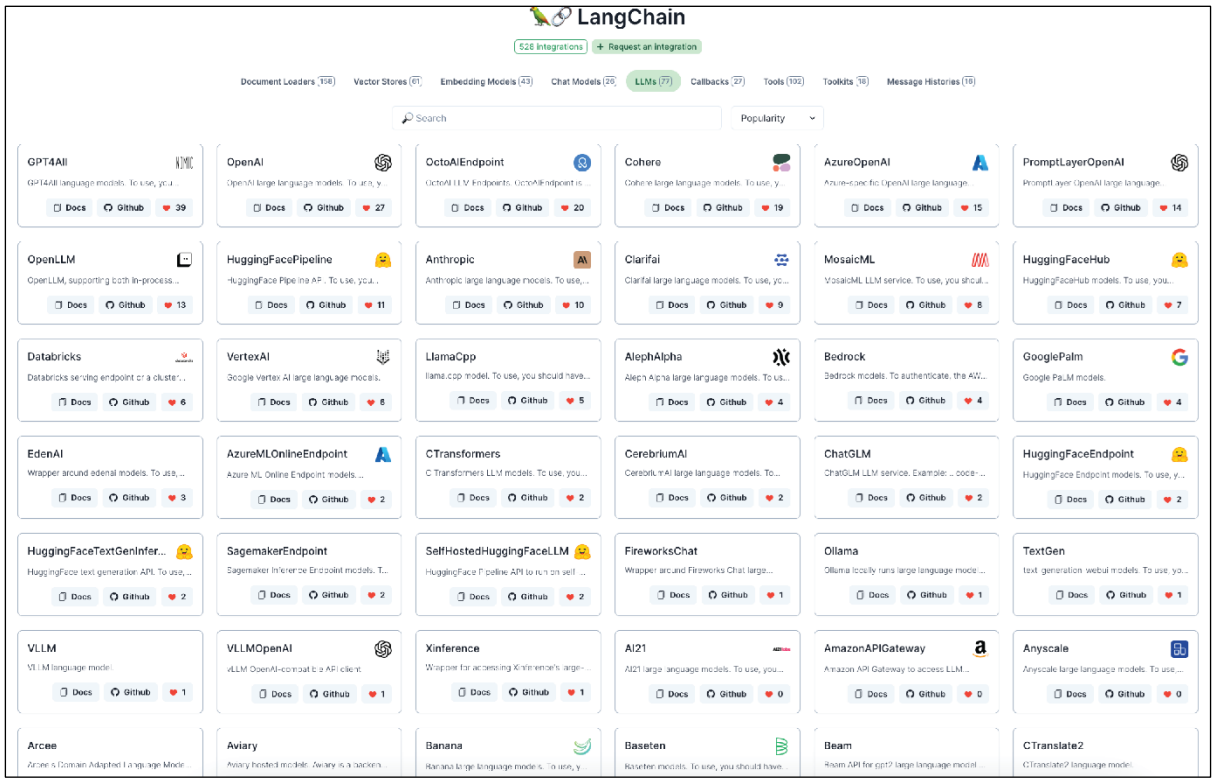

LangChain
507 integrations
+ Request an integration
Document Loaders 157
Vector Stores 57
Embedding Models 43
Chat Models 19
LLMs 73
Callbacks 26
Tools 101
Toolkits 18
Message Histories 13
Search
Popularity
AirbyteJSONLoader
Load local `Airbyte` json files.
ApifyDatasetLoader
Load datasets from `Apify` web scrapin...
UnstructuredHTMLLoader
Load `HTML` files using `Unstructured`...
UnstructuredPDFLoader
Load `PDF` files using `Unstructured`....
UnstructuredCSVLoader
Load `CSV` files using `Unstructured`....
UnstructuredURLLoader
Load files from remote URLs using...
OnlinePDFLoader
Load online `PDF`.
UnstructuredMarkdownL...
Load `Markdown` files using...
UnstructuredFileLoader
Load files using `Unstructured`. The fil...
UnstructuredExcelLoader
Load Microsoft Excel files using...
UnstructuredFileIOLoader
Load files using `Unstructured`. The fil...
UnstructuredODTLoader
Load `OpenOffice ODT` files using...
UnstructuredAPIFileIOLo...
Load files using `Unstructured` API. By...
UnstructuredAPIFileLoader
Load files using `Unstructured` API. By...
UnstructuredEPubLoader
Load `EPub` files using `Unstructured`....
UnstructuredEmailLoader
Load email files using `Unstructured`....
UnstructuredImageLoader
Load `PNG` and `JPG` files using...
UnstructuredOrgModeLo...
Load `Org-Mode` files using...
UnstructuredWordDocum...
Load `Microsoft Word` file using...
UnstructuredXMLLoader
Load `XML` file using `Unstructured`....
UnstructuredPowerPoint...
Load `Microsoft PowerPoint` files using...
UnstructuredRSTLoader
Load `RST` files using `Unstructured`....
UnstructuredRTFLoader
Load `RTF` files using `Unstructured`....
JSONLoader
Load a `JSON` file using a `jq` schema....
GoogleDriveLoader
Load Google Docs from `Google Drive`.
Docs
Github

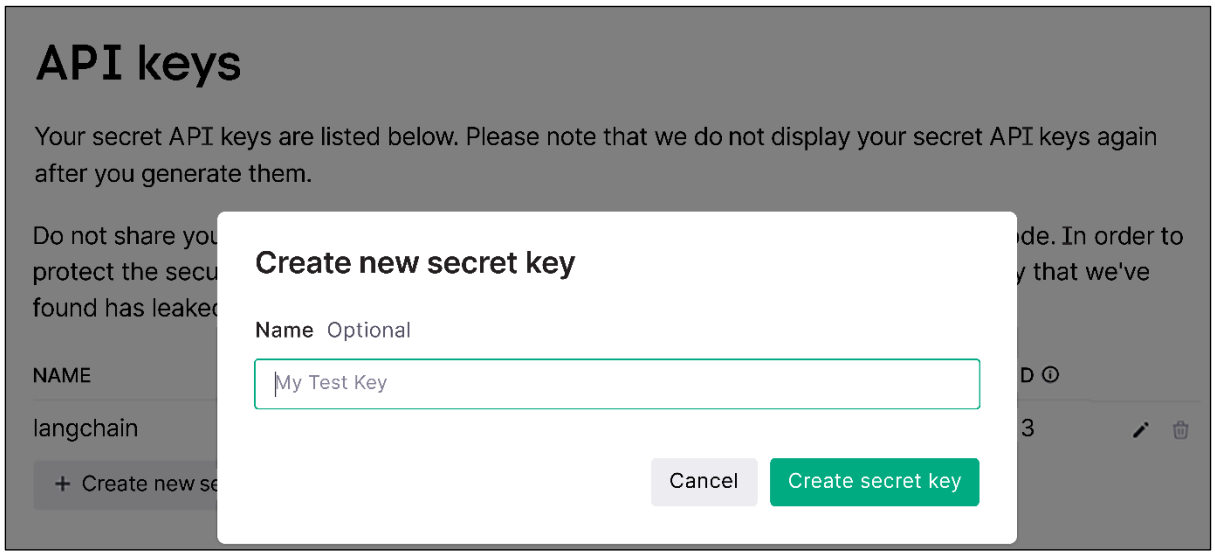

Serp API
Inputs
Serp Api Key*
Output
SerpAPI
Calculator
Output
Calculator
OpenAI
Inputs
OpenAI Api Key*
Model Name
text-davinci-003
Temperature
0
Output
OpenAI
Allowed
Hi there! How can i help?
Who is Elon Musk? What is his current net worth raised to the 0.23 power?
Elon Musk's net worth raised to the 0.23 power is 3.354089266219425.
Type your question...


PythonFunctionTool
Python function to be executed.
Code *
def python_function(text: str) -> str: """T...
Description *
Returns the Text you send. This is a testin...
Name *
PythonFunction
Return Direct *
PythonFunctionTool
ChatOpenAI
Wrapper around OpenAI Chat large language models.
Max Tokens
Type a integer number
Model Name
gpt-3.5-turbo
OpenAI API Base
Type something...
OpenAI API Key
Type something...
Temperature
0.2
ChatOpenAI
AgentInitializer
Construct a zero shot agent from an LLM and tools.
LLM *
Memory
Tools
Agent *
zero-shot-react-description
AgentInitializer


TransformerOptimus/SuperAGI
deepset-ai/haystack
langchain-ai/langchain
microsoft/autogen
run-llama/llama_index
Stars
60000
50000
40000
30000
20000
10000
0
2020-01
2020-07
2021-01
2021-07
2022-01
2022-07
2023-01
2023-07
Times

