INFO7375 - Prompt Engineering & AI

Summer 2024

Local LLM Deployment and Interaction

By – Kaushikee Bhawsar (002704590)

Master of Science in Information Systems

College of Engineering

Northeastern University, Toronto

Summer 2024

Table of Contents

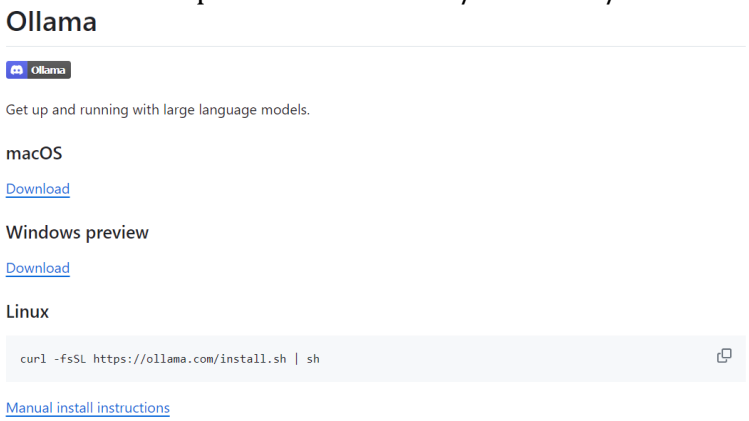
Ollama

Ollama is a language model platform that offers customizable and efficient generative AI models for various applications.

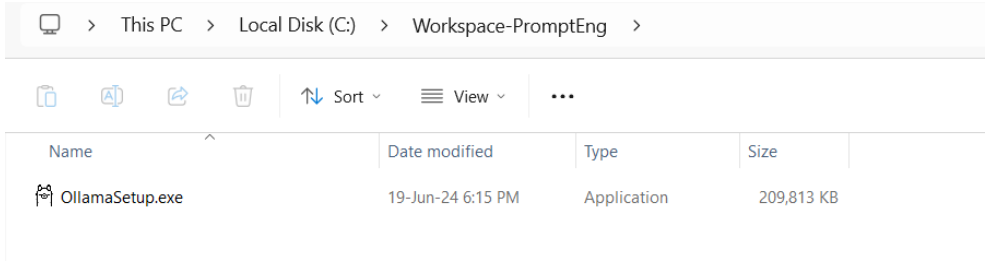
Setting up Ollama

Download and Run Ollama as per steps given below:

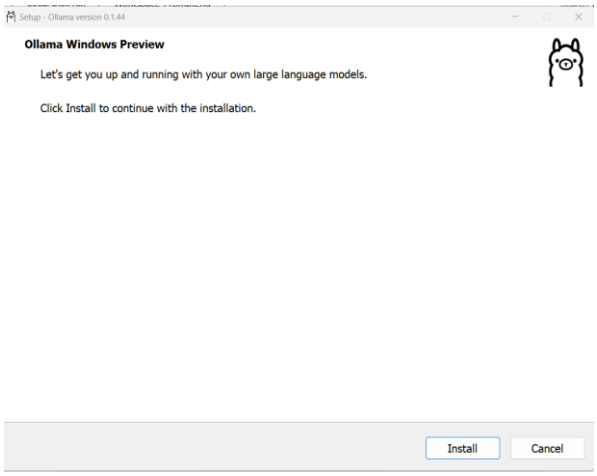
1. Go to Ollama github page - https://github.com/Ollama/llama3
2. Download Ollama for the respective OS – macOS/Windows/Linux



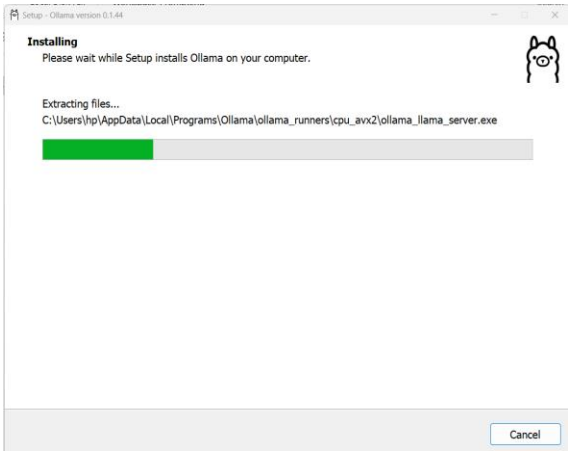
3. I am downloading for Windows
4. Run the downloaded OllamaSetup.exe file

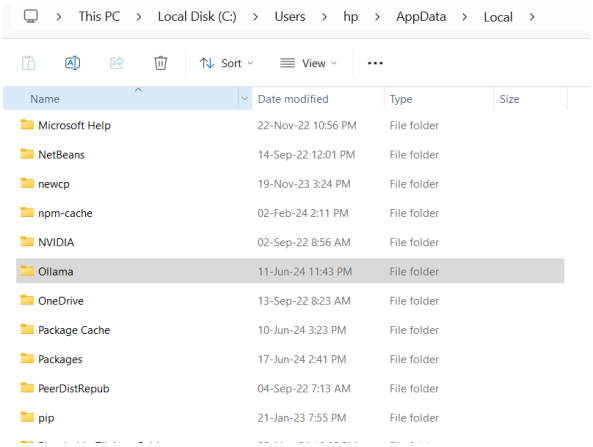


5. Install Ollama by clicking "install"



6. It will Install Ollama in your Programs folder in C:\Users

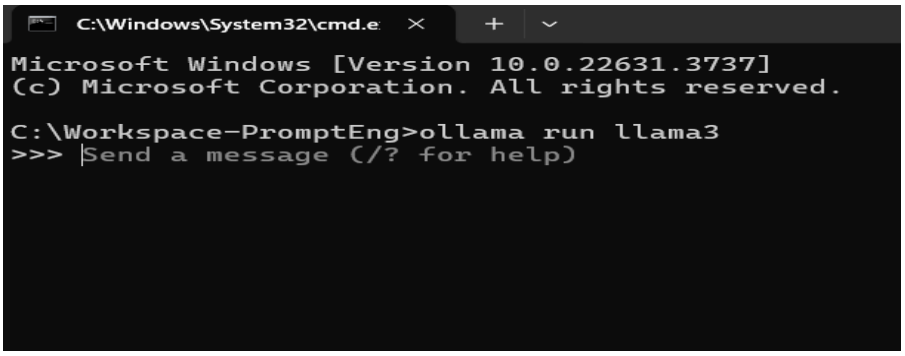




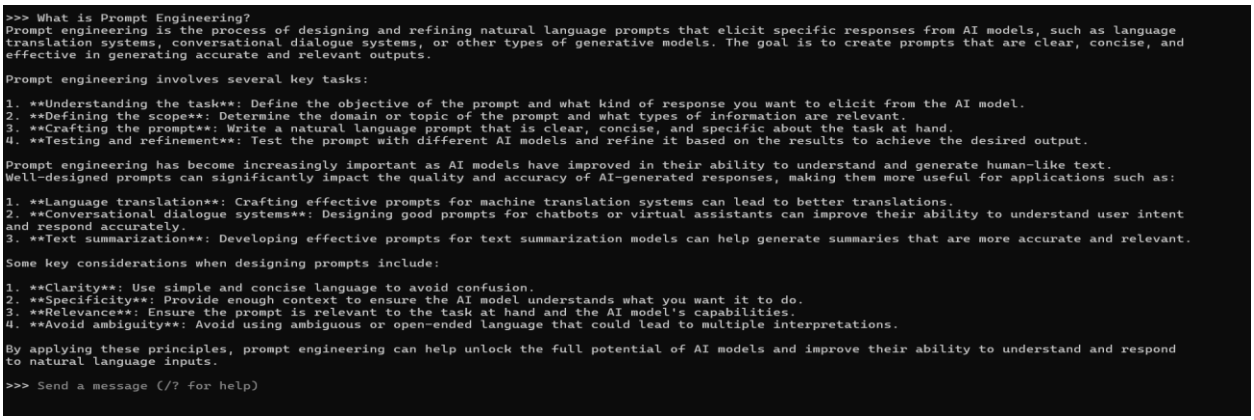
➢ Model 1: llama3 using Ollama

Llama3: Interaction with the Model

1. To chat with Llama 3, run below command:
   Run: **ollama run llama3**



2. Now, Ask a question
   Question: What is prompt engineering?



```
>>> What is Prompt Engineering?
Prompt engineering is the process of designing and refining natural language prompts that elicit specific responses from AI models, such as language
translation systems, conversational dialogue systems, or other types of generative models. The goal is to create prompts that are clear, concise, and
effective in generating accurate and relevant outputs.

Prompt engineering involves several key tasks:

1. **Understanding the task**: Define the objective of the prompt and what kind of response you want to elicit from the AI model.
2. **Defining the scope**: Determine the domain or topic of the prompt and what types of information are relevant.
3. **Crafting the prompt**: Write a natural language prompt that is clear, concise, and specific about the task at hand.
4. **Testing and refinement**: Test the prompt with different AI models and refine it based on the results to achieve the desired output.

Prompt engineering has become increasingly important as AI models have improved in their ability to understand and generate human-like text.
Well-designed prompts can significantly impact the quality and accuracy of AI-generated responses, making them more useful for applications such as:

1. **Language translation**: Crafting effective prompts for machine translation systems can lead to better translations.
2. **Conversational dialogue systems**: Designing good prompts for chatbots or virtual assistants can improve their ability to understand user intent
and respond accurately.
3. **Text summarization**: Developing effective prompts for text summarization models can help generate summaries that are more accurate and relevant.

Some key considerations when designing prompts include:

1. **Clarity**: Use simple and concise language to avoid confusion.
2. **Specificity**: Provide enough context to ensure the AI model understands what you want it to do.
3. **Relevance**: Ensure the prompt is relevant to the task at hand and the AI model's capabilities.
4. **Avoid ambiguity**: Avoid using ambiguous or open-ended language that could lead to multiple interpretations.

By applying these principles, prompt engineering can help unlock the full potential of AI models and improve their ability to understand and respond
to natural language inputs.

>>> Send a message (/? for help)
```