Chapter 3: Getting Started with LangChain





< Back

Create Inference API

* Inference API name

langchain

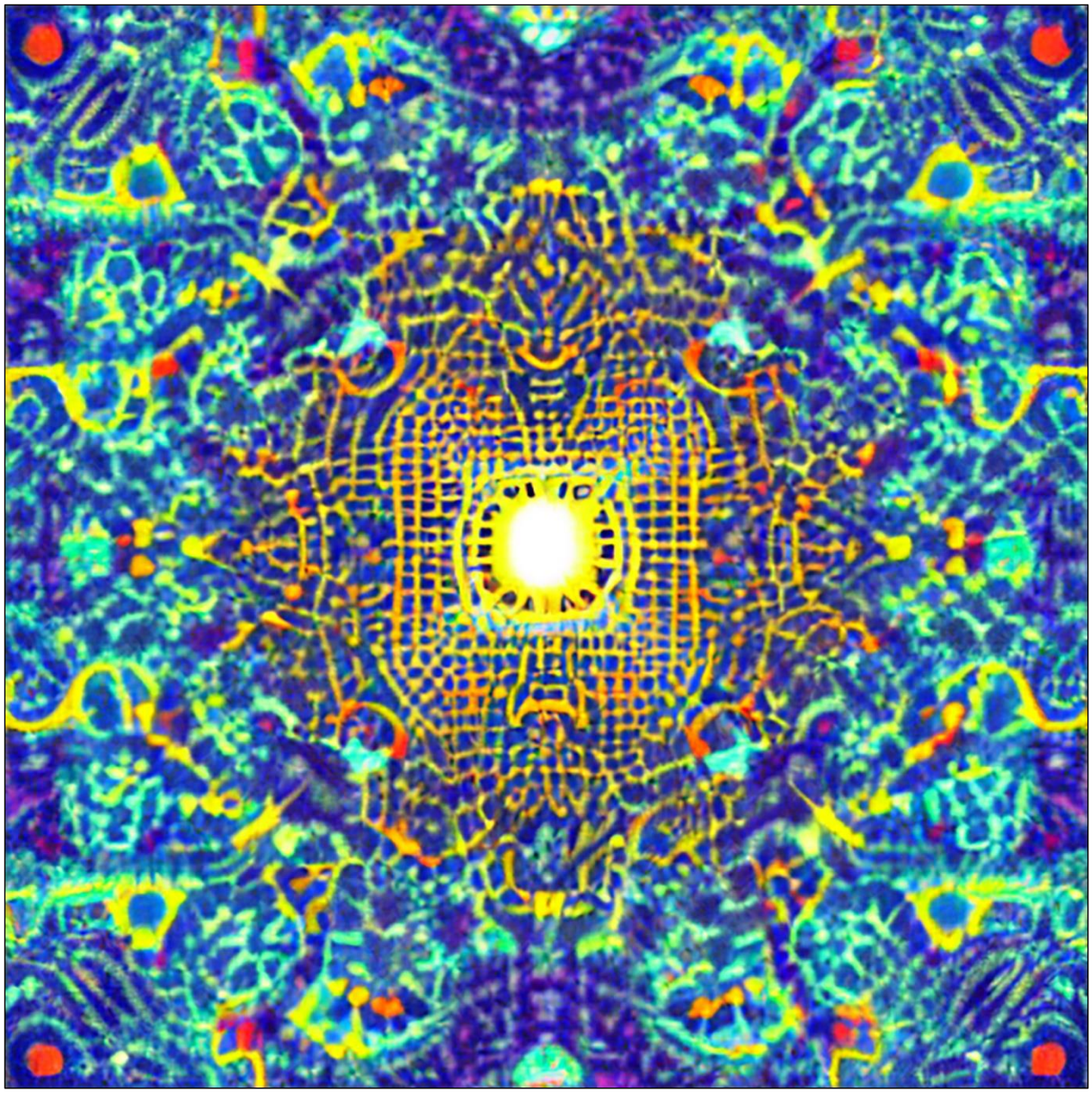
* Task

Visual Question Answering

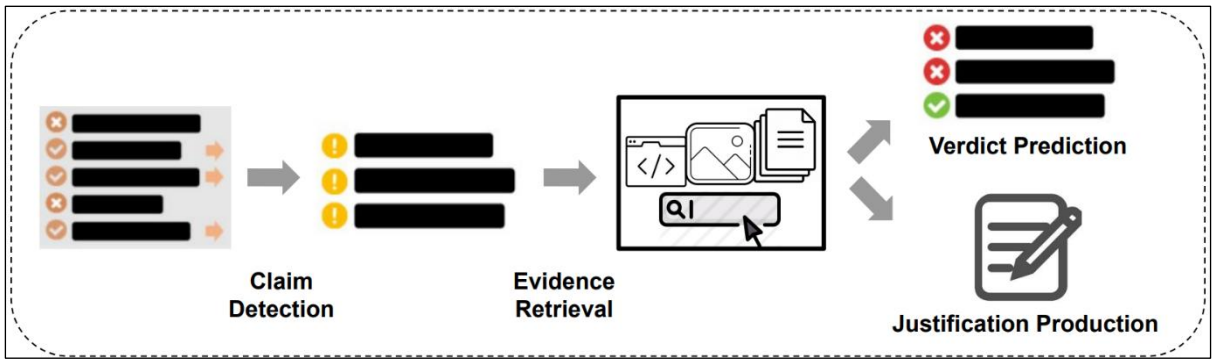
Understand the content of an image and answer questions about it in natural language.

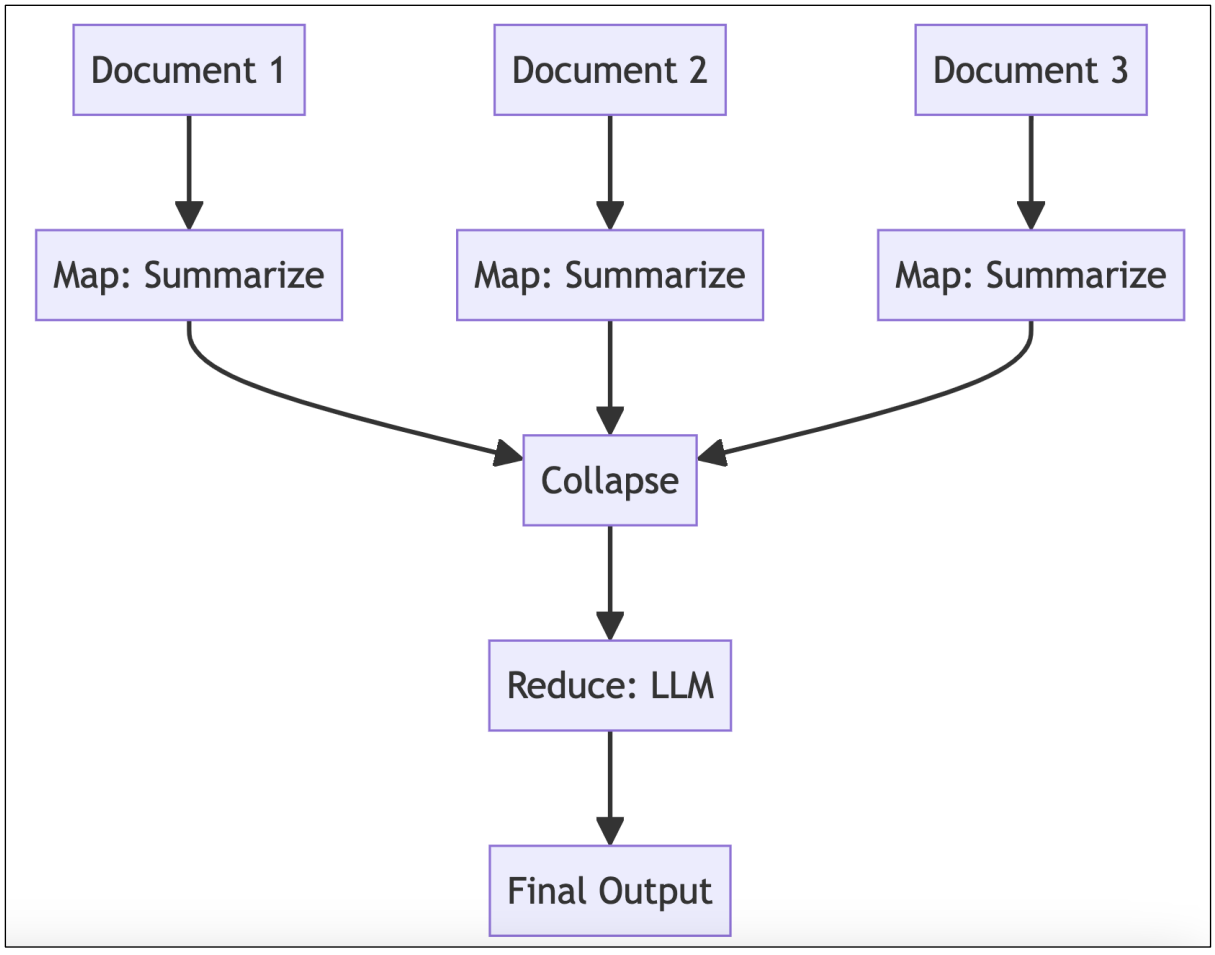
* Model

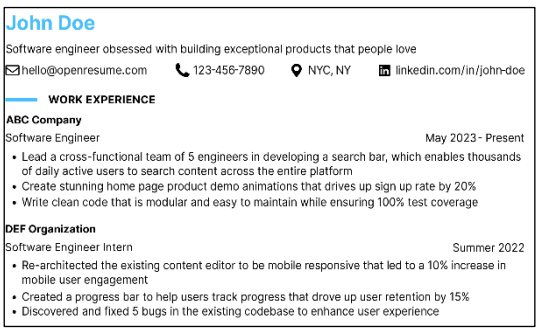
Salesforce/blip2-flan-t5-xl



Chapter 4: Building Capable Assistants







John Doe

Software engineer obsessed with building exceptional products that people love

hello@openresume.com | 123-456-7890 | NYC, NY | linkedin.com/in/john-doe

WORK EXPERIENCE

ABC Company

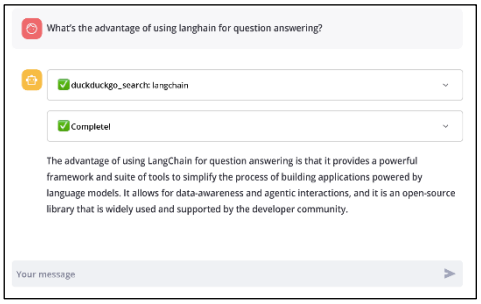
Software Engineer — May 2023 - Present

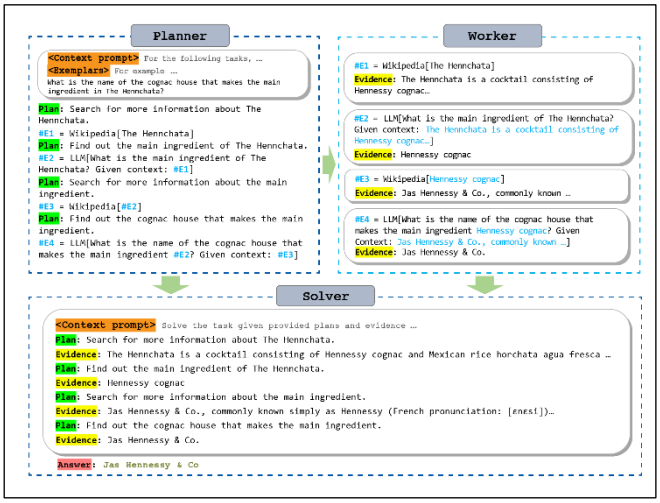
- Lead a cross-functional team of 5 engineers in developing a search bar, which enables thousands of daily active users to search content across the entire platform
- Create stunning home page product demo animations that drives up sign up rate by 20%
- Write clean code that is modular and easy to maintain while ensuring 100% test coverage

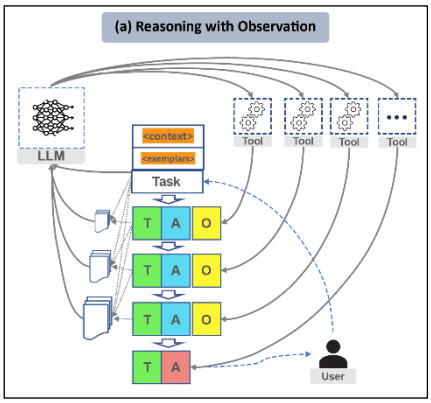
DEF Organization

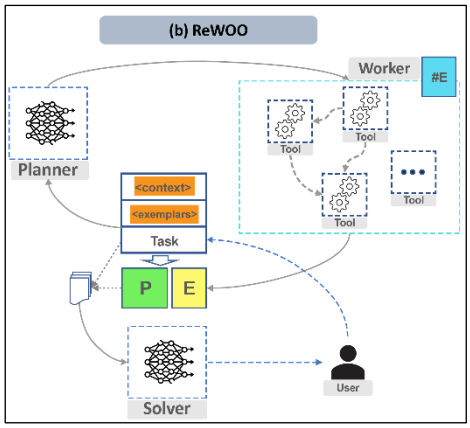
Software Engineer Intern — Summer 2022

- Re-architected the existing content editor to be mobile responsive that led to a 10% increase in mobile user engagement
- Created a progress bar to help users track progress that drove up user retention by 15%
- Discovered and fixed 5 bugs in the existing codebase to enhance user experience


What's the advantage of using langhain for question answering?
duckduckgo_search: langchain
Complete!
The advantage of using LangChain for question answering is that it provides a powerful framework and suite of tools to simplify the process of building applications powered by language models. It allows for data-awareness and agentic interactions, and it is an open-source library that is widely used and supported by the developer community.
Your message


Planner
<Context prompt> For the following tasks, …
<Exemplars> For example …
What is the name of the cognac house that makes the main ingredient in the Hennchata?
Plan: Search for more information about The Hennchata.
#E1 = Wikipedia[The Hennchata]
Plan: Find out the main ingredient of The Hennchata.
#E2 = LLM[What is the main ingredient of The Hennchata? Given context: #E1]
Plan: Search for more information about the main ingredient.
#E3 = Wikipedia[#E2]
Plan: Find out the cognac house that makes the main ingredient.
#E4 = LLM[What is the name of the cognac house that makes the main ingredient #E2? Given context: #E3]
Worker
#E1 = Wikipedia[The Hennchata]
Evidence: The Hennchata is a cocktail consisting of Hennessy cognac…
#E2 = LLM[What is the main ingredient of The Hennchata? Given context: The Hennchata is a cocktail consisting of Hennessy cognac…]
Evidence: Hennessy cognac
#E3 = Wikipedia[Hennessy cognac]
Evidence: Jas Hennessy & Co., commonly known …
#E4 = LLM[What is the name of the cognac house that makes the main ingredient Hennessy cognac? Given Context: Jas Hennessy & Co., commonly known …]
Evidence: Jas Hennessy & Co.
Solver
<Context prompt> Solve the task given provided plans and evidence …
Plan: Search for more information about The Hennchata.
Evidence: The Hennchata is a cocktail consisting of Hennessy cognac and Mexican rice horchata agua fresca …
Plan: Find out the main ingredient of The Hennchata.
Evidence: Hennessy cognac
Plan: Search for more information about the main ingredient.
Evidence: Jas Hennessy & Co., commonly known simply as Hennessy (French pronunciation: [ɛnesi])…
Plan: Find out the cognac house that makes the main ingredient.
Evidence: Jas Hennessy & Co.
Answer: Jas Hennessy & Co


(a) Reasoning with Observation
LLM
Tool
Tool
Tool
Tool
<context>
<exemplars>
Task
T
A
O
User



Ask a research question!