Llama3: Interaction using curl command

1. Open a command prompt and Run:

ollama serve

```
C:\Windows\System32\cmd.e

Microsoft Windows [Version 10.0.22631.3737]
(c) Microsoft Corporation. All rights reserved.

C:\Workspace-PromptEng>ollama serve
2024/06/19 20:31:33 routes.go:1011: INFO server config env="map[OLLAMA_DEBUG:false OLLAMA_FLASH_ATTENTION:fals
e OLLAMA_HOST:http://127.0.0.1:11434 OLLAMA_KEEP_ALIVE: OLLAMA_LLM_LIBRARY: OLLAMA_MAX_LOADED_MODELS:1 OLLAMA_
MAX_QUEUE:512 OLLAMA_MAX_VRAM:0 OLLAMA_MODELS:C:\\Users\\hp\\.ollama\\models OLLAMA_NOHISTORY:false OLLAMA_NOP
RUNE:false OLLAMA_NUM_PARALLEL:1 OLLAMA_ORIGINS:[http://localhost https://localhost http://localhost:* https:/
/localhost:* http://127.0.0.1 https://127.0.0.1 http://127.0.0.1:* https://127.0.0.1:* http://0.0.0.0 https://
0.0.0.0 http://0.0.0.0:* https://0.0.0.0:* app://* file://* tauri://*] OLLAMA_RUNNERS_DIR:C:\\Users\\hp\\AppDa
ta\\Local\\Programs\\Ollama\\ollama_runners OLLAMA_TMPDIR:]"
time=2024-06-19T20:31:33.136-04:00 level=INFO source=images.go:725 msg="total blobs: 11"
time=2024-06-19T20:31:33.136-04:00 level=INFO source=images.go:732 msg="total unused blobs removed: 0"
time=2024-06-19T20:31:33.136-04:00 level=INFO source=routes.go:1057 msg="Listening on 127.0.0.1:11434 (version
 0.1.44)"
time=2024-06-19T20:31:33.136-04:00 level=INFO source=payload.go:44 msg="Dynamic LLM libraries [cpu cpu_avx cpu
_avx2 cuda_v11.3 rocm_v5.7]"
time=2024-06-19T20:31:33.625-04:00 level=INFO source=types.go:71 msg="inference compute" id=GPU-67369c47-b988-
130c-ad58-6a44ed175a68 library=cuda compute=5.0 driver=11.6 name="NVIDIA GeForce MX110" total="2.0 GiB" availa
ble="1.7 GiB"
[GIN] 2024/06/19 - 20:40:46 | 200 |            0s |       127.0.0.1 | GET      "/"
[GIN] 2024/06/19 - 20:47:57 | 200 |       847.6µs |       127.0.0.1 | GET      "/"
```

2. To check if ollama is running or not, open windows power shell/command prompt and run:

curl http://127.0.0.1:11434/

```
Command Prompt

C:\Users\hp>curl http://127.0.0.1:11434/
Ollama is running
C:\Users\hp>
```

In Window's powershell it generate detailed response given below

```
PS C:\Users\hp> curl http://127.0.0.1:11434/

StatusCode        : 200
StatusDescription : OK
Content           : Ollama is running
RawContent        : HTTP/1.1 200 OK
                    Content-Length: 17
                    Content-Type: text/plain; charset=utf-8
                    Date: Thu, 20 Jun 2024 00:47:57 GMT

                    Ollama is running
Forms             : {}
Headers           : {[Content-Length, 17], [Content-Type, text/plain; charset=utf-8], [Date, Thu, 20 Jun 2024 00:47:57
                    GMT]}
Images            : {}
InputFields       : {}
Links             : {}
ParsedHtml        : mshtml.HTMLDocumentClass
RawContentLength  : 17
```

3. Generate a completion (Streaming & No Streaming)

   POST /api/generate

Generate a response for a given prompt with a provided model. This is a streaming endpoint, so there will be a series of responses. The final response object will include statistics and additional data from the request.

Parameters

- `model` : (required) the model name
- `prompt` : the prompt to generate a response for
- `images` : (optional) a list of base64-encoded images (for multimodal models such as `llava`)

Advanced parameters (optional):

- `format` : the format to return a response in. Currently the only accepted value is `json`
- `options` : additional model parameters listed in the documentation for the Modelfile such as `temperature`
- `system` : system message to (overrides what is defined in the `Modelfile`)
- `template` : the prompt template to use (overrides what is defined in the `Modelfile`)
- `context` : the context parameter returned from a previous request to `/generate` , this can be used to keep a short conversational memory
- `stream` : if `false` the response will be returned as a single response object, rather than a stream of objects
- `raw` : if `true` no formatting will be applied to the prompt. You may choose to use the `raw` parameter if you are specifying a full templated prompt in your request to the API
- `keep_alive` : controls how long the model will stay loaded into memory following the request (default: `5m`)

a. Generate request (Streaming)

Request:

curl -X POST http://127.0.0.1:11434/api/generate -H "Content-Type: application/json" -d "{\"model\": \"llama3\", \"prompt\": \"Why is the sky blue?\"}"

Response:

```
C:\Users\hp>curl http://127.0.0.1:11434/
Ollama is running
C:\Users\hp>curl -X POST http://localhost:11434/api/generate -H "Content-Type: application/json" -d "{\"model\": \"llama3\", \"prompt\": \"What is the large
st planet in our solar system?\"}"
{"model":"llama3","created_at":"2024-06-20T02:33:26.5745617Z","response":"The","done":false}
{"model":"llama3","created_at":"2024-06-20T02:33:32.9568595Z","response":" largest","done":false}
{"model":"llama3","created_at":"2024-06-20T02:33:33.7729135Z","response":" planet","done":false}
{"model":"llama3","created_at":"2024-06-20T02:33:34.6351528Z","response":" in","done":false}
{"model":"llama3","created_at":"2024-06-20T02:33:35.4516446Z","response":" our","done":false}
{"model":"llama3","created_at":"2024-06-20T02:33:36.2818032Z","response":" solar","done":false}
{"model":"llama3","created_at":"2024-06-20T02:33:37.1133449Z","response":" system","done":false}
{"model":"llama3","created_at":"2024-06-20T02:33:37.9442496Z","response":" is","done":false}
{"model":"llama3","created_at":"2024-06-20T02:33:38.7914526Z","response":" Jupiter","done":false}
{"model":"llama3","created_at":"2024-06-20T02:33:39.6388918Z","response":".","done":false}
{"model":"llama3","created_at":"2024-06-20T02:33:40.4703553Z","response":" It","done":false}
{"model":"llama3","created_at":"2024-06-20T02:33:41.3021133Z","response":" is","done":false}
{"model":"llama3","created_at":"2024-06-20T02:33:42.1498289Z","response":" a","done":false}
{"model":"llama3","created_at":"2024-06-20T02:33:42.9832288Z","response":" gas","done":false}
{"model":"llama3","created_at":"2024-06-20T02:33:43.8130447Z","response":" giant","done":false}
{"model":"llama3","created_at":"2024-06-20T02:33:44.832887Z","response":",","done":false}
{"model":"llama3","created_at":"2024-06-20T02:33:45.6952178Z","response":" meaning","done":false}
{"model":"llama3","created_at":"2024-06-20T02:33:46.4955442Z","response":" it","done":false}
{"model":"llama3","created_at":"2024-06-20T02:33:47.3277787Z","response":" is","done":false}
{"model":"llama3","created_at":"2024-06-20T02:33:48.1595478Z","response":" primarily","done":false}
{"model":"llama3","created_at":"2024-06-20T02:33:48.9940926Z","response":" composed","done":false}
{"model":"llama3","created_at":"2024-06-20T02:33:49.8375696Z","response":" of","done":false}
{"model":"llama3","created_at":"2024-06-20T02:33:50.6689059Z","response":" hydrogen","done":false}
{"model":"llama3","created_at":"2024-06-20T02:33:51.5157408Z","response":" and","done":false}
{"model":"llama3","created_at":"2024-06-20T02:33:52.347212Z","response":" helium","done":false}
{"model":"llama3","created_at":"2024-06-20T02:33:53.2020423Z","response":" gases","done":false}
{"model":"llama3","created_at":"2024-06-20T02:33:54.0263658Z","response":".","done":false}
{"model":"llama3","created_at":"2024-06-20T02:33:54.8577144Z","response":" Jupiter","done":false}
{"model":"llama3","created_at":"2024-06-20T02:33:55.7048598Z","response":" has","done":false}
{"model":"llama3","created_at":"2024-06-20T02:33:56.536167Z","response":" a","done":false}
{"model":"llama3","created_at":"2024-06-20T02:33:57.3673531Z","response":" diameter","done":false}
{"model":"llama3","created_at":"2024-06-20T02:33:58.2148915Z","response":" of","done":false}
{"model":"llama3","created_at":"2024-06-20T02:33:59.0313973Z","response":" approximately","done":false}
{"model":"llama3","created_at":"2024-06-20T02:33:59.8784642Z","response":" ","done":false}
{"model":"llama3","created_at":"2024-06-20T02:34:00.7260379Z","response":"142","done":false}
{"model":"llama3","created_at":"2024-06-20T02:34:01.5574439Z","response":",","done":false}
```

```
{"model":"llama3","created_at":"2024-06-20T02:35:56.3975543Z","response":"Overall","done":false}
{"model":"llama3","created_at":"2024-06-20T02:35:57.3861851Z","response":",","done":false}
{"model":"llama3","created_at":"2024-06-20T02:35:58.4059984Z","response":" Jupiter","done":false}
{"model":"llama3","created_at":"2024-06-20T02:35:59.3939277Z","response":" is","done":false}
{"model":"llama3","created_at":"2024-06-20T02:36:00.4289193Z","response":" an","done":false}
{"model":"llama3","created_at":"2024-06-20T02:36:01.7455401Z","response":" impressive","done":false}
{"model":"llama3","created_at":"2024-06-20T02:36:02.9068476Z","response":" and","done":false}
{"model":"llama3","created_at":"2024-06-20T02:36:03.7538684Z","response":" fascinating","done":false}
{"model":"llama3","created_at":"2024-06-20T02:36:04.7422695Z","response":" planet","done":false}
{"model":"llama3","created_at":"2024-06-20T02:36:05.5899471Z","response":" that","done":false}
{"model":"llama3","created_at":"2024-06-20T02:36:06.578521Z","response":" continues","done":false}
{"model":"llama3","created_at":"2024-06-20T02:36:07.434421Z","response":" to","done":false}
{"model":"llama3","created_at":"2024-06-20T02:36:08.4769084Z","response":" capture","done":false}
{"model":"llama3","created_at":"2024-06-20T02:36:09.4381675Z","response":" the","done":false}
{"model":"llama3","created_at":"2024-06-20T02:36:10.2961622Z","response":" imagination","done":false}
{"model":"llama3","created_at":"2024-06-20T02:36:11.3003188Z","response":" of","done":false}
{"model":"llama3","created_at":"2024-06-20T02:36:12.4615075Z","response":" scientists","done":false}
{"model":"llama3","created_at":"2024-06-20T02:36:13.4650644Z","response":" and","done":false}
{"model":"llama3","created_at":"2024-06-20T02:36:14.2903283Z","response":" space","done":false}
{"model":"llama3","created_at":"2024-06-20T02:36:15.2946988Z","response":" enthusiasts","done":false}
{"model":"llama3","created_at":"2024-06-20T02:36:16.518392Z","response":" alike","done":false}
{"model":"llama3","created_at":"2024-06-20T02:36:17.8359888Z","response":"!","done":false}
{"model":"llama3","created_at":"2024-06-20T02:36:18.8407407Z","response":"","done":true,"done_reason":"stop","context":[128006,882,128007,271,3923,374,279,7
928,11841,304,1057,13238,1887,30,128009,128006,78191,128007,271,791,7928,11841,304,1057,13238,1887,374,50789,13,1102,374,264,6962,14880,11,7438,433,374,1587
1,24306,315,35784,323,97607,45612,13,50789,706,264,23899,315,13489,220,10239,11,23812,41668,320,2421,11,26563,8931,705,902,374,810,1109,220,806,3115,8294,11
09,279,23899,315,9420,382,41,20089,374,779,11191,430,433,706,264,28254,5536,389,1202,40190,13,11699,3831,71019,6958,22223,279,86228,315,14373,33975,11,2737,
9420,11,323,433,374,3463,311,617,6476,264,3560,304,46620,279,13238,1887,439,584,1440,433,3432,382,41,20089,596,1404,323,3148,1101,1304,433,264,6962,14880,11
,7438,433,1587,539,617,264,6573,7479,1093,9420,477,1023,56617,33975,13,12361,11,1202,15135,374,1903,709,315,13931,315,6962,11,14812,9501,11,323,32426,11,449
,264,93378,13766,1887,3967,439,279,8681,3816,24178,520,1202,4219,382,28589,11,50789,374,459,16358,323,27387,11841,430,9731,311,12602,279,28899,315,14248,323
,3634,43448,27083,0,128009],"total_duration":216707032700,"load_duration":26952862700,"prompt_eval_count":20,"prompt_eval_duration":17470381000,"eval_count"
:191,"eval_duration":172262476000}

C:\Users\hp>
```

Lets break it down,

A stream of JSON objects is returned:

{
 "model": "llama3",
 "created_at": "2024-06-20T00:22:14.2056465Z ",
 "response": " What ",
 "done": false
}

The final response in the stream also includes additional data about the generation:

- `total_duration` : time spent generating the response
- `load_duration` : time spent in nanoseconds loading the model
- `prompt_eval_count` : number of tokens in the prompt
- `prompt_eval_duration` : time spent in nanoseconds evaluating the prompt
- `eval_count` : number of tokens in the response
- `eval_duration` : time in nanoseconds spent generating the response
- `context` : an encoding of the conversation used in this response, this can be sent in the next request to keep a conversational memory
- `response` : empty if the response was streamed, if not streamed, this will contain the full response


```
{
 "model": "llama3",
 "created_at": "2024-06-20T00:22:14.2056465Z ",
 "response": " What ",
 "done": true,
 "context": [1, 2, 3],
 "total_duration": 10706818083,
 "load_duration": 6338219291,
 "prompt_eval_count": 26,
 "prompt_eval_duration": 130079000,
 "eval_count": 259,
 "eval_duration": 4232710000
}
```

```
{"model":"llama3","created_at":"2024-06-20T02:36:18.8407407Z","response":"","done":true,"done_reason":"stop","context":[128006,882,128007,271,3923,374,279,7
928,11841,304,1057,13238,1887,30,128009,128006,78191,128007,271,791,7928,11841,304,1057,13238,1887,374,50789,13,1102,374,264,6962,14880,11,7438,433,374,1587
1,24306,315,35784,323,97607,45612,13,50789,706,264,23899,315,13489,220,10239,11,23812,41668,320,2421,11,26563,8931,705,902,374,810,1109,220,806,3115,8294,11
09,279,23899,315,9420,382,41,20089,374,779,11191,430,433,706,264,28254,5536,389,1202,40190,13,11699,3831,71019,6958,22223,279,86228,315,14373,33975,11,2737,
9420,11,323,433,374,3463,311,617,6476,264,3560,304,46620,279,13238,1887,439,584,1440,433,3432,382,41,20089,596,1404,323,3148,1101,1304,433,264,6962,14880,11
,7438,433,1587,539,617,264,6573,7479,1093,9420,477,1023,56617,33975,13,12361,11,1202,15135,374,1903,709,315,13931,315,6962,11,14812,9501,11,323,32426,11,449
,264,93378,13766,1887,3967,439,279,8681,3816,24178,520,1202,4219,382,28589,11,50789,374,459,16358,323,27387,11841,430,9731,311,12602,279,28899,315,14248,323
,3634,43448,27083,0,128009],"total_duration":216707032700,"load_duration":26952862700,"prompt_eval_count":20,"prompt_eval_duration":17470381000,"eval_count"
:191,"eval_duration":172262476000}
```


b. <u>Generate request (No Streaming)</u>

Request:

curl -X POST http://localhost:11434/api/generate -H "Content-Type: application/json" -d "{\"model\": \"llama3\", \"prompt\": \"What is the largest planet in our solar system?\", \"stream\": false}"

Response:


```
C:\Users\hp>curl -X POST http://localhost:11434/api/generate -H "Content-Type: application/json" -d "{\"model\": \"llama3\", \"prompt\": \"What is the large
st planet in our solar system?\", \"stream\": false}"
{"model":"llama3","created_at":"2024-06-20T02:54:03.3039913Z","response":"The largest planet in our solar system is Jupiter. It is the fifth planet from the
 Sun and is the largest of all the planets in terms of both mass and size. It has a diameter of approximately 142,984 kilometers (88,846 miles), which is mo
re than 11 times larger than the Earth's diameter.\n\nJupiter is a gas giant, meaning it is primarily composed of hydrogen and helium gases. It has a massiv
e atmosphere with storm systems that can be thousands of kilometers wide, including the famous Great Red Spot, which has been raging for centuries.\n\nJupit
er's size and mass are truly impressive:\n\n* Diameter: 142,984 km (88,846 miles)\n* Mass: approximately 318 times that of Earth\n* Volume: over 1,300 times
 larger than Earth's volume\n\nJupiter is so massive that it is often referred to as a \"failed star,\" meaning that it would have become a star if it had b
een slightly more massive. Instead, it became the largest planet in our solar system!","done":true,"done_reason":"stop","context":[128006,882,128007,271,392
3,374,279,7928,11841,304,1057,13238,1887,30,128009,128006,78191,128007,271,791,7928,11841,304,1057,13238,1887,374,50789,13,1102,374,279,18172,11841,505,279,
8219,323,374,279,7928,315,682,279,33975,304,3878,315,2225,3148,323,1404,13,1102,706,264,23899,315,13489,220,10239,11,23812,41668,320,2421,11,26563,8931,705,
902,374,810,1109,220,806,3115,8294,1109,279,9420,596,23899,382,41,20089,374,264,6962,14880,11,7438,433,374,15871,24306,315,35784,323,97607,45612,13,1102,706
,264,11191,16975,449,13766,6067,430,649,387,9214,315,41668,7029,11,2737,279,11495,8681,3816,24178,11,902,706,1027,70517,369,24552,382,41,20089,596,1404,323,
3148,527,9615,16358,1473,9,76661,25,220,10239,11,23812,13437,320,2421,11,26563,8931,340,9,9346,25,13489,220,17592,3115,430,315,9420,198,9,20880,25,927,220,1
6,11,3101,3115,8294,1109,9420,596,8286,271,41,20089,374,779,11191,430,433,374,3629,14183,311,439,264,330,16479,6917,1359,7438,430,433,1053,617,3719,264,6917
,422,433,1047,1027,10284,810,11191,13,12361,11,433,6244,279,7928,11841,304,1057,13238,1887,0,128009],"total_duration":253908598800,"load_duration":277209418
00,"prompt_eval_count":20,"prompt_eval_duration":17667475000,"eval_count":209,"eval_duration":208494375000}
C:\Users\hp>
```


4. Generate a chat completion

POST /api/chat

Generate the next message in a chat with a provided model. This is a streaming endpoint, so there will be a series of responses. Streaming can be disabled using "stream": false. The final response object will include statistics and additional data from the request.

Request:

curl -X POST http://localhost:11434/api/chat -H "Content-Type: application/json" -d "{\"model\": \"llama3\", \"messages\": [{\"role\": \"user\", \"content\": \"Hello!\"}], \"stream\": false}"

Response:

```
C:\Users\hp>curl -X POST http://localhost:11434/api/chat -H "Content-Type: application/json" -d "{\"model\": \"llama3\", \"messages\": [{\"role\": \"user\",
 \"content\": \"Hello!\"}], \"stream\": false}"
{"model":"llama3","created_at":"2024-06-20T03:05:15.024105Z","message":{"role":"assistant","content":"Hello! It's nice to meet you. Is there something I can
 help you with, or would you like to chat?"},"done_reason":"stop","done":true,"total_duration":73409711300,"load_duration":27606221200,"prompt_eval_count":1
2,"prompt_eval_duration":11802459000,"eval_count":26,"eval_duration":33989141000}
C:\Users\hp>
```

5. List Local Models

   GET /api/tags

List models that are available locally.

Request:

curl http://localhost:11434/api/tags

Response:

```
C:\Users\hp>curl http://localhost:11434/api/tags
{"models":[{"name":"llama2:latest","model":"llama2:latest","modified_at":"2024-06-13T13:26:00.3541328-04:00","size":3826793677,"digest":"78e26419b4469263f75
331927a00a0284ef6544c1975b826b15abdaef17bb962","details":{"parent_model":"","format":"gguf","family":"llama","families":["llama"],"parameter_size":"7B","qua
ntization_level":"Q4_0"}},{"name":"llama3:latest","model":"llama3:latest","modified_at":"2024-06-11T15:26:55.9793058-04:00","size":4661224676,"digest":"365c
0bd3c000a25d28ddbf732fe1c6add414de7275464c4e4d1c3b5fcb5d8ad1","details":{"parent_model":"","format":"gguf","family":"llama","families":["llama"],"parameter_
size":"8.0B","quantization_level":"Q4_0"}}]}
C:\Users\hp>
```

6. Show Model Information

   POST /api/show

Show information about a model including details, modelfile, template, parameters, license, system prompt.

Request:

curl -X POST http://localhost:11434/api/show -H "Content-Type: application/json" -d "{\"name\": \"llama3\"}"

Response:

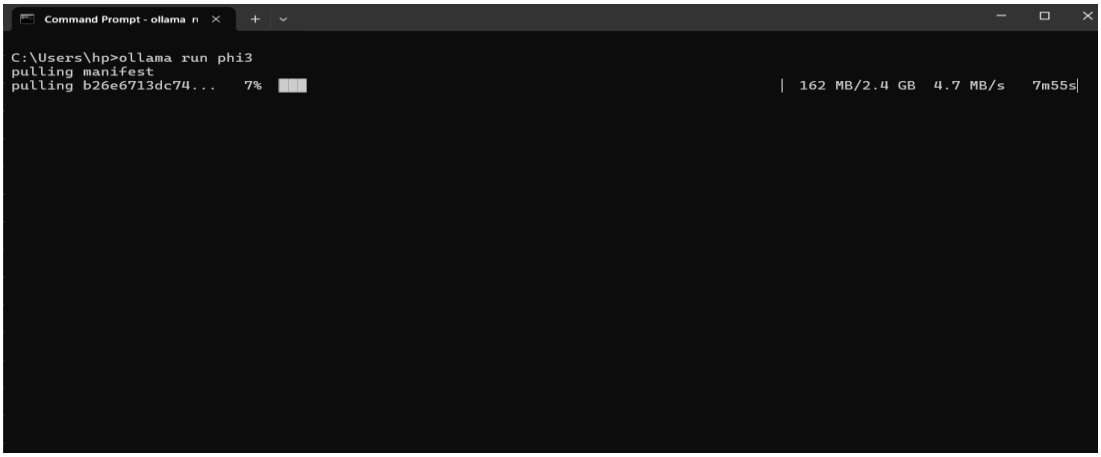
```
C:\Users\hp>curl -X POST http://localhost:11434/api/show -H "Content-Type: application/json" -d "{\"name\": \"llama3\"}"
{"license":"META LLAMA 3 COMMUNITY LICENSE AGREEMENT\n\nMeta Llama 3 Version Release Date: April 18, 2024\n“Agreement” means the terms and conditions for us
e, reproduction, distribution and modification of the Llama Materials set forth herein.\n\n“Documentation” means the specifications, manuals and documentati
on accompanying Meta Llama 3 distributed by Meta at https://llama.meta.com/get-started/.\n\n“Licensee” or “you” means you, or your employer or any other per
son or entity (if you are entering into this Agreement on such person or entity’s behalf), of the age required under applicable laws, rules or regulations t
o provide legal consent and that has legal authority to bind your employer or such other person or entity if you are entering in this Agreement on their beh
alf.\n\n“Meta Llama 3” means the foundational large language models and software and algorithms, including machine-learning model code, trained model weight
s, inference-enabling code, training-enabling code, fine-tuning enabling code and other elements of the foregoing distributed by Meta at https://llama.meta.
com/llama-downloads.\n\n“Llama Materials” means, collectively, Meta’s proprietary Meta Llama 3 and Documentation (and any portion thereof) made available un
der this Agreement.\n\n“Meta” or “we” means Meta Platforms Ireland Limited (if you are located in or, if you are an entity, your principal place of business
 is in the EEA or Switzerland) and Meta Platforms, Inc. (if you are located outside of the EEA or Switzerland). \n\nBy clicking “I Accept” below or by using
 or distributing any portion or element of the Llama Materials, you agree to be bound by this Agreement.\n\n1. License Rights and Redistribution.\n\n\ta. Gr
ant of Rights. You are granted a non-exclusive, worldwide, non-transferable and royalty-free limited license under Meta’s intellectual property or other rig
hts owned by Meta embodied in the Llama Materials to use, reproduce, distribute, copy, create derivative works of, and make modifications to the Llama Mater
ials. \n\tb. Redistribution and Use.  \n\t\ti. If you distribute or make available the Llama Materials (or any derivative works thereof), or a product or se
rvice that uses any of them, including another AI model, you shall (A) provide a copy of this Agreement with any such Llama Materials; and (B) prominently d
isplay “Built with Meta Llama 3” on a related website, user interface, blogpost, about page, or product documentation. If you use the Llama Materials to cre
ate, train, fine tune, or otherwise improve an AI model, which is distributed or made available, you shall also include “Llama 3” at the beginning of any su
ch AI model name.\n\t\tii. If you receive Llama Materials, or any derivative works thereof, from a Licensee as part of an integrated end user product, then
Section 2 of this Agreement will not apply to you. \n\t\tiii. You must retain in all copies of the Llama Materials that you distribute the following attribu
tion notice within a “Notice” text file distributed as a part of such copies: “Meta Llama 3 is licensed under the Meta Llama 3 Community License, Copyright
© Meta Platforms, Inc. All Rights Reserved.”\n\t\tiv. Your use of the Llama Materials must comply with applicable laws and regulations (including trade comp
```

```
ion of this Policy through one of the following means:\n   \t* Reporting issues with the model: https://github.com/meta-llama/llama3\n   \t* Reporting risky
 content generated by the model: developers.facebook.com/llama_output_feedback\n   \t* Reporting bugs and security concerns: facebook.com/whitehat/info\n
\t* Reporting violations of the Acceptable Use Policy or unlicensed uses of Meta Llama 3: LlamaUseReport@meta.com\n\"\n","parameters":"num_keep
          24\nstop                                \"\u003c|start_header_id|\u003e\"\nstop                                \"\u003c|end_header_id|\u003e\"\nstop
                    \"\u003c|eot_id|\u003e\"","template":"{{ if .System }}\u003c|start_header_id|\u003esystem\u003c|end_header_id|\u003e\n\n{{ .System }}\u00
3c|eot_id|\u003e{{ end }}{{ if .Prompt }}\u003c|start_header_id|\u003euser\u003c|end_header_id|\u003e\n\n{{ .Prompt }}\u003c|eot_id|\u003e{{ end }}\u003c|st
art_header_id|\u003eassistant\u003c|end_header_id|\u003e\n\n{{ .Response }}\u003c|eot_id|\u003e","details":{"parent_model":"","format":"gguf","family":"llam
a","families":["llama"],"parameter_size":"8.0B","quantization_level":"Q4_0"}}
C:\Users\hp>
```

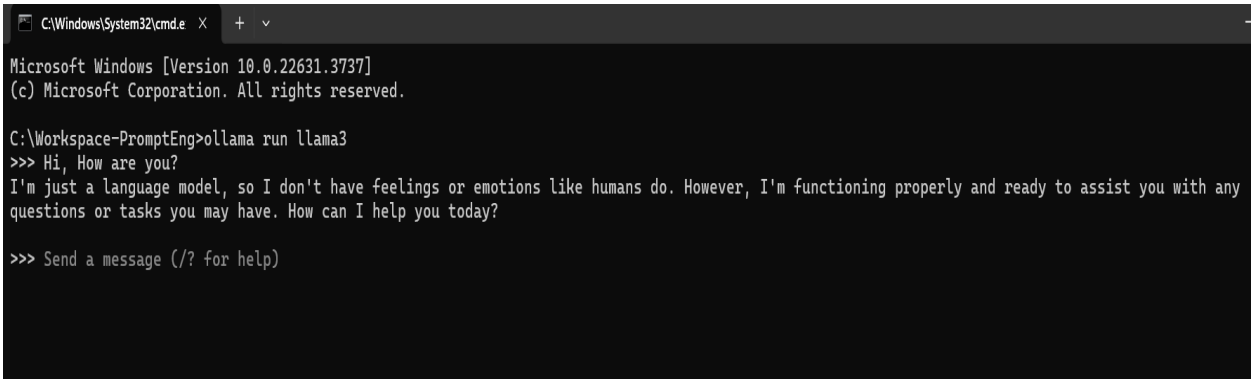
➢ Model 2: Phi 3 using Ollama

Phi3: Interaction with the Model

1. To chat with Phi 3, run below command:
   Run: ollama run phi3



2. Now, Ask a question
   Question: Hi, How are you?



Phi3: Interaction using curl command

1. Generate a completion

   a. Generate request (Streaming)

Request:

curl -X POST http://127.0.0.1:11434/api/generate -H "Content-Type: application/json" -d "{\"model\": \"phi3\", \"prompt\": \"Why is the sky blue?\"}"

Response:

```
C:\Users\hp>
C:\Users\hp>
C:\Users\hp>curl -X POST http://127.0.0.1:11434/api/generate -H "Content-Type: application/json" -d "{\"model\": \"phi3\", \"prompt\": \"Why is the sky blue
?\"}"
{"model":"phi3","created_at":"2024-06-20T05:05:30.5429704Z","response":" The","done":false}
{"model":"phi3","created_at":"2024-06-20T05:05:30.9628299Z","response":" reason","done":false}
{"model":"phi3","created_at":"2024-06-20T05:05:31.4416259Z","response":" the","done":false}
{"model":"phi3","created_at":"2024-06-20T05:05:31.9364246Z","response":" sky","done":false}
{"model":"phi3","created_at":"2024-06-20T05:05:32.4455542Z","response":" appears","done":false}
{"model":"phi3","created_at":"2024-06-20T05:05:32.9445Z","response":" blue","done":false}
{"model":"phi3","created_at":"2024-06-20T05:05:33.4415048Z","response":" to","done":false}
{"model":"phi3","created_at":"2024-06-20T05:05:33.9473006Z","response":" us","done":false}
{"model":"phi3","created_at":"2024-06-20T05:05:34.457899Z","response":" is","done":false}
{"model":"phi3","created_at":"2024-06-20T05:05:34.9588509Z","response":" due","done":false}
{"model":"phi3","created_at":"2024-06-20T05:05:35.4695575Z","response":" to","done":false}
{"model":"phi3","created_at":"2024-06-20T05:05:35.9612605Z","response":" a","done":false}
{"model":"phi3","created_at":"2024-06-20T05:05:36.462676Z","response":" phenomen","done":false}
{"model":"phi3","created_at":"2024-06-20T05:05:36.9578432Z","response":"on","done":false}
{"model":"phi3","created_at":"2024-06-20T05:05:37.4677301Z","response":" called","done":false}
{"model":"phi3","created_at":"2024-06-20T05:05:37.9628169Z","response":" Ray","done":false}
{"model":"phi3","created_at":"2024-06-20T05:05:38.4655524Z","response":"le","done":false}
{"model":"phi3","created_at":"2024-06-20T05:05:38.9687551Z","response":"igh","done":false}
{"model":"phi3","created_at":"2024-06-20T05:05:39.4719235Z","response":" scatter","done":false}
{"model":"phi3","created_at":"2024-06-20T05:05:39.9765052Z","response":"ing","done":false}
{"model":"phi3","created_at":"2024-06-20T05:05:40.4783281Z","response":".","done":false}
{"model":"phi3","created_at":"2024-06-20T05:05:40.9783782Z","response":" When","done":false}
{"model":"phi3","created_at":"2024-06-20T05:05:41.4734899Z","response":" sun","done":false}
{"model":"phi3","created_at":"2024-06-20T05:05:41.9822063Z","response":"light","done":false}
{"model":"phi3","created_at":"2024-06-20T05:05:42.4876675Z","response":" enters","done":false}
{"model":"phi3","created_at":"2024-06-20T05:05:42.991793Z","response":" Earth","done":false}
{"model":"phi3","created_at":"2024-06-20T05:05:43.4919066Z","response":"'","done":false}
{"model":"phi3","created_at":"2024-06-20T05:05:43.9853487Z","response":"s","done":false}
{"model":"phi3","created_at":"2024-06-20T05:05:44.4940385Z","response":" atmosphere","done":false}
{"model":"phi3","created_at":"2024-06-20T05:05:44.9884368Z","response":",","done":false}
{"model":"phi3","created_at":"2024-06-20T05:05:45.495434Z","response":" it","done":false}
{"model":"phi3","created_at":"2024-06-20T05:05:45.9884704Z","response":" coll","done":false}
{"model":"phi3","created_at":"2024-06-20T05:05:46.5005379Z","response":"ides","done":false}
{"model":"phi3","created_at":"2024-06-20T05:05:46.9969331Z","response":" with","done":false}
{"model":"phi3","created_at":"2024-06-20T05:05:47.4926492Z","response":" gas","done":false}
```

```
{"model":"phi3","created_at":"2024-06-20T05:08:26.2468065Z","response":"","done":true,"done_reason":"stop","context":[32010,3750,338,278,14744,7254,29973,32
007,32001,450,2769,278,14744,5692,7254,304,502,338,2861,304,263,27791,265,2000,9596,280,1141,14801,292,29889,1932,6575,4366,24395,11563,29915,29879,25005,29
892,372,5321,2247,411,10489,13206,21337,322,2319,17105,297,278,4799,29889,4525,17105,526,1568,7968,1135,278,281,6447,1477,310,7962,3578,29892,10805,20511,29
899,29893,6447,1477,11955,313,4561,7254,29897,304,14801,901,1135,5520,29899,29893,6447,1477,11955,313,14565,408,2654,467,13,13,1576,8991,953,1169,4796,3578,
29892,607,338,263,29544,310,599,278,1422,11955,2629,278,3546,26097,293,18272,29889,1932,445,3578,22170,11563,29915,29879,25005,29892,372,16254,29879,411,479
9,13206,21337,322,916,17105,297,5164,5837,29889,10924,3578,756,20511,281,6447,1477,29879,1135,916,7962,11955,322,4550,885,271,2153,901,28520,297,2999,18112,
2861,304,967,22060,411,2319,15489,8096,293,17105,29889,13,13,2887,263,1121,310,278,11664,14801,292,2779,363,7254,3578,9401,304,2654,29892,746,591,1106,701,4
72,278,14744,2645,2462,4366,6199,313,8256,278,8991,338,2038,1749,28205,511,278,29574,7254,3578,338,825,758,24130,1078,29892,4969,278,10097,310,263,7254,1474
4,29889,450,26171,310,445,27791,265,884,18568,2020,278,14744,5692,901,325,3640,368,7254,373,2821,3841,1135,372,947,373,447,1537,470,9570,29891,3841,746,727,
526,7200,17105,297,278,25005,393,14801,1422,281,6447,1477,29879,322,11105,777,310,278,7254,3578,29889,13,13,3112,29915,29879,4100,304,4443,393,278,2927,310,
278,14744,508,2615,17587,1090,5164,5855,29892,1316,408,472,6575,29878,895,470,6575,842,29892,988,591,4049,16277,263,3464,310,2654,29879,29892,282,19363,2989
2,322,3708,2701,2861,304,3620,297,15489,8096,293,14801,292,9545,4475,304,278,8991,29915,29879,2602,29889,32007],"total_duration":193161651100,"load_duration
":15062657700,"prompt_eval_count":10,"prompt_eval_duration":2380638000,"eval_count":340,"eval_duration":175703896000}
```