Reasoning strategy

- (•) plan-and-solve
- ( ) zero-shot-react

Which tools do you want to use?

ddg-search × wolfram-alpha × wikipedia ×

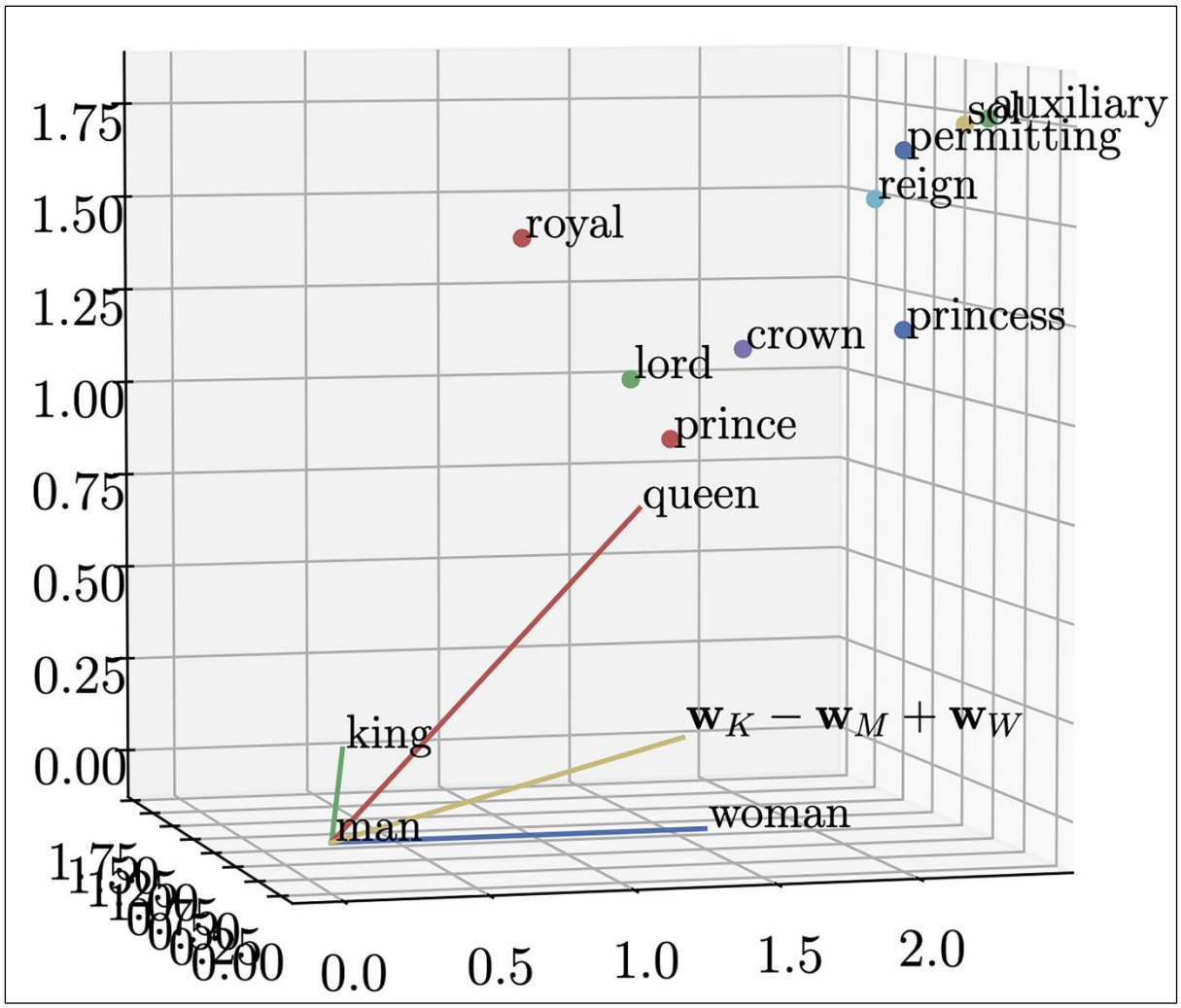
✓ **Complete!**

✓ **Complete!**

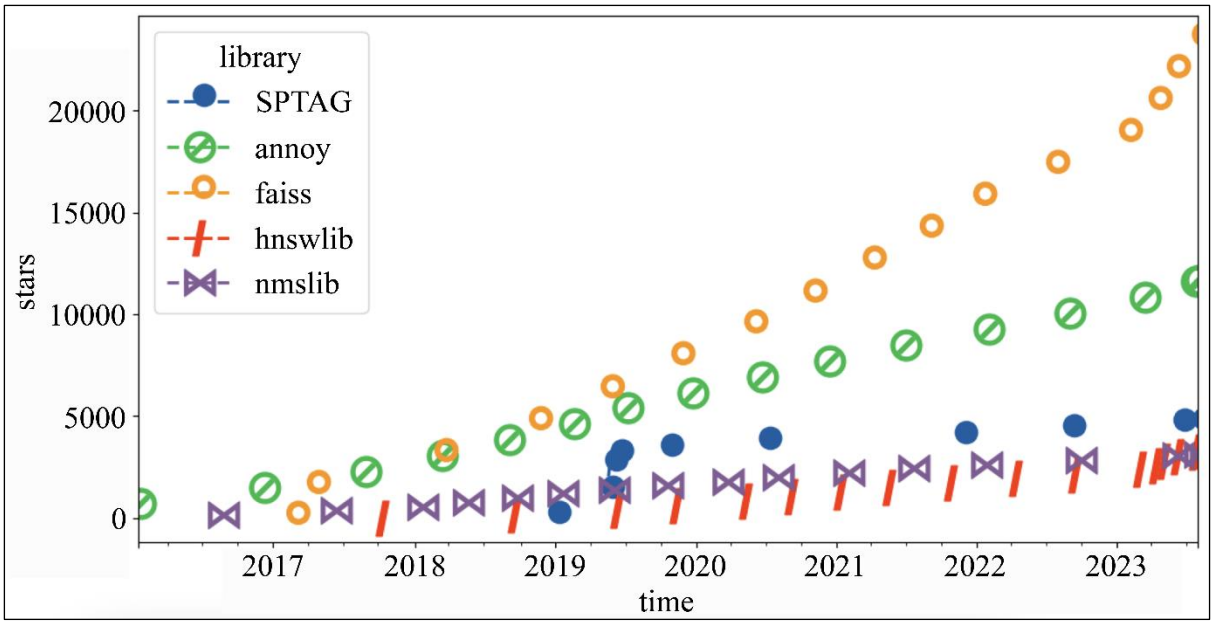
✓ **Wikipedia:** solve agent (large language models)

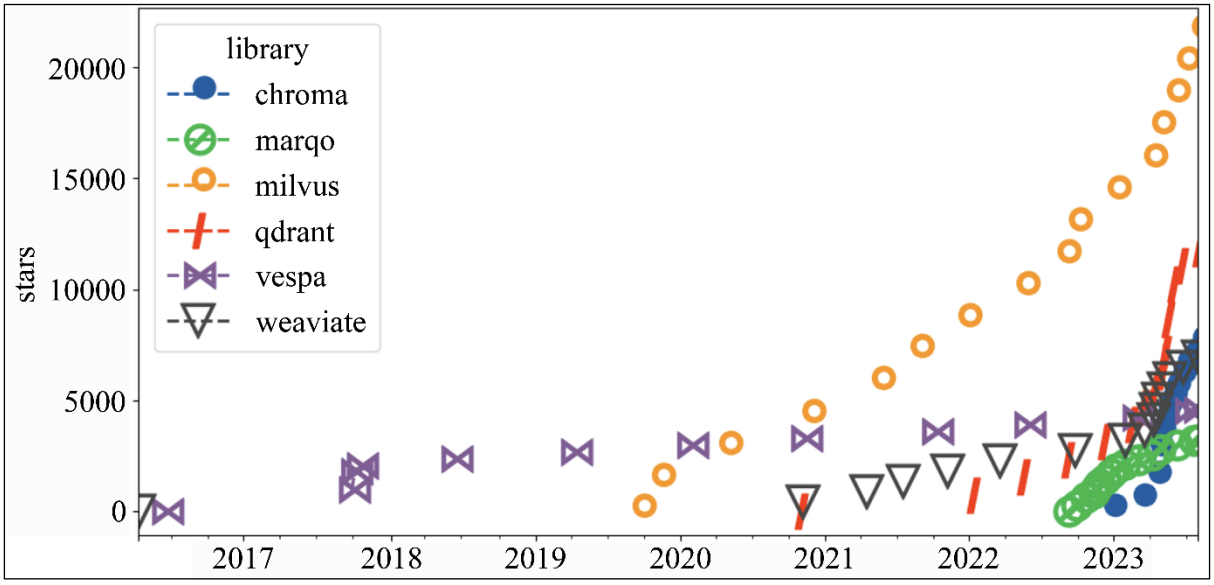
Ask me anything!

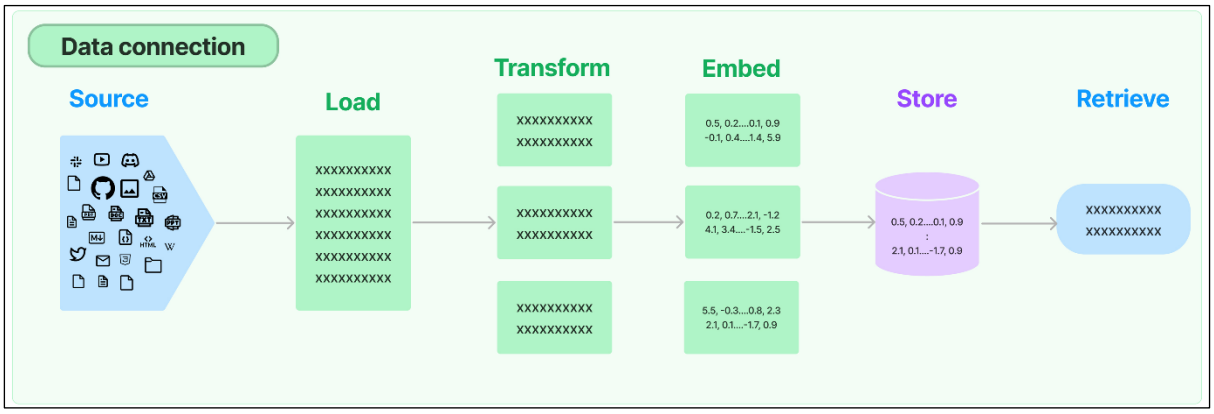
Chapter 5: Building a Chatbot like ChatGPT

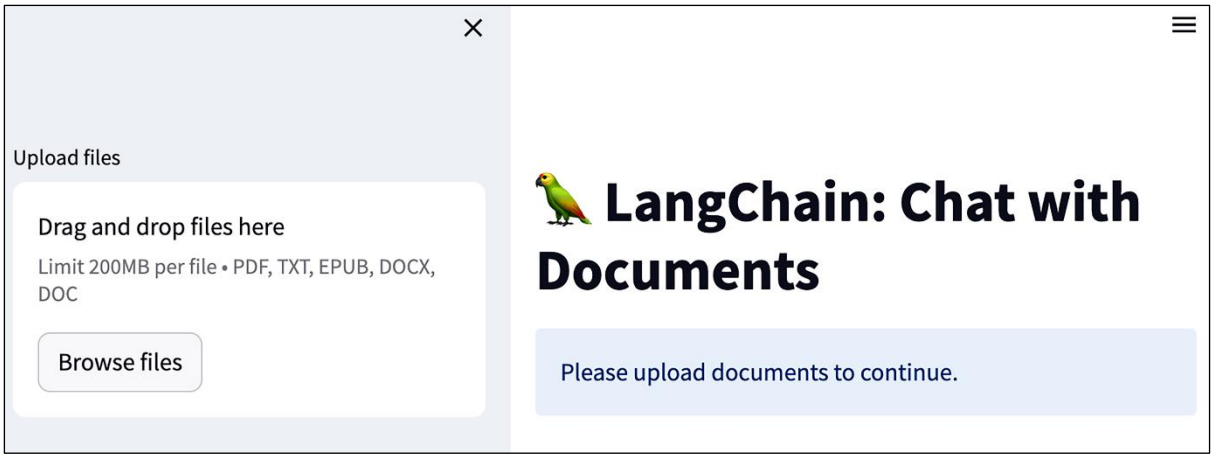


	cat	dog	computer	animal
cat	0.000000	0.522352	0.575285	0.521214
dog	0.522352	0.000000	0.581203	0.478794
computer	0.575285	0.581203	0.000000	0.591435
animal	0.521214	0.478794	0.591435	0.000000

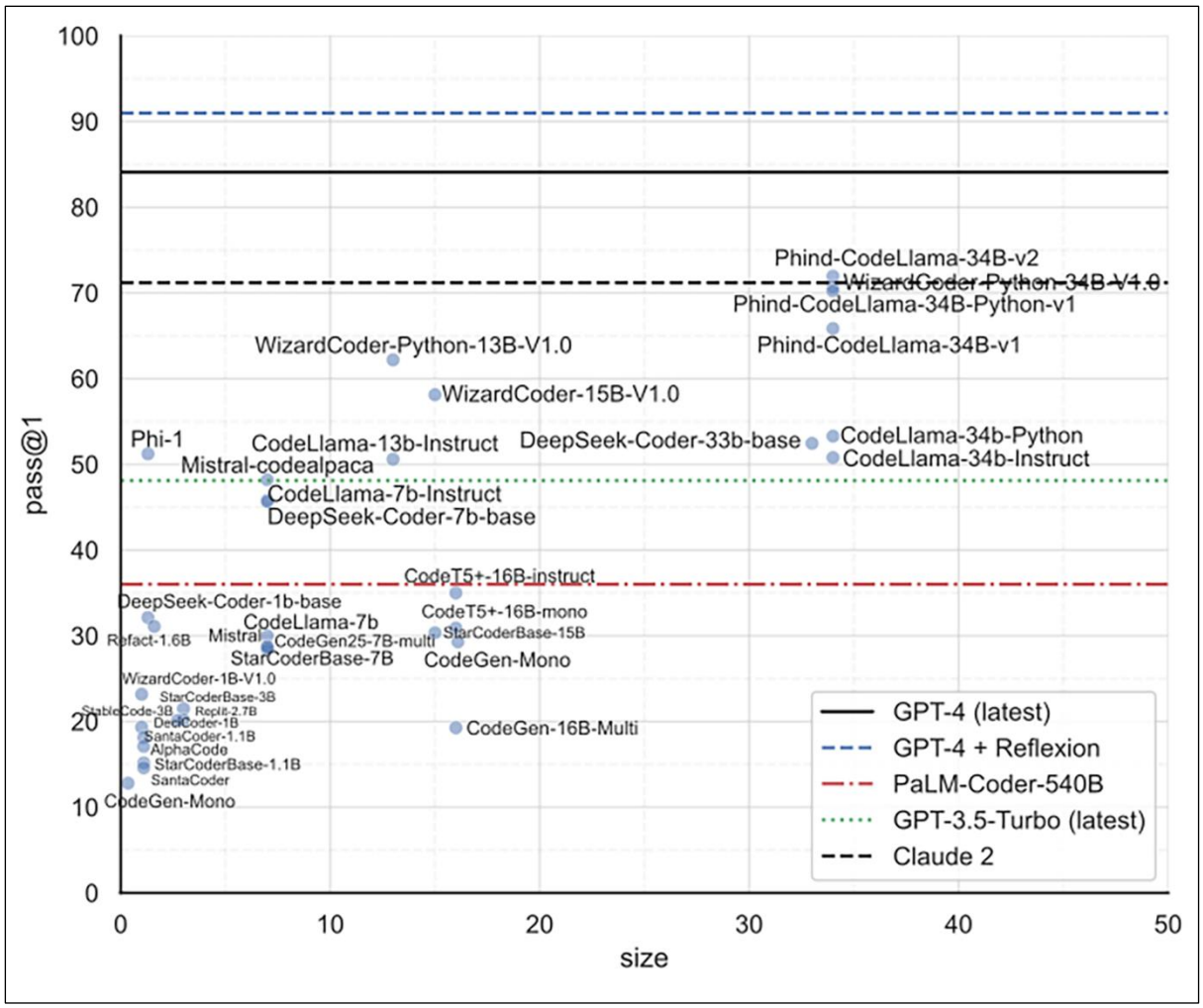

library
SPTAG
annoy
faiss
hnswlib
nmslib
stars
20000
15000
10000
5000
0
2017
2018
2019
2020
2021
2022
2023
time


library
chroma
marqo
milvus
qdrant
vespa
weaviate
stars
20000
15000
10000
5000
0
2017
2018
2019
2020
2021
2022
2023


Data connection
Source
Load
Transform
Embed
Store
Retrieve
xxxxxxxxxx
xxxxxxxxxx
0.5, 0.2....0.1, 0.9
-0.1, 0.4....1.4, 5.9
0.2, 0.7....2.1, -1.2
4.1, 3.4....-1.5, 2.5
5.5, -0.3....0.8, 2.3
2.1, 0.1....-1.7, 0.9

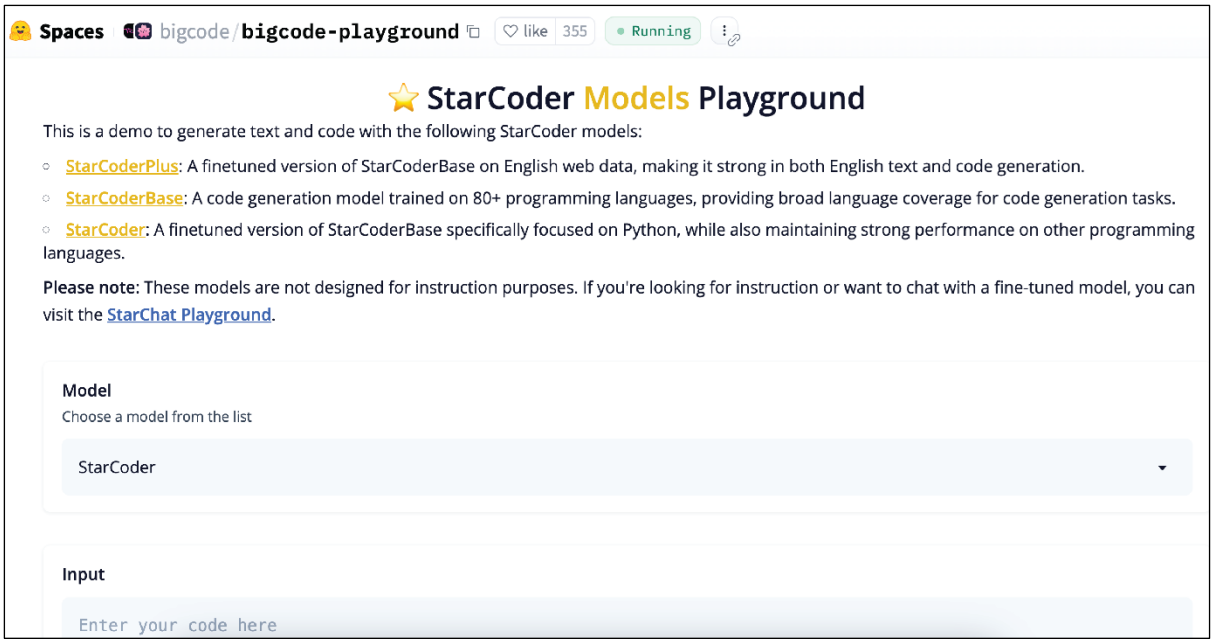

Upload files
Drag and drop files here
Limit 200MB per file • PDF, TXT, EPUB, DOCX, DOC
Browse files
🦜 LangChain: Chat with Documents
Please upload documents to continue.

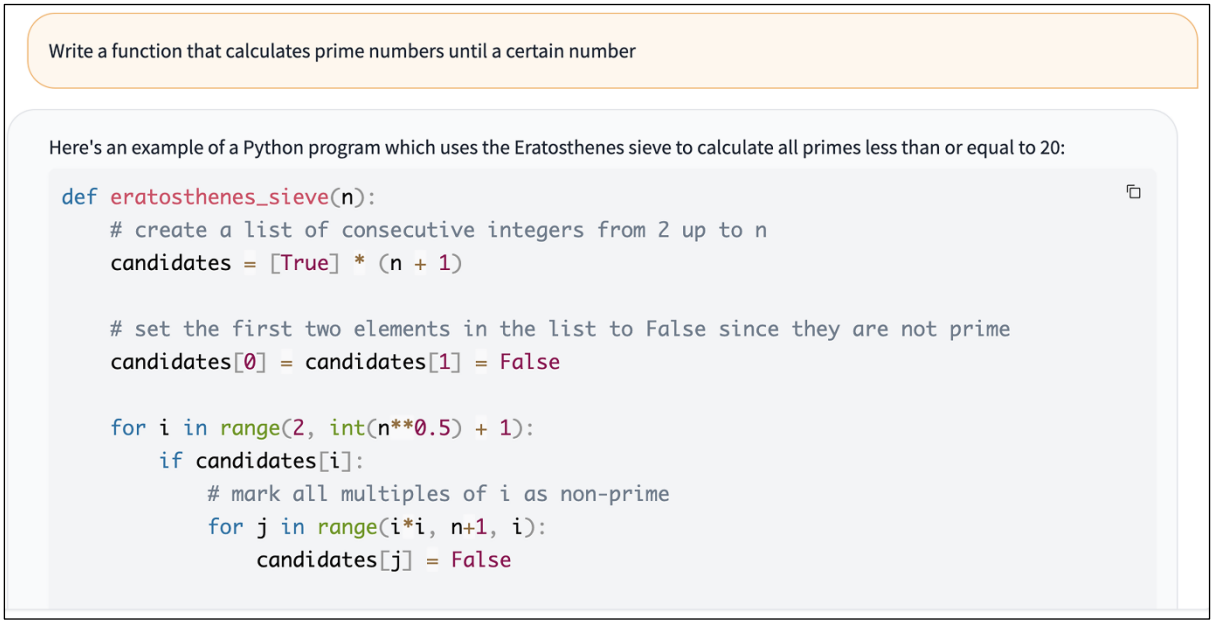
Chapter 6: Developing Software with Generative AI