   b. Generate request (No Streaming)

Request:

curl -X POST http://localhost:11434/api/generate -H "Content-Type: application/json" -d "{\"model\": \"phi3\", \"prompt\": \"What is the largest planet in our solar system?\", \"stream\": false}"

Response:

```
C:\Users\hp>curl -X POST http://localhost:11434/api/generate -H "Content-Type: application/json" -d "{\"model\": \"phi3\", \"prompt\": \"What is the largest
 planet in our solar system?\", \"stream\": false}"
{"model":"phi3","created_at":"2024-06-20T05:10:26.8857304Z","response":" The largest planet in our solar system is Jupiter. It is the fifth planet from the
Sun and is classified as a gas giant. Jupiter has a diameter of about 139,820 kilometers (86,881 miles), making it more than twice as wide as Earth's moon.
It is known for its prominent Great Red Spot, which is a giant storm that has been raging on the planet for at least 400 years. Jupiter also possesses a str
ong magnetic field and over 79 moons, including four large \"Galilean\" moons: Io, Europa, Ganymede, and Callisto.","done":true,"done_reason":"stop","contex
t":[32010,1724,338,278,10150,15754,297,1749,21635,1788,29973,32007,32001,450,10150,15754,297,1749,21635,1788,338,27441,1524,29889,739,338,278,18615,15754,51
5,278,8991,322,338,770,2164,408,263,10489,28396,29889,27441,1524,756,263,24235,310,1048,29871,29896,29941,29929,29892,29947,29906,29900,20052,2699,313,29947
,29953,29892,29947,29947,29896,7800,511,3907,372,901,1135,8951,408,9377,408,11563,29915,29879,18786,29889,739,338,2998,363,967,19555,7027,4367,1706,327,2989
2,607,338,263,28396,14280,393,756,1063,1153,3460,373,278,15754,363,472,3203,29871,29946,29900,29900,2440,29889,27441,1524,884,3119,15322,263,4549,15611,1746
,322,975,29871,29955,29929,2730,787,29892,3704,3023,2919,376,29954,284,488,273,29908,2730,787,29901,22244,29892,9646,29892,28245,962,2742,29892,322,8251,513
7,29889,32007],"total_duration":74748469800,"load_duration":72110000,"prompt_eval_count":12,"prompt_eval_duration":2854041000,"eval_count":142,"eval_duratio
n":71817825000}
C:\Users\hp>
```

2. Generate a chat completion

POST /api/chat

Request:

curl -X POST http://localhost:11434/api/chat -H "Content-Type: application/json" -d "{\"model\": \"phi3\", \"messages\": [{\"role\": \"user\", \"content\": \"Hello!\"}], \"stream\": false}"

Response:

```
C:\Users\hp>
C:\Users\hp>curl -X POST http://localhost:11434/api/chat -H "Content-Type: application/json" -d "{\"model\": \"phi3\", \"messages\": [{\"role\": \"user\", \
"content\": \"Hello!\"}], \"stream\": false}"
{"model":"phi3","created_at":"2024-06-20T05:12:26.6974151Z","message":{"role":"assistant","content":" Hello! How can I help you today? Whether it's answerin
g a question, providing information on a particular topic, or assisting with general inquiries, feel free to ask. I'm here to ensure that your experience is
 as smooth and helpful as possible!"},"done_reason":"stop","done":true,"total_duration":29385000100,"load_duration":15425300,"prompt_eval_count":4,"prompt_e
val_duration":1963741000,"eval_count":56,"eval_duration":27394321000}
C:\Users\hp>
```

3. List Local Models

List models that are available locally.

GET /api/tags

Request:

curl http://localhost:11434/api/tags

Response:

```
C:\Users\hp>
C:\Users\hp>curl http://localhost:11434/api/tags
{"models":[{"name":"phi3:latest","model":"phi3:latest","modified_at":"2024-06-20T00:24:20.8932058-04:00","size":2393232963,"digest":"64c1188f2485448235b2d37
1639a127fc0e4dc2cd3c041152368883c42eb2bd1","details":{"parent_model":"","format":"gguf","family":"phi3","families":["phi3"],"parameter_size":"3.8B","quantiz
ation_level":"Q4_K_M"}},{"name":"llama2:latest","model":"llama2:latest","modified_at":"2024-06-13T13:26:00.3541328-04:00","size":3826793677,"digest":"78e264
19b4469263f75331927a00a0284ef6544c1975b826b15abdaef17bb962","details":{"parent_model":"","format":"gguf","family":"llama","families":["llama"],"parameter_si
ze":"7B","quantization_level":"Q4_0"}},{"name":"llama3:latest","model":"llama3:latest","modified_at":"2024-06-11T15:26:55.9793058-04:00","size":4661224676,"
digest":"365c0bd3c000a25d28ddbf732fe1c6add414de7275464c4e4d1c3b5fcb5d8ad1","details":{"parent_model":"","format":"gguf","family":"llama","families":["llama"
],"parameter_size":"8.0B","quantization_level":"Q4_0"}}]}
C:\Users\hp>
```

4. Show Model Information

POST /api/show

Request:

curl -X POST http://localhost:11434/api/show -H "Content-Type: application/json" -d "{\"name\": \"phi3\"}"

Response:

```
C:\Users\hp>
C:\Users\hp>curl -X POST http://localhost:11434/api/show -H "Content-Type: application/json" -d "{\"name\": \"phi3\"}"
{"license":"Microsoft.\nCopyright (c) Microsoft Corporation.\n\nMIT License\n\nPermission is hereby granted, free of charge, to any person obtaining a copy\
nof this software and associated documentation files (the \"Software\"), to deal\nin the Software without restriction, including without limitation the righ
ts\nto use, copy, modify, merge, publish, distribute, sublicense, and/or sell\ncopies of the Software, and to permit persons to whom the Software is\nfurnis
hed to do so, subject to the following conditions:\n\nThe above copyright notice and this permission notice shall be included in all\ncopies or substantial
portions of the Software.\n\nTHE SOFTWARE IS PROVIDED *AS IS*, WITHOUT WARRANTY OF ANY KIND, EXPRESS OR\nIMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIE
S OF MERCHANTABILITY,\nFITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE\nAUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DA
MAGES OR OTHER\nLIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM,\nOUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER
 DEALINGS IN THE\nSOFTWARE.","modelfile":"# Modelfile generated by \"ollama show\"\n# To build a new Modelfile based on this, replace FROM with:\n# FROM phi
3:latest\n\nFROM C:\\Users\\hp\\.ollama\\models\\blobs\\sha256-b26e6713dc749dda35872713fa19a568040f475cc71cb132cff332fe7e216462\nTEMPLATE \"{{ if .System }}
\u003c|system|\u003e\n{{ .System }}\u003c|end|\u003e\n{{ end }}{{ if .Prompt }}\u003c|user|\u003e\n{{ .Prompt }}\u003c|end|\u003e\n{{ end }}\u003c|assistant
|\u003e\n{{ .Response }}\u003c|end|\u003e\"\nPARAMETER stop \u003c|end|\u003e\nPARAMETER stop \u003c|user|\u003e\nPARAMETER stop \u003c|assistant|\u003e\nLI
CENSE \"\"\"Microsoft.\nCopyright (c) Microsoft Corporation.\n\nMIT License\n\nPermission is hereby granted, free of charge, to any person obtaining a copy\
nof this software and associated documentation files (the \"Software\"), to deal\nin the Software without restriction, including without limitation the righ
ts\nto use, copy, modify, merge, publish, distribute, sublicense, and/or sell\ncopies of the Software, and to permit persons to whom the Software is\nfurnis
hed to do so, subject to the following conditions:\n\nThe above copyright notice and this permission notice shall be included in all\ncopies or substantial
portions of the Software.\n\nTHE SOFTWARE IS PROVIDED *AS IS*, WITHOUT WARRANTY OF ANY KIND, EXPRESS OR\nIMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIE
S OF MERCHANTABILITY,\nFITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE\nAUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DA
MAGES OR OTHER\nLIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM,\nOUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER
 DEALINGS IN THE\nSOFTWARE.\"\"\"\n","parameters":"stop                                    \"\u003c|end|\u003e\"\nstop                                    \"\u003c|user|\u003e
\"\nstop                                    \"\u003c|assistant|\u003e\"","template":"{{ if .System }}\u003c|system|\u003e\n{{ .System }}\u003c|end|\u003e\n{{ end }}{
{ if .Prompt }}\u003c|user|\u003e\n{{ .Prompt }}\u003c|end|\u003e\n{{ end }}\u003c|assistant|\u003e\n{{ .Response }}\u003c|end|\u003e","details":{"parent_mo
del":"","format":"gguf","family":"phi3","families":["phi3"],"parameter_size":"3.8B","quantization_level":"Q4_K_M"}}
C:\Users\hp>
```

GPT4ALL

GPT4All is an open-source project that provides accessible, versatile, and privacy-friendly large language models (LLMs) for various applications. It aims to democratize the use of LLMs by offering models that can be run on local devices without requiring extensive cloud infrastructure.