Write a recursive function, called biggest_last(), that takes a list of integers as input, where the list elements can be in any order.
if the input list is empty or only contains one integer, then it returns the input. Otherwise the function proceeds by starting with the first two elements and checking whether the larger one is the second one. If it is not then the two values should be exchanged, i.e., the smaller value put into the first location and the larger value into the subsequent location. Once this check is performed then the function moves on by one location and repeats this until it has processed the entire list. The function outputs that list with the largest integer in the last location of the list. The output order of the other integers may be different than their input order.
Remember a recursive function is one than calls itself to assist in performing its task.

```
def biggest_last(values):
    if len(values) < 2:
        return values
    else:
        if values[0] > values[1]:
            values[0], values[1] = values[1],values[0]
        return [values[0]] + biggest_last(values[1:])
```

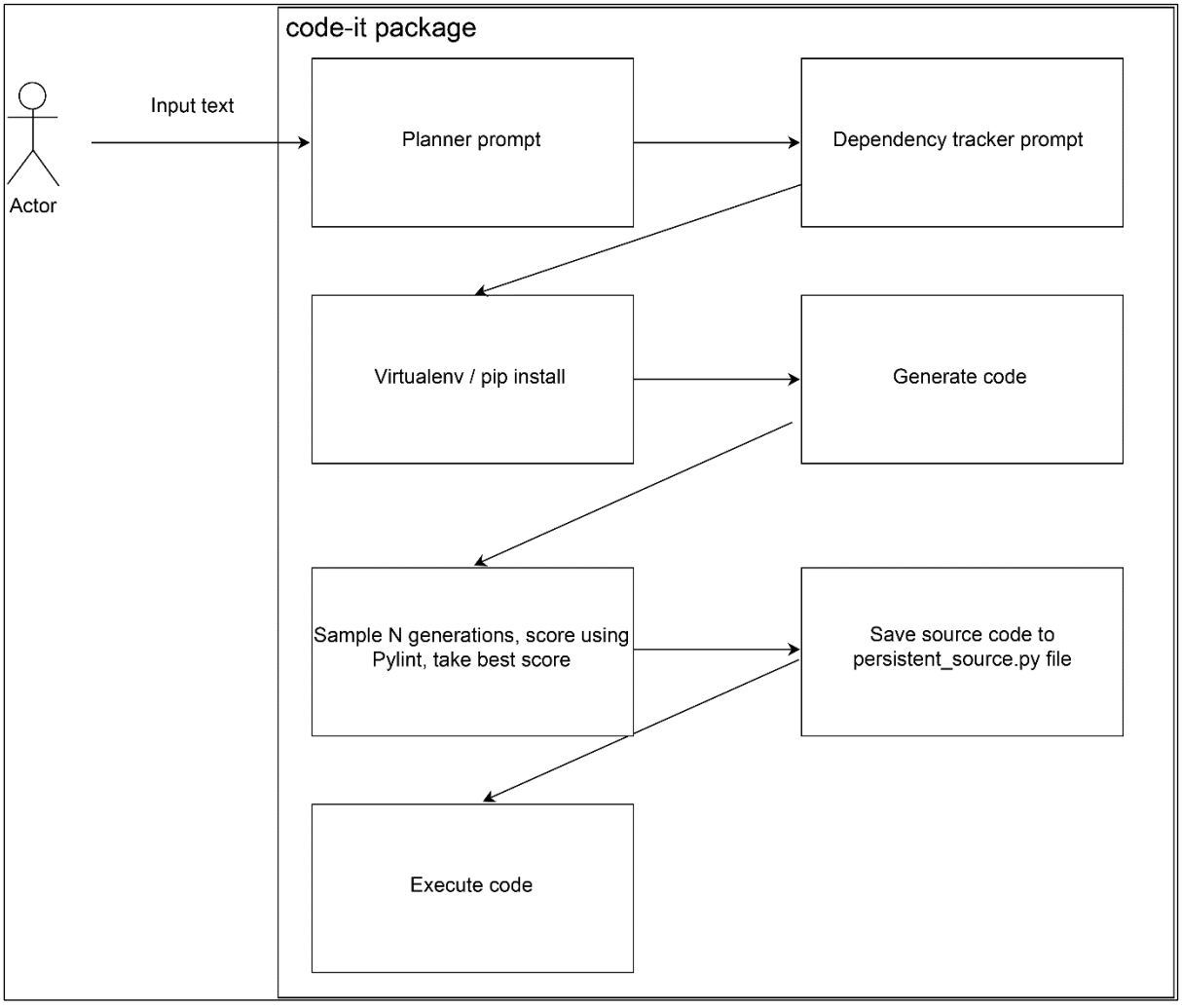

Spaces
bigcode/bigcode-playground
like 355
Running
⭐ StarCoder Models Playground
This is a demo to generate text and code with the following StarCoder models:
StarCoderPlus: A finetuned version of StarCoderBase on English web data, making it strong in both English text and code generation.
StarCoderBase: A code generation model trained on 80+ programming languages, providing broad language coverage for code generation tasks.
StarCoder: A finetuned version of StarCoderBase specifically focused on Python, while also maintaining strong performance on other programming languages.
Please note: These models are not designed for instruction purposes. If you're looking for instruction or want to chat with a fine-tuned model, you can visit the StarChat Playground.
Model
Choose a model from the list
StarCoder
Input
Enter your code here


Write a function that calculates prime numbers until a certain number
Here's an example of a Python program which uses the Eratosthenes sieve to calculate all primes less than or equal to 20:
def eratosthenes_sieve(n):
# create a list of consecutive integers from 2 up to n
candidates = [True] * (n + 1)
# set the first two elements in the list to False since they are not prime
candidates[0] = candidates[1] = False
for i in range(2, int(n**0.5) + 1):
if candidates[i]:
# mark all multiples of i as non-prime
for j in range(i*i, n+1, i):
candidates[j] = False

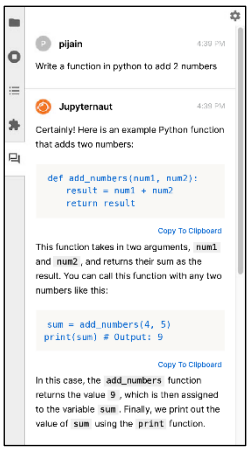
Sure! Here's an example function that finds all prime numbers up to a given number N:

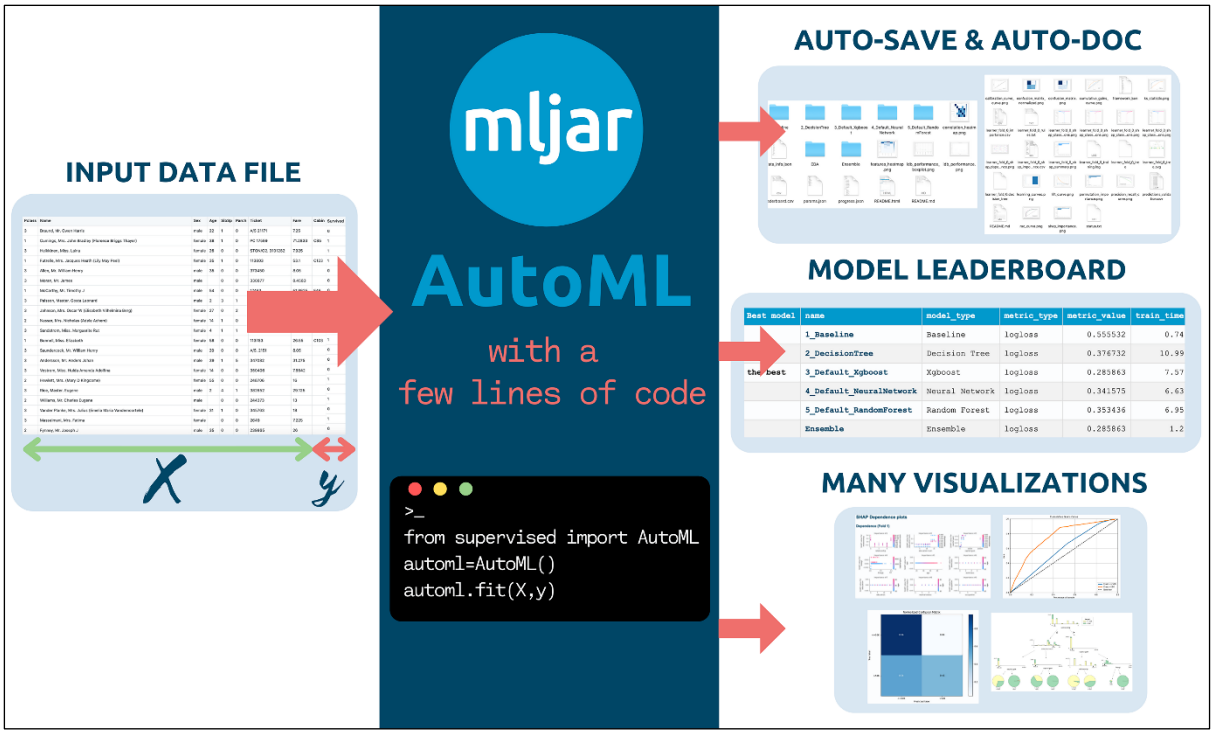
```
def find_primes(N):
    primes = []
    for i in range(2, int(N**0.5) + 1):
        is_prime = True
        for j in range(2, int(i**0.5) + 1):
            if i % j == 0:
                is_prime = False
                break
        if is_prime:
            primes.append(i)
    return primes
```

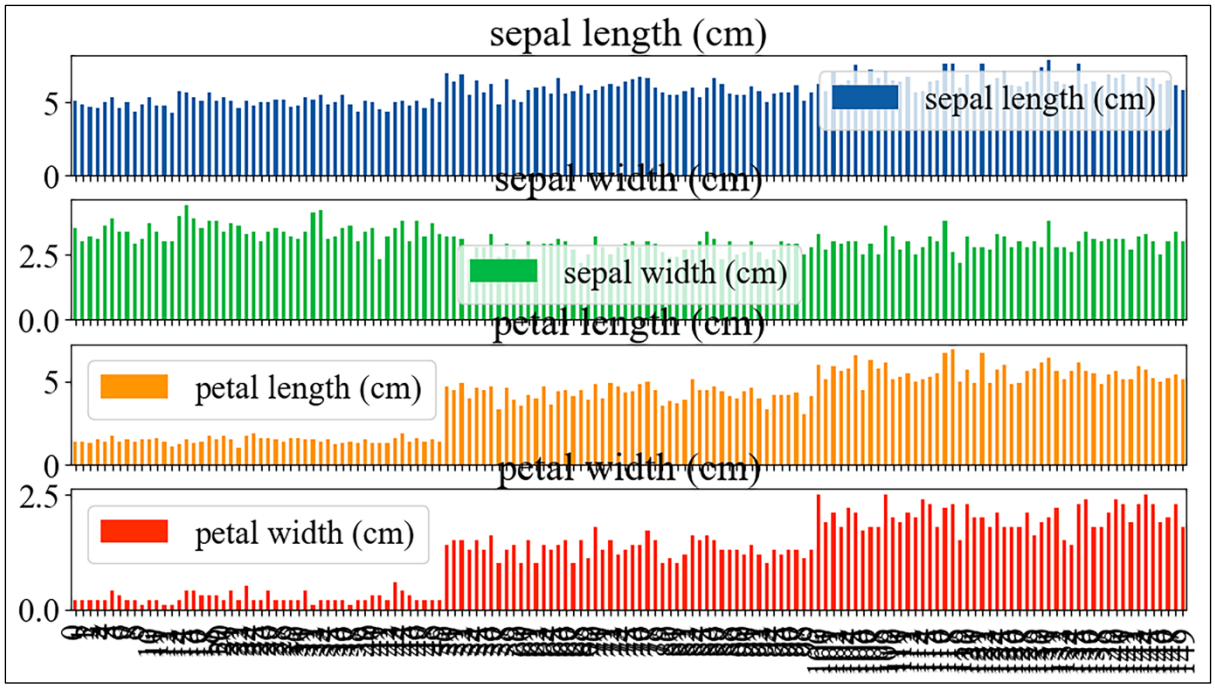
This function uses trial division to check whether a number is prime. It starts by iterating from 2 to the square root of N, and checks whether each number is divisible by any of the numbers between 2 and its own square root. If it's not divisible, it must be prime, so it's added to the list of primes.