Setting up GPT4ALL

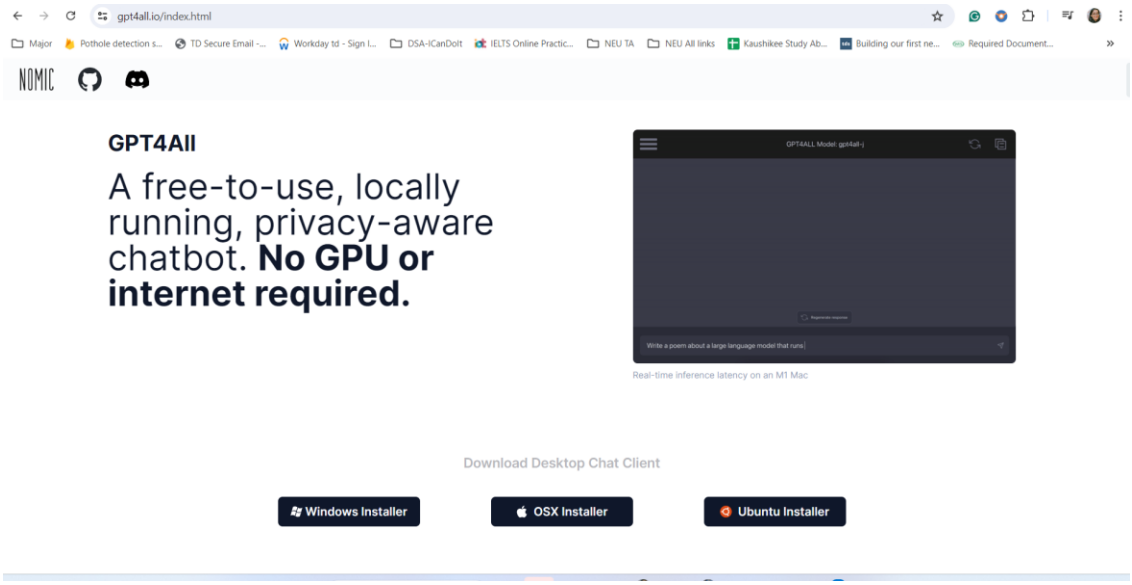
Download the GPT4All Chat Client or GPT4All Python Client based on preference:

1. GPT4All Chat Client: Provides a GUI-based interface.
2. GPT4All Python Client: Allows you to interact via Python scripts and run a local server.

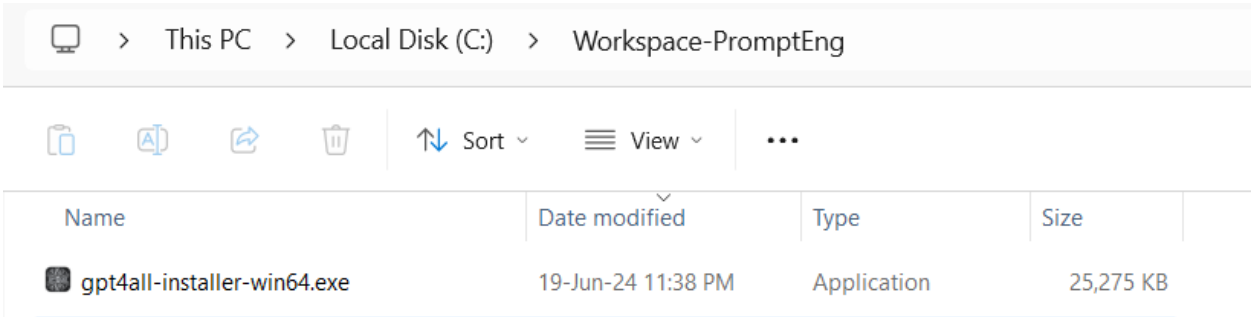
➢ Model 3: Phi3 Mini using GPT4ALL Chat Client

Install GPT4All Chat Client:

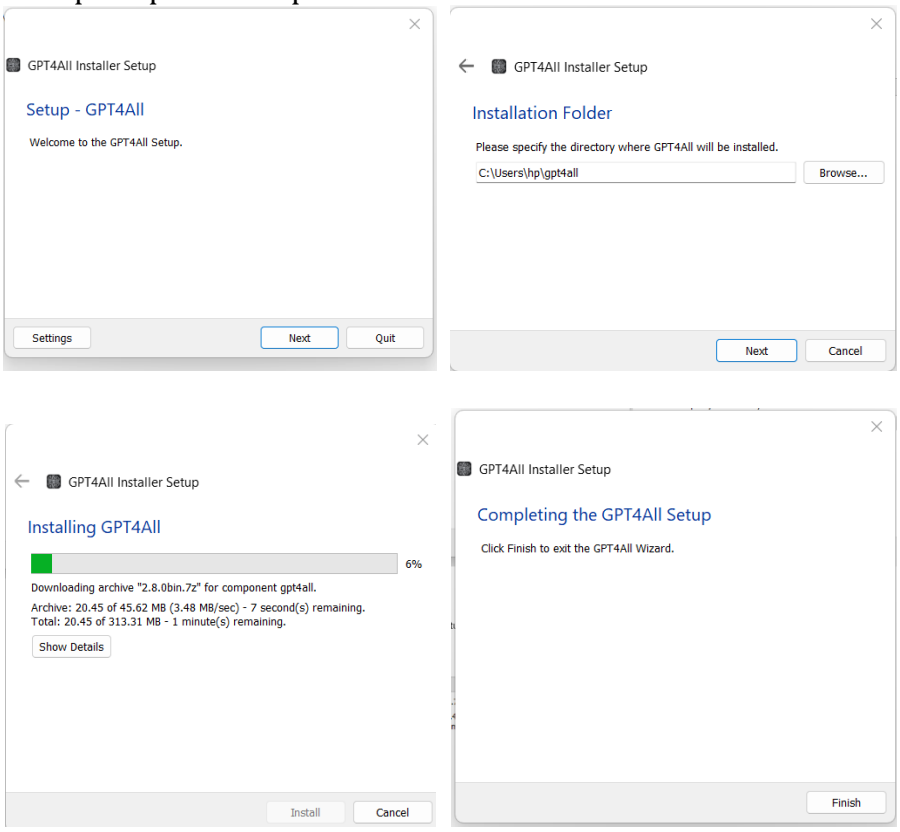
1. Download and install GPT4ALL from https://gpt4all.io/index.html



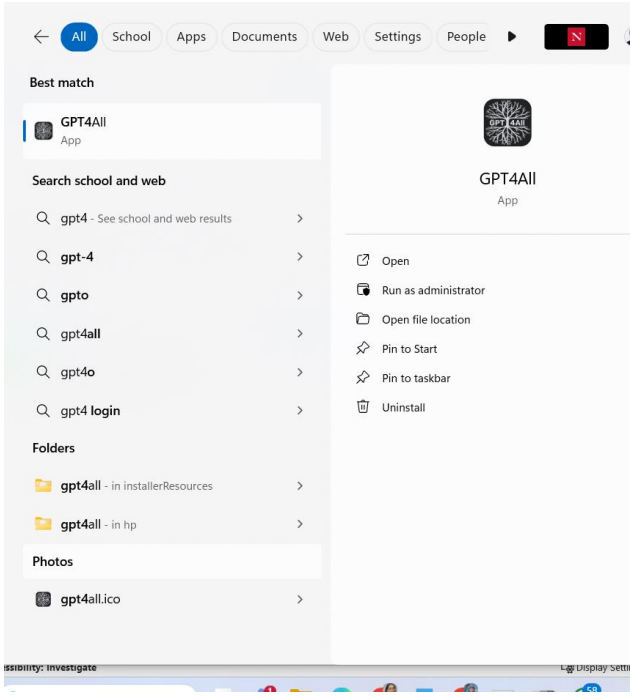
2. Run .exe file



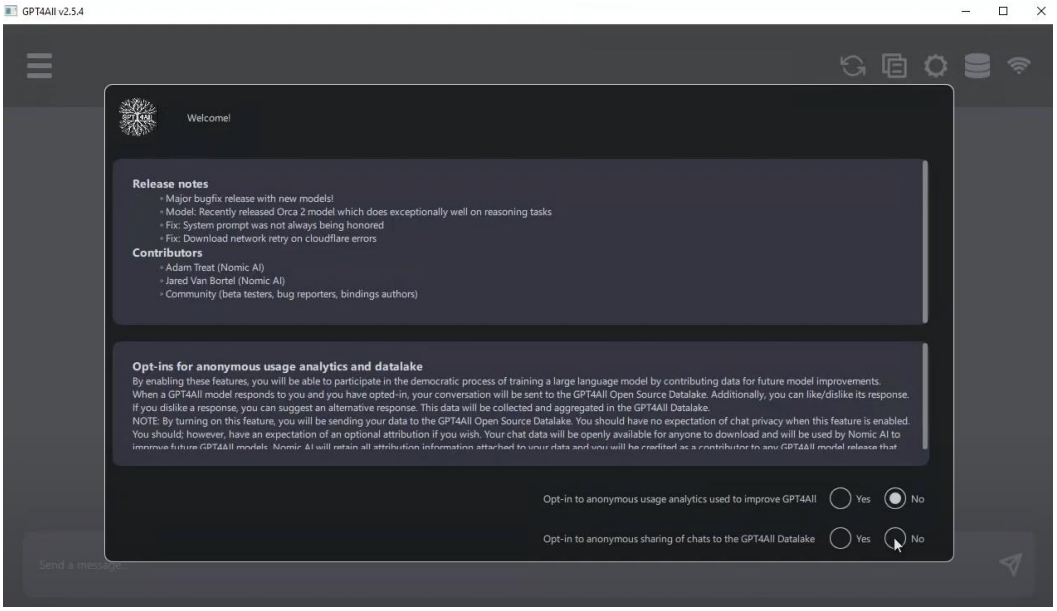
3. Follow the prompts to complete the installation



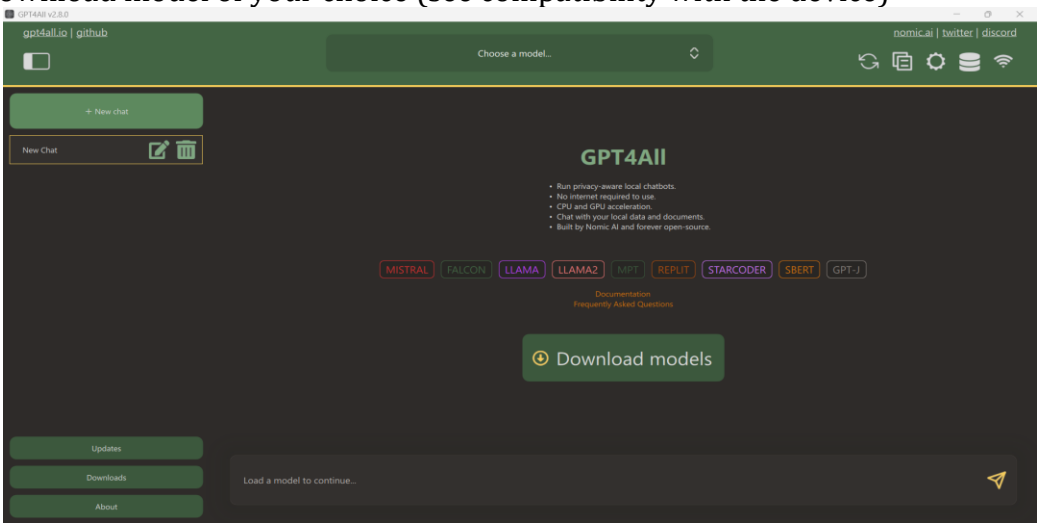
4. Now, after successful installation, search GPT4All and open