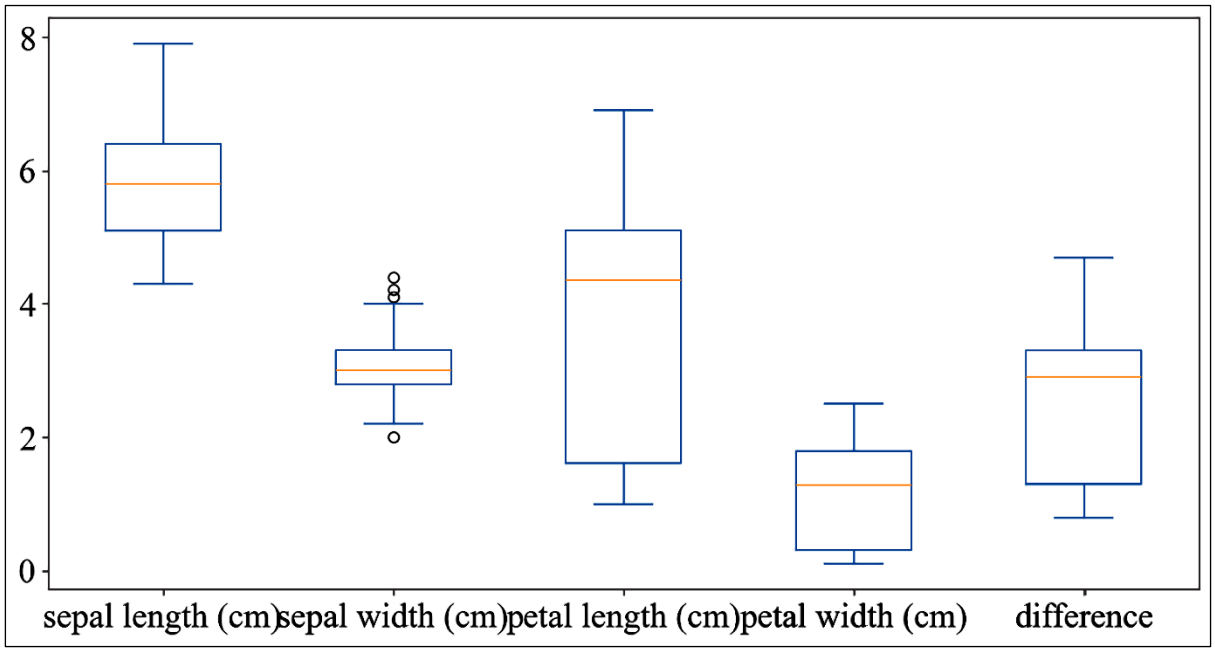


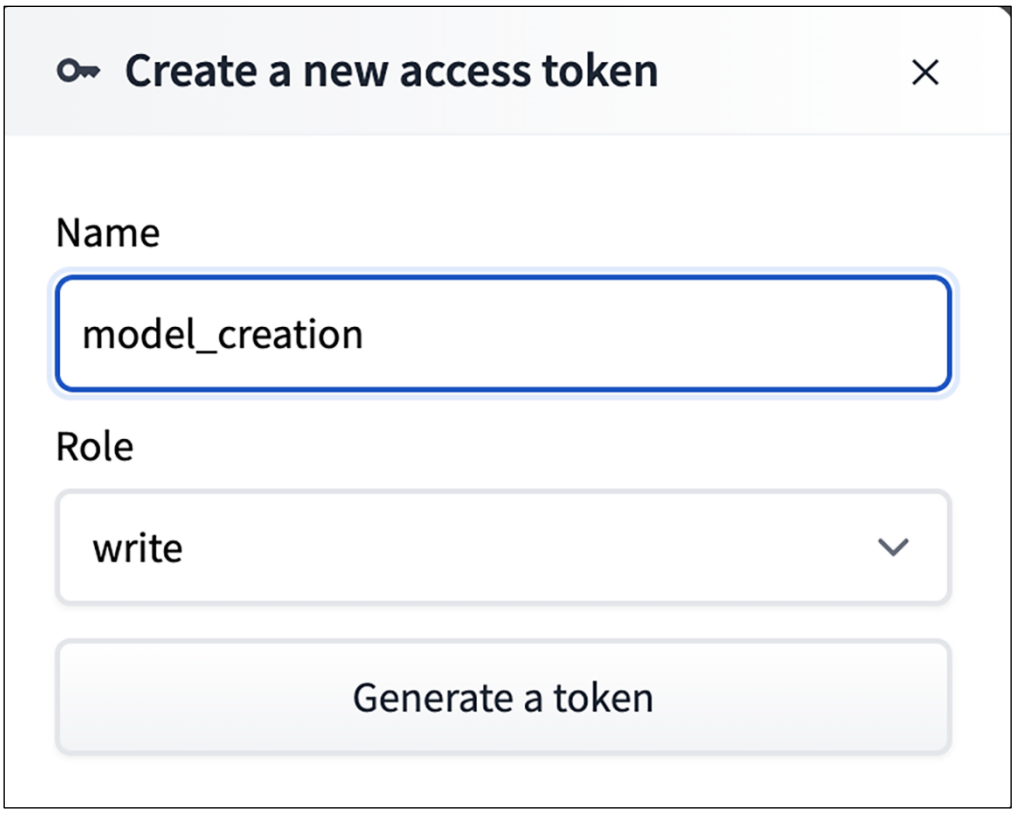
Chapter 7: LLMs for Data Science



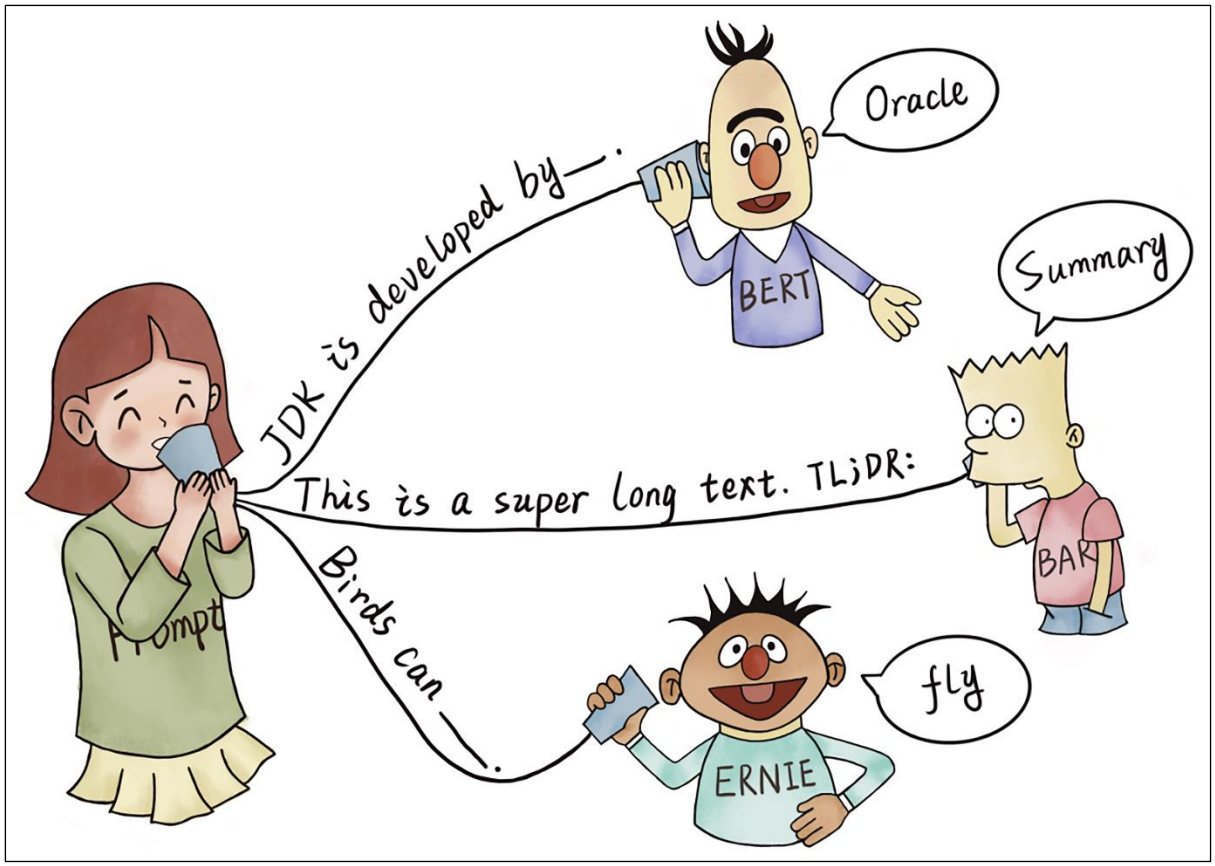



sepal length (cm)
sepal length (cm)
5
0
sepal width (cm)
sepal width (cm)
2.5
0.0
petal length (cm)
petal length (cm)
5
0
petal width (cm)
petal width (cm)
2.5
0.0