5. Select No in both

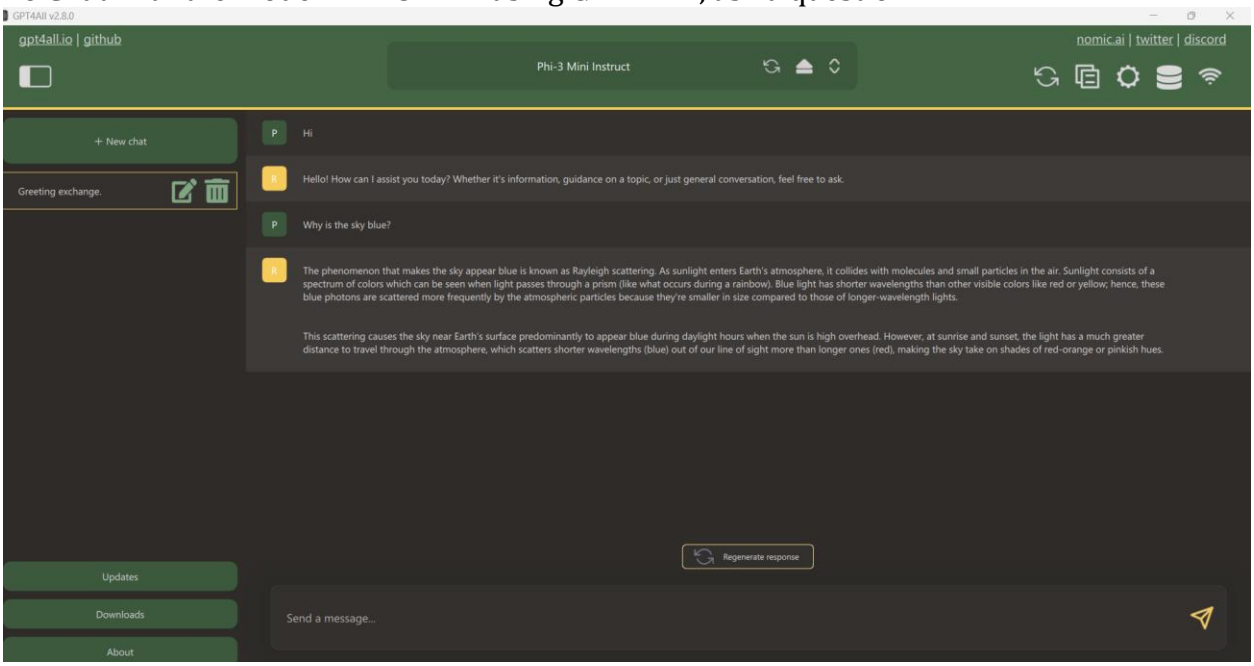


6. Download model of your choice (see compatibility with the device)



7. I am installing Phi3

8. To Chat with the model – Phi3 Mini using GPT4ALL, ask a question



➢ Model 4: orca-mini-3b using GPT4ALL Python Client

Install GPT4ALL Python Client–

1. Create a virtual environment using command –
   python -m venv gpt4all-venv

2. Activate virtual environment –
   gpt4all-venv\Scripts\activate

```
C:\Workspace-PromptEng>python -m venv gpt4all-venv

C:\Workspace-PromptEng>gpt4all-venv\Scripts\activate
```

3. (after activating, you will see (gpt4all-venv) before the project directory) like this –
```
(gpt4all-venv) C:\Workspace-PromptEng>
```

4. Then install GPT4ALL –
(gpt4all-venv) C:\Workspace-PromptEng>pip install gpt4all

```
(gpt4all-venv) C:\Workspace-PromptEng>pip install gpt4all
Collecting gpt4all
  Downloading gpt4all-2.7.0-py3-none-win_amd64.whl.metadata (4.7 kB)
Collecting requests (from gpt4all)
  Using cached requests-2.32.3-py3-none-any.whl.metadata (4.6 kB)
Collecting tqdm (from gpt4all)
  Using cached tqdm-4.66.4-py3-none-any.whl.metadata (57 kB)
Collecting charset-normalizer<4,>=2 (from requests->gpt4all)
  Using cached charset_normalizer-3.3.2-cp312-cp312-win_amd64.whl.metadata (34 kB)
Collecting idna<4,>=2.5 (from requests->gpt4all)
  Using cached idna-3.7-py3-none-any.whl.metadata (9.9 kB)
Collecting urllib3<3,>=1.21.1 (from requests->gpt4all)
  Downloading urllib3-2.2.2-py3-none-any.whl.metadata (6.4 kB)
Collecting certifi>=2017.4.17 (from requests->gpt4all)
  Using cached certifi-2024.6.2-py3-none-any.whl.metadata (2.2 kB)
Collecting colorama (from tqdm->gpt4all)
  Using cached colorama-0.4.6-py2.py3-none-any.whl.metadata (17 kB)
Downloading gpt4all-2.7.0-py3-none-win_amd64.whl (28.6 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 28.6/28.6 MB 3.5 MB/s eta 0:00:00
Using cached requests-2.32.3-py3-none-any.whl (64 kB)
Using cached tqdm-4.66.4-py3-none-any.whl (78 kB)
Using cached certifi-2024.6.2-py3-none-any.whl (164 kB)
Using cached charset_normalizer-3.3.2-cp312-cp312-win_amd64.whl (100 kB)
```

5. Then install flask –
(gpt4all-venv) C:\Workspace-PromptEng\gpt4all-venv>pip install flask

```
(gpt4all-venv) C:\Workspace-PromptEng\gpt4all-venv>pip install flask
Collecting flask
  Downloading flask-3.0.3-py3-none-any.whl.metadata (3.2 kB)
Collecting Werkzeug>=3.0.0 (from flask)
  Downloading werkzeug-3.0.3-py3-none-any.whl.metadata (3.7 kB)
Collecting Jinja2>=3.1.2 (from flask)
  Using cached jinja2-3.1.4-py3-none-any.whl.metadata (2.6 kB)
Collecting itsdangerous>=2.1.2 (from flask)
  Downloading itsdangerous-2.2.0-py3-none-any.whl.metadata (1.9 kB)
Collecting click>=8.1.3 (from flask)
  Using cached click-8.1.7-py3-none-any.whl.metadata (3.0 kB)
Collecting blinker>=1.6.2 (from flask)
  Using cached blinker-1.8.2-py3-none-any.whl.metadata (1.6 kB)
Requirement already satisfied: colorama in c:\workspace-prompteng\gpt4all-venv\lib\site-packages (from click>=8.1.3->fla
sk) (0.4.6)
Collecting MarkupSafe>=2.0 (from Jinja2>=3.1.2->flask)
  Using cached MarkupSafe-2.1.5-cp312-cp312-win_amd64.whl.metadata (3.1 kB)
Downloading flask-3.0.3-py3-none-any.whl (101 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 101.7/101.7 kB 973.4 kB/s eta 0:00:00
Using cached blinker-1.8.2-py3-none-any.whl (9.5 kB)
Using cached click-8.1.7-py3-none-any.whl (97 kB)
```

6. Check flask version –

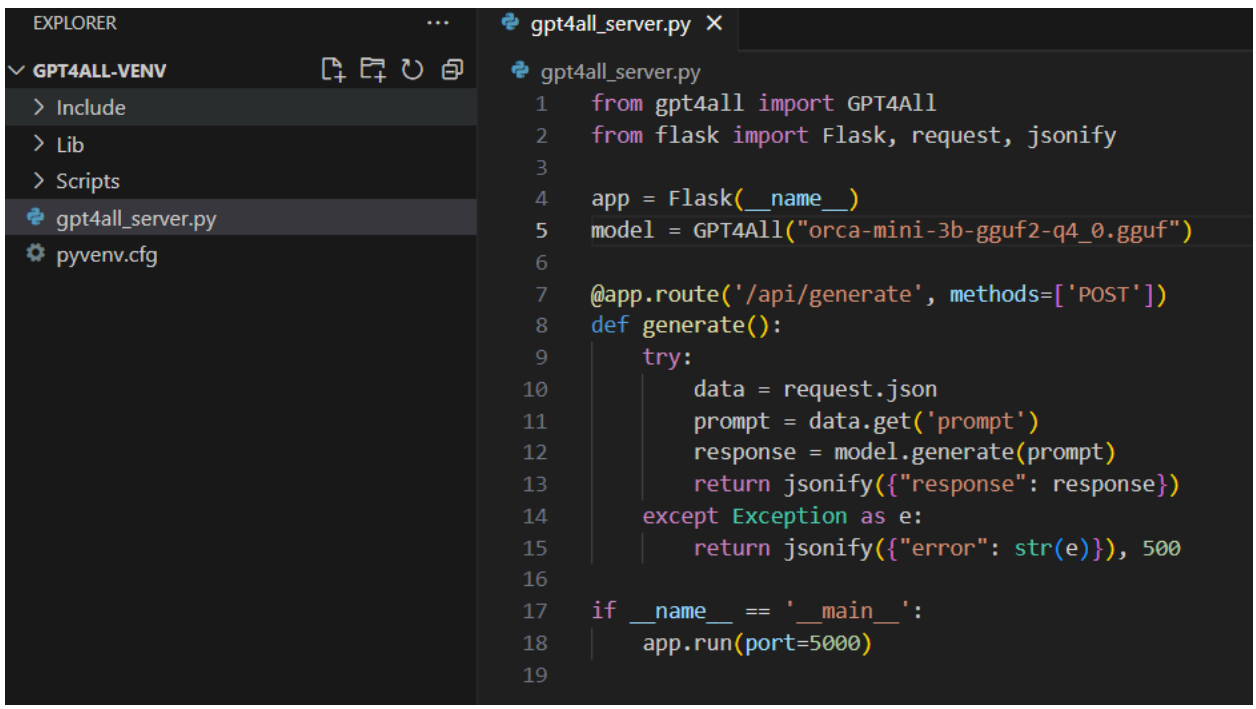
(gpt4all-venv) C:\Workspace-PromptEng\gpt4all-venv>flask –version

```
(gpt4all-venv) C:\Workspace-PromptEng\gpt4all-venv>flask --version
Python 3.12.4
Flask 3.0.3
Werkzeug 3.0.3
```

orca-mini-3b: Interaction using curl command

1. Firstly, create a simple python script "gpt4all_server.py" as shown below in GPT4ALL-VENV folder and use the below GPT4ALL model

GPT4All model = "orca-mini-3b-gguf2-q4_0.gguf"