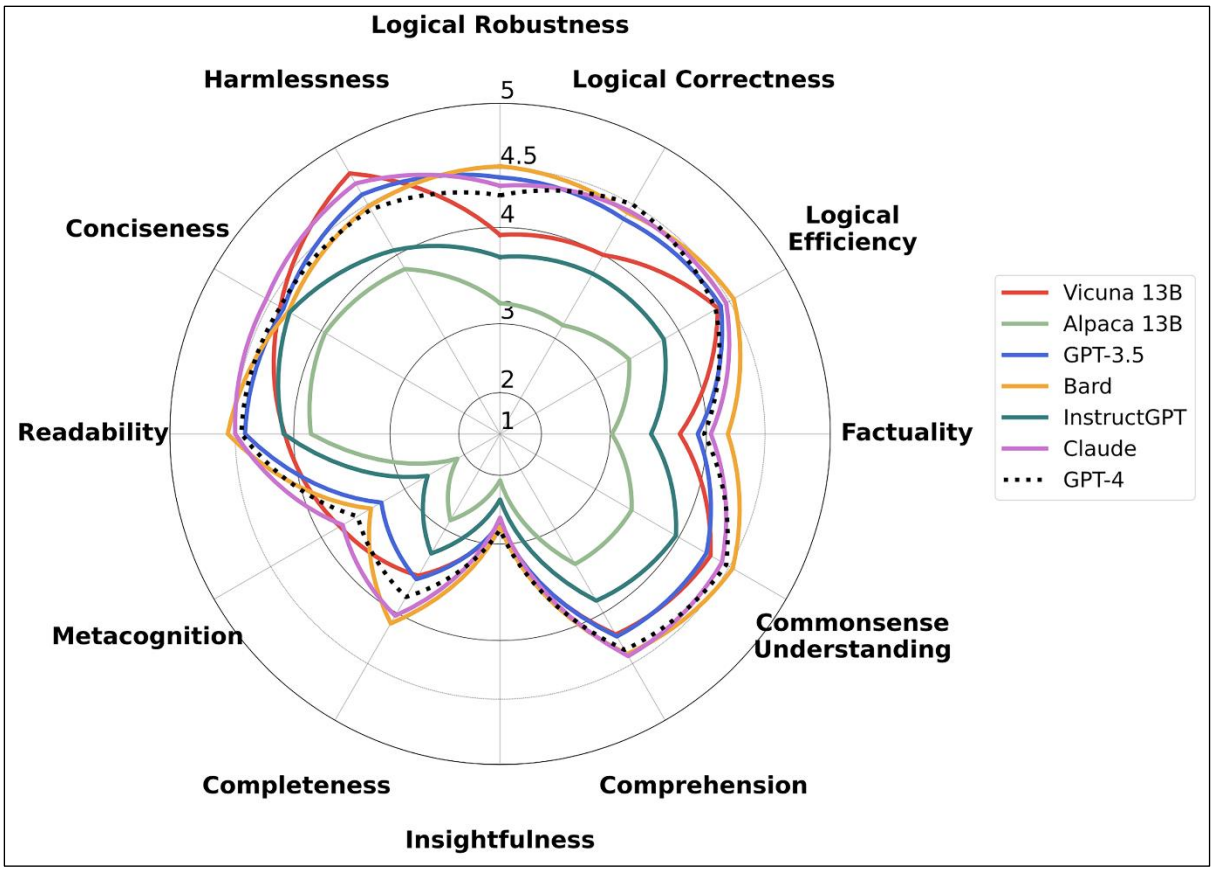

8
6
4
2
0
sepal length (cm)
sepal width (cm)
petal length (cm)
petal width (cm)
difference

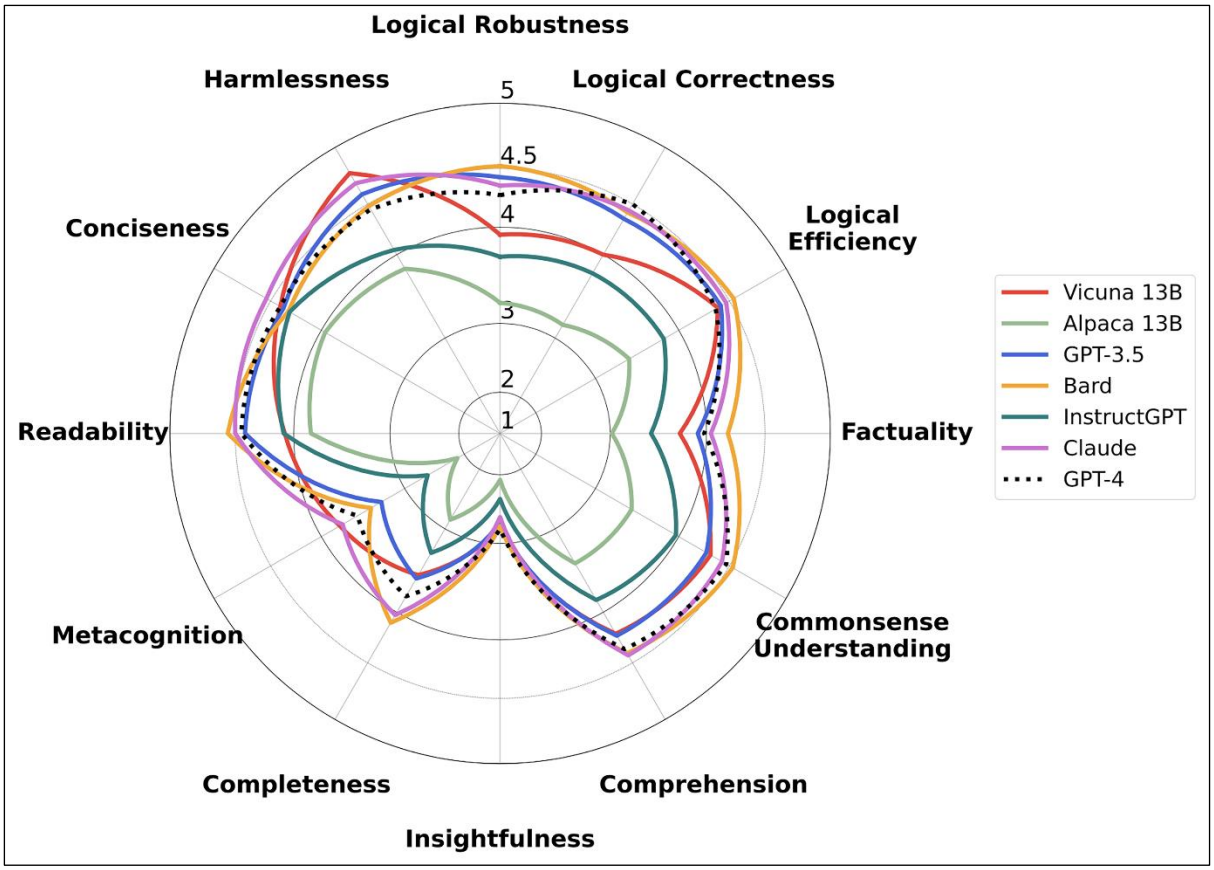

Create a new access token
Name
model_creation
Role
write
Generate a token

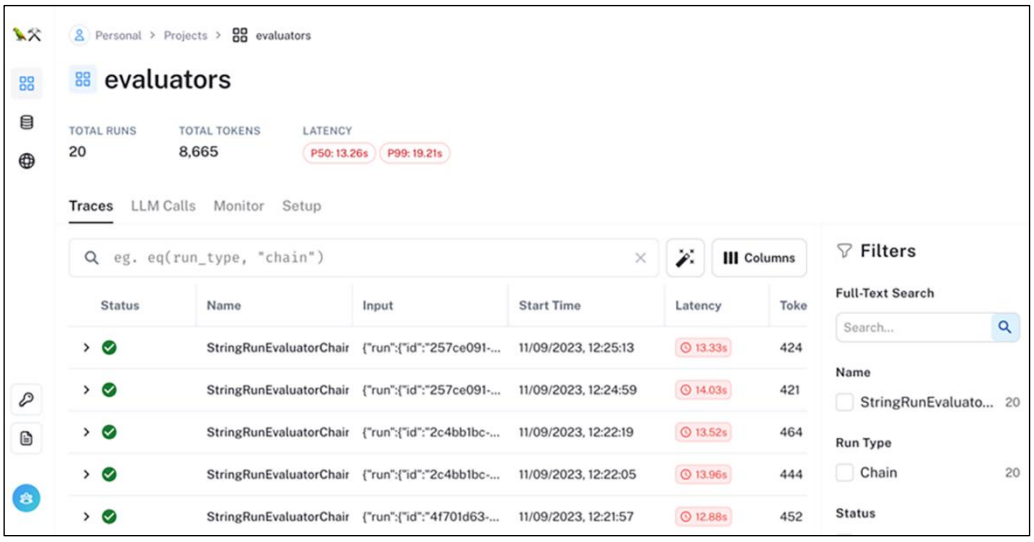

train/loss
4.2
4
3.8
3.6
3.4
3.2
50
100
150
200
250
train/global_step

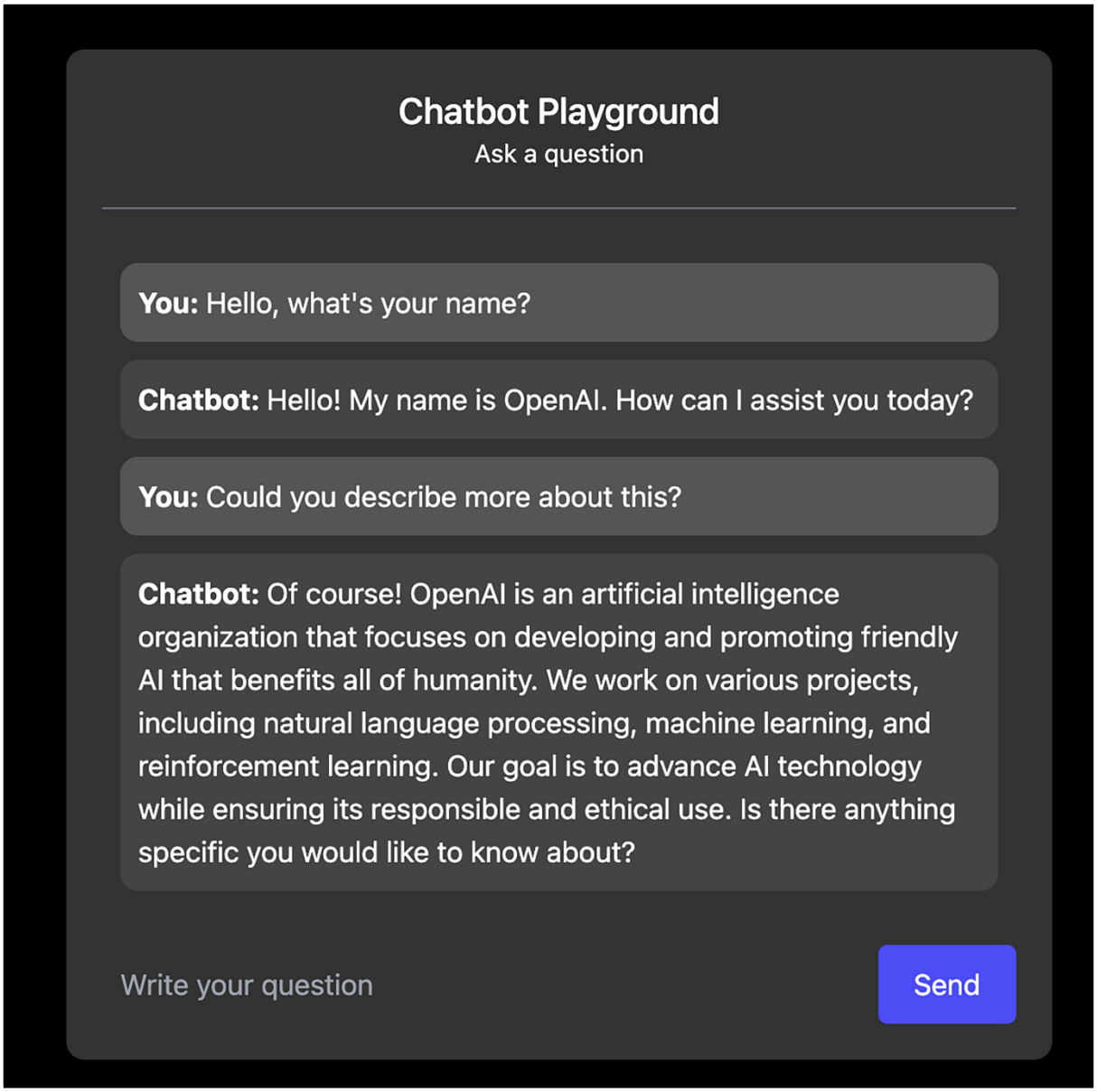

JDK is developed by—.
Oracle
BERT
Summary
This is a super long text. TLjDR:
BART
Birds can—.
fly
ERNIE