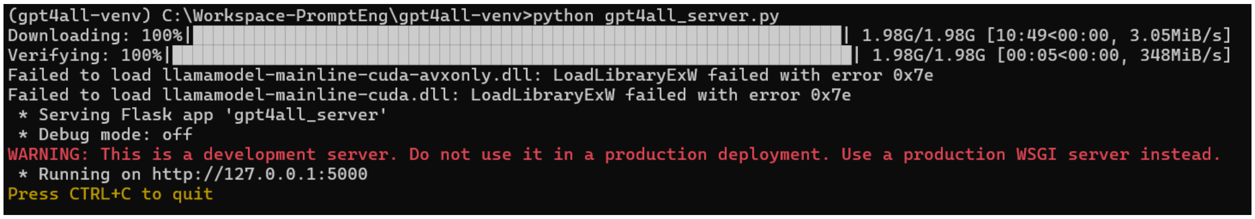
2. Run the python script gpt4all_server.py using below command

(gpt4all-venv) C:\Workspace-PromptEng\gpt4all-venv>python gpt4all_server.py

It will start downloading

```
(gpt4all-venv) C:\Workspace-PromptEng\gpt4all-venv>python gpt4all_server.py
Downloading:  44%|                                                  | 872M/1.98G [04:56<04:46, 3.87MiB/s]
```

And then our server will be running on http://127.0.0.1:5000

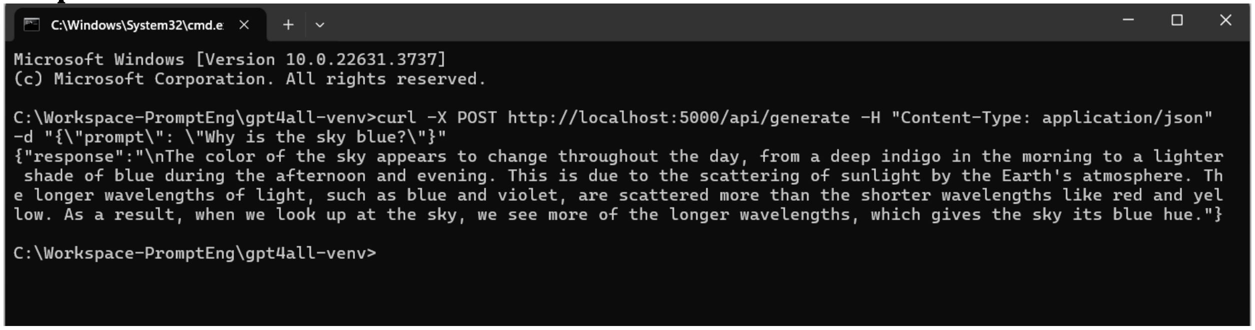


```
(gpt4all-venv) C:\Workspace-PromptEng\gpt4all-venv>python gpt4all_server.py
Downloading: 100%|████████████████████████████████████████████████████████████████| 1.98G/1.98G [10:49<00:00, 3.05MiB/s]
Verifying: 100%|██████████████████████████████████████████████████████████████████| 1.98G/1.98G [00:05<00:00, 348MiB/s]
Failed to load llamamodel-mainline-cuda-avxonly.dll: LoadLibraryExW failed with error 0x7e
Failed to load llamamodel-mainline-cuda.dll: LoadLibraryExW failed with error 0x7e
 * Serving Flask app 'gpt4all_server'
 * Debug mode: off
WARNING: This is a development server. Do not use it in a production deployment. Use a production WSGI server instead.
 * Running on http://127.0.0.1:5000
Press CTRL+C to quit
```

3. Now open new command prompt to interact with curl command

Request: curl -X POST http://localhost:5000/api/generate -H "Content-Type: application/json" -d "{\"prompt\": \"Why is the sky blue?\"}"

Response:

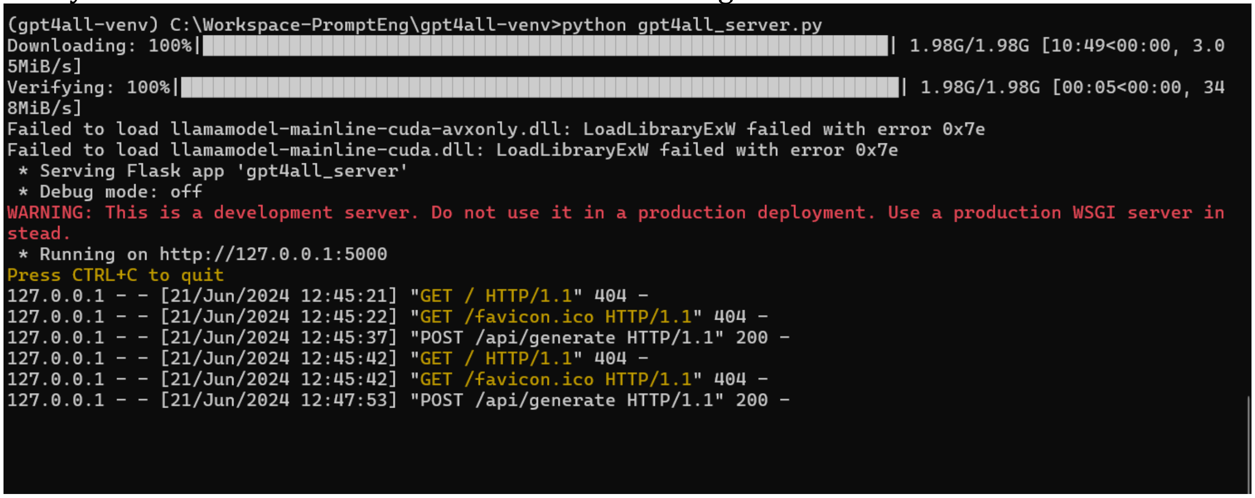


```
Microsoft Windows [Version 10.0.22631.3737]
(c) Microsoft Corporation. All rights reserved.

C:\Workspace-PromptEng\gpt4all-venv>curl -X POST http://localhost:5000/api/generate -H "Content-Type: application/json"
-d "{\"prompt\": \"Why is the sky blue?\"}"
{"response":"\nThe color of the sky appears to change throughout the day, from a deep indigo in the morning to a lighter
 shade of blue during the afternoon and evening. This is due to the scattering of sunlight by the Earth's atmosphere. Th
e longer wavelengths of light, such as blue and violet, are scattered more than the shorter wavelengths like red and yel
low. As a result, when we look up at the sky, we see more of the longer wavelengths, which gives the sky its blue hue."}

C:\Workspace-PromptEng\gpt4all-venv>
```

Also you can see the cmd where our server is running



```
(gpt4all-venv) C:\Workspace-PromptEng\gpt4all-venv>python gpt4all_server.py
Downloading: 100%|████████████████████████████████████████████████████████| 1.98G/1.98G [10:49<00:00, 3.0
5MiB/s]
Verifying: 100%|███████████████████████████████████████████████████████████| 1.98G/1.98G [00:05<00:00, 34
8MiB/s]
Failed to load llamamodel-mainline-cuda-avxonly.dll: LoadLibraryExW failed with error 0x7e
Failed to load llamamodel-mainline-cuda.dll: LoadLibraryExW failed with error 0x7e
 * Serving Flask app 'gpt4all_server'
 * Debug mode: off
WARNING: This is a development server. Do not use it in a production deployment. Use a production WSGI server in
stead.
 * Running on http://127.0.0.1:5000
Press CTRL+C to quit
127.0.0.1 - - [21/Jun/2024 12:45:21] "GET / HTTP/1.1" 404 -
127.0.0.1 - - [21/Jun/2024 12:45:22] "GET /favicon.ico HTTP/1.1" 404 -
127.0.0.1 - - [21/Jun/2024 12:45:37] "POST /api/generate HTTP/1.1" 200 -
127.0.0.1 - - [21/Jun/2024 12:45:42] "GET / HTTP/1.1" 404 -
127.0.0.1 - - [21/Jun/2024 12:45:42] "GET /favicon.ico HTTP/1.1" 404 -
127.0.0.1 - - [21/Jun/2024 12:47:53] "POST /api/generate HTTP/1.1" 200 -
```

Request:
curl -X POST http://localhost:5000/api/generate -H "Content-Type: application/json" -d "{\"prompt\": \"Explain the theory of relativity.\", \"max_tokens\": 100}"

Request: curl -X POST http://localhost:5000/api/generate -H "Content-Type: application/json" -d "{\"prompt\": \"Write a Python function to calculate the factorial of a number.\", \"max_tokens\": 50}"

Response:

```
C:\Workspace-PromptEng\gpt4all-venv>curl -X POST http://localhost:5000/api/generate -H "Content-Type: application/json"
-d "{\"prompt\": \"Explain the theory of relativity.\", \"max_tokens\": 100}"
{"response":"\nThe theory of relativity is a scientific theory proposed by Albert Einstein in 1905 and further developed
 in 1915. The theory of relativity explains how space, time, and gravity work together in our universe. According to thi
s theory, time and space are not absolute entities but are relative to the observer's position and motion. Gravity is al
so a fundamental force that exists between any two objects with mass or energy.\nThe theory of relativity has had a sign
ificant impact on modern physics and has led to many technological advancements, such as the development of GPS technolo
gy and the understanding of black holes. The theory of relativity has also been used in various fields, including astron
omy, particle physics, and cosmology."}

C:\Workspace-PromptEng\gpt4all-venv>curl -X POST http://localhost:5000/api/generate -H "Content-Type: application/json"
-d "{\"prompt\": \"Write a Python function to calculate the factorial of a number.\", \"max_tokens\": 50}"
{"response":"\n\n```python\ndef factorial(n):\n if n == 0:\n return 1\n else:\n return n * factorial(n-1)\n```\n\nThis f
unction takes in an integer `n`, and returns the result of multiplying `n` by the factorial of `n-1`. The factorial of a
 number is the product of all positive integers from 1 to that number. For example, the factorial of 5 is 5 x 4 x 3 x 2
x 1 = 120."}
```

Request:

curl -X POST http://localhost:5000/api/generate -H "Content-Type: application/json" -d "{\"prompt\": \"Tell me how to cook pasta?\", \"max_tokens\": 50}"

Response:

```
C:\Workspace-PromptEng\gpt4all-venv>curl -X POST http://localhost:5000/api/generate -H "Content-Type: application/json"
-d "{\"prompt\": \"Tell me how to cook pasta?\", \"max_tokens\": 50}"
{"response":"\nTo cook pasta, follow these simple steps:\n1. Bring a large pot of salted water to a boil. Add the pasta
and cook until al dente, according to package instructions. Drain the pasta and set it aside.\n2. While the pasta is coo
king, heat a tablespoon of olive oil in a large skillet over medium-high heat. Add your choice of vegetables (such as mu
shrooms, onions, bell peppers, etc.) and cook until they are tender, stirring occasionally.\n3. Once the vegetables are
cooked, add a cup of pasta sauce to the skillet and stir everything together. Cook for about 5 minutes, or until the sau
ce has thickened slightly.\n4. Add the cooked pasta back into the skillet with the vegetables and sauce. Stir everything
 together and cook for another few minutes until the pasta is heated through.\n5. Serve hot and enjoy your delicious pas
ta dish!"}
```