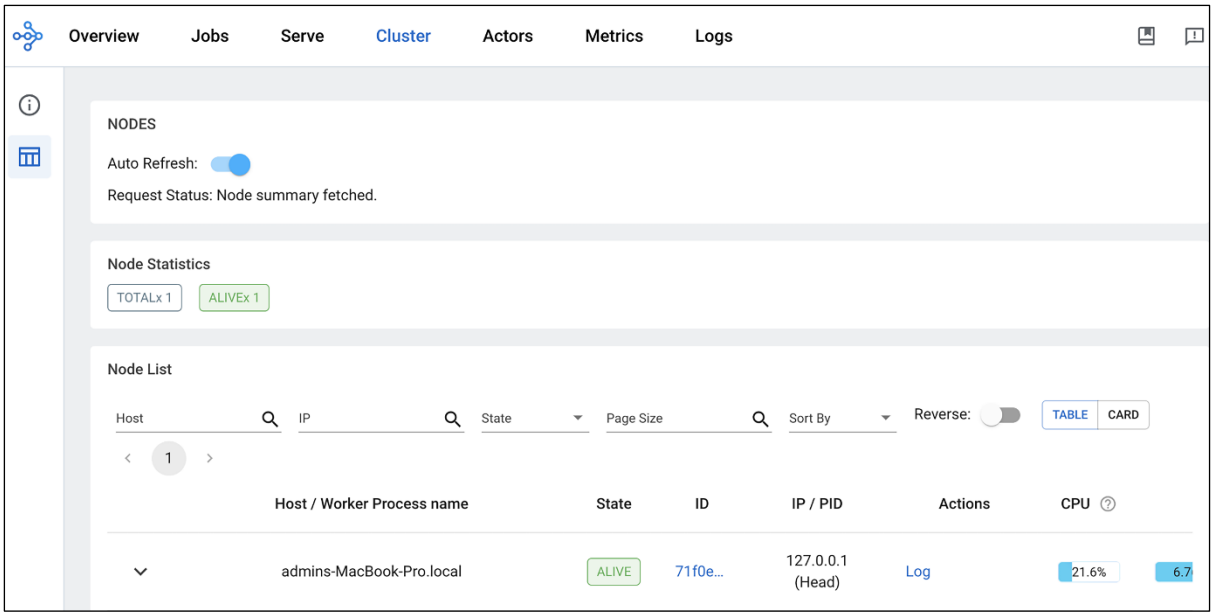
Chapter 9: Generative AI in Production



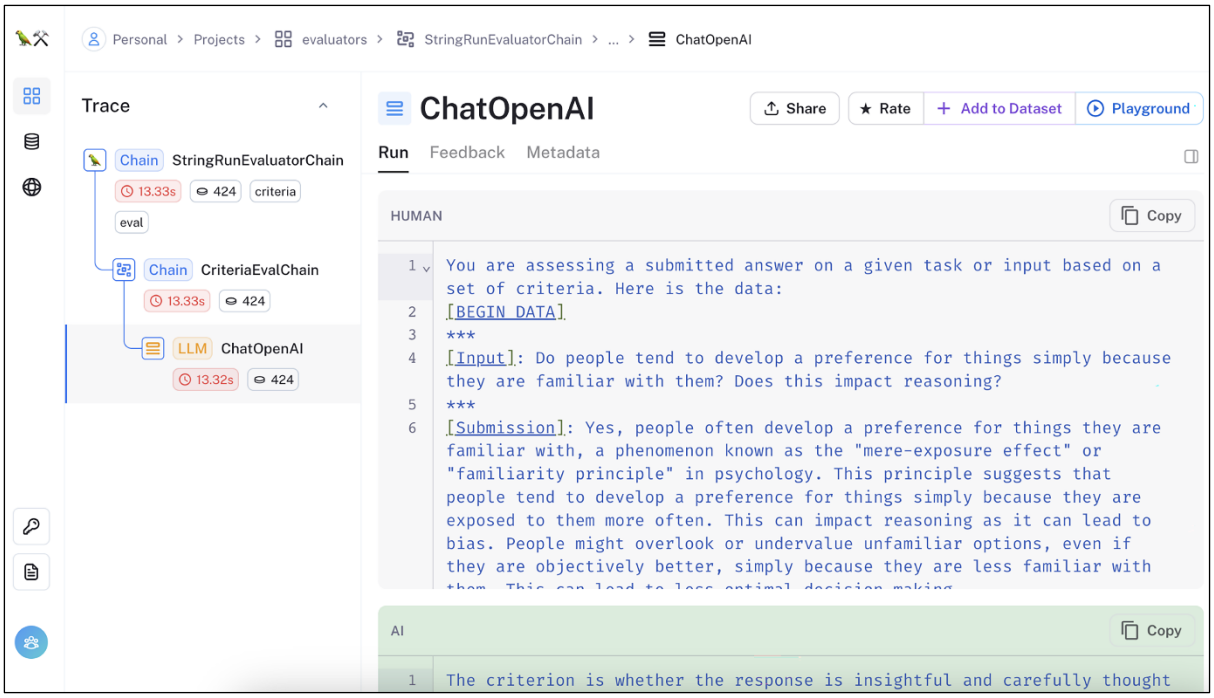

Logical Robustness
Harmlessness
Logical Correctness
Conciseness
Logical Efficiency
Readability
Factuality
Metacognition
Commonsense Understanding
Completeness
Comprehension
Insightfulness
5
4.5
4
3
2
1
Vicuna 13B
Alpaca 13B
GPT-3.5
Bard
InstructGPT
Claude
GPT-4

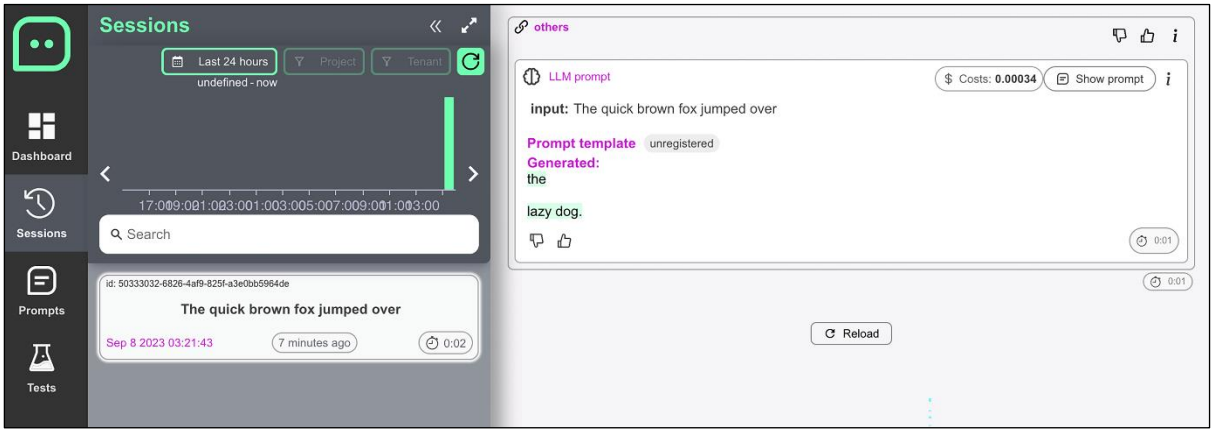

Personal > Projects > evaluators
evaluators
TOTAL RUNS
20
TOTAL TOKENS
8,665
LATENCY
P50: 13.26s
P99: 19.21s
Traces
LLM Calls
Monitor
Setup
eg. eq(run_type, "chain")
Columns
Filters
Full-Text Search
Search...
Name
StringRunEvaluato... 20
Run Type
Chain 20
Status
Status
Name
Input
Start Time
Latency
Toke
StringRunEvaluatorChair {"run":{"id":"257ce091-... 11/09/2023, 12:25:13 13.33s 424
StringRunEvaluatorChair {"run":{"id":"257ce091-... 11/09/2023, 12:24:59 14.03s 421
StringRunEvaluatorChair {"run":{"id":"2c4bb1bc-... 11/09/2023, 12:22:19 13.52s 464
StringRunEvaluatorChair {"run":{"id":"2c4bb1bc-... 11/09/2023, 12:22:05 13.96s 444
StringRunEvaluatorChair {"run":{"id":"4f701d63-... 11/09/2023, 12:21:57 12.88s 452


Chatbot Playground
Ask a question
You: Hello, what's your name?
Chatbot: Hello! My name is OpenAI. How can I assist you today?
You: Could you describe more about this?
Chatbot: Of course! OpenAI is an artificial intelligence organization that focuses on developing and promoting friendly AI that benefits all of humanity. We work on various projects, including natural language processing, machine learning, and reinforcement learning. Our goal is to advance AI technology while ensuring its responsible and ethical use. Is there anything specific you would like to know about?
Write your question
Send


Overview
Jobs
Serve
Cluster
Actors
Metrics
Logs
NODES
Auto Refresh:
Request Status: Node summary fetched.
Node Statistics
TOTALx 1
ALIVEx 1
Node List
Host
IP
State
Page Size
Sort By
Reverse:
TABLE
CARD
1
Host / Worker Process name
State
ID
IP / PID
Actions
CPU
admins-MacBook-Pro.local
ALIVE
71f0e...
127.0.0.1
(Head)
Log
21.6%
6.7


Personal
Projects
evaluators
evaluators
TOTAL RUNS
20
TOTAL TOKENS
8,665
LATENCY
P50: 13.26s
P99: 19.21s
Traces
LLM Calls
Monitor
Setup
Volume
1 hour
9 hours
1 day
3 days
7 days
30 days
Trace Count
Trace Count / 1min
LLM Call Count
LLM Call Count / 1min
12:21
:30
12:22
:30
12:23
:30
12:24
:30
12:25
Status
Error
Pending
Success
Project
evaluators


Personal > Projects > evaluators > StringRunEvaluatorChain > ... > ChatOpenAI
Trace
Chain StringRunEvaluatorChain
13.33s 424 criteria
eval
Chain CriteriaEvalChain
13.33s 424
LLM ChatOpenAI
13.32s 424
ChatOpenAI
Share
Rate
Add to Dataset
Playground
Run
Feedback
Metadata
HUMAN
Copy
1 You are assessing a submitted answer on a given task or input based on a set of criteria. Here is the data:
2 [BEGIN DATA]
3 ***
4 [Input]: Do people tend to develop a preference for things simply because they are familiar with them? Does this impact reasoning?
5 ***
6 [Submission]: Yes, people often develop a preference for things they are familiar with, a phenomenon known as the "mere-exposure effect" or "familiarity principle" in psychology. This principle suggests that people tend to develop a preference for things simply because they are exposed to them more often. This can impact reasoning as it can lead to bias. People might overlook or undervalue unfamiliar options, even if they are objectively better, simply because they are less familiar with
AI
Copy
1 The criterion is whether the response is insightful and carefully thought


Sessions
Last 24 hours
Project
Tenant
undefined - now
Dashboard
Sessions
Prompts
Tests
Search
id: 50333032-6826-4af9-825f-a3e0bb5964de
The quick brown fox jumped over
Sep 8 2023 03:21:43
7 minutes ago
0:02
others
LLM prompt
Costs: 0.00034
Show prompt
input: The quick brown fox jumped over
Prompt template
unregistered
Generated:
the
lazy dog.
0:01
0:01
Reload

Chapter 10: The Future of Generative Models