SYMBOLA

PrivateGPT on Xubuntu

Table of Contents

Disclaimer

This guide is provided for informational purposes only. Every effort has been made to ensure the accuracy and completeness of this guide, but it is provided “as is” without warranty of any kind, express or implied. The author shall not be held liable for any direct, indirect, incidental, or consequential damages or losses arising from the use of this guide.

The procedures and software described in this guide are subject to change and may not be up-to-date. Users are advised to exercise caution and consider their specific circumstances when following the instructions.

This guide may contain links to external websites. The author is not responsible for the content or accuracy of any external site.

Please use this guide responsibly and at your own risk.

Contact the Author

If you've spotted an error or simply wish to make contact, feel free to leave a message at:

https://symbola.co.uk/contact/

Your feedback and inquiries are always welcome!

Preface

This installation guide is designed to assist users in setting up and utilizing PrivateGPT, an innovative AI project that marks a significant step forward in addressing data privacy concerns associated with the use of Large Language Models (LLMs). For this installation, we will be using Xubuntu, a lightweight version of Ubuntu, known for its efficiency and compatibility with a wide range of hardware. It is important to note that the setup requires a Nvidia GPU with a sufficient amount of RAM to effectively load and operate the LLM models. This hardware prerequisite is crucial for harnessing the full capabilities of PrivateGPT. In the current technological landscape, the integration of AI into various sectors is increasingly common, yet it frequently raises critical issues regarding data security. PrivateGPT emerges as a noteworthy solution, offering a unique, production-ready platform for querying documents with the advanced capabilities of LLMs in a completely private, offline environment.

Overview of PrivateGPT

PrivateGPT is characterized by its sophisticated API, which is thoughtfully crafted to enable the development of private, context-aware AI applications. This API not only adheres to but also expands upon OpenAI API standards, facilitating a comprehensive and flexible experience for both beginners and experienced users alike. The inclusion of features for both normal and streaming responses adds to its utility.

The API is divided into two primary segments:

1. High-level API: This segment is designed to ease the complexity involved in deploying a Retrieval Augmented Generation (RAG) pipeline. It encompasses a wide range of functionalities, including:
   - ⇒ Document ingestion, which handles parsing, splitting, metadata extraction, embedding generation, and storage.
   - ⇒ Chat and Completions, utilizing the context from ingested documents and simplifying prompt engineering and response generation.
2. Low-level API: Tailored for advanced users, this section allows for the creation of complex pipelines and offers:
   - ⇒ Generation of embeddings based on textual inputs.
   - ⇒ Retrieval of contextual chunks to fetch relevant text segments from ingested documents in response to specific queries.

Moreover, PrivateGPT includes a practical Gradio UI client for API testing and a suite of tools to enhance user experience, such as a bulk model download script, ingestion script, and a documents folder watch feature.

The Genesis and Evolution of PrivateGPT

Initially launched in May 2023, PrivateGPT was primarily targeted at industries where data privacy is paramount, such as healthcare and legal. This early version laid the groundwork for today's more advanced and user-friendly PrivateGPT, preserving the essence of a private, locally-operated ChatGPT-like tool.

PrivateGPT is now on a path toward becoming a more inclusive gateway for generative AI models and primitives. The goal is to simplify the development of AI applications and provide a robust architecture for community contributions.

Documentation and Architecture

For detailed guidance on installation, configuration, and API usage, users are directed to the comprehensive PrivateGPT Documentation at docs.privategpt.dev. The API, developed using FastAPI, is built around a RAG pipeline based on LlamaIndex, reflecting a commitment to simplicity, adaptability, and efficient utilization of existing abstractions.

Community Engagement and Contributions

Contributions to PrivateGPT are highly encouraged. The project benefits from a collaborative community, and systems are in place to maintain code quality and support effective contributions.

Conclusion

PrivateGPT stands as a significant development in the AI domain, particularly for those seeking to leverage AI capabilities while ensuring data privacy. This installation guide is intended to facilitate a smooth setup and effective use of PrivateGPT, catering to a diverse range of users, from AI novices to experts. By engaging with PrivateGPT, users join a journey into the realm of secure, powerful, and accessible AI technologies.

Overview of Xubuntu

Xubuntu is a lightweight and efficient variant of Ubuntu, itself based on the Debian operating system. As a derivative of Ubuntu, Xubuntu shares the same robust architecture and reliable core system but is distinguished by its use of the XFCE desktop environment. This choice of desktop environment makes Xubuntu significantly lighter on system resources, thereby offering a faster and more responsive experience, especially on older or less powerful hardware.

The design philosophy of Xubuntu focuses on simplicity and elegance, providing users with a user-friendly interface and a customizable desktop experience. Despite its lightweight nature, Xubuntu comes packed with all the essential features and applications one would expect from a full-fledged desktop operating system. This includes a wide range of standard applications for web browsing, email, office productivity, and multimedia playback.

For developers and users interested in AI and machine learning, Xubuntu offers a stable and efficient platform. Its compatibility with a variety of hardware, including Nvidia GPUs, makes it an excellent choice for computing-intensive tasks like loading and operating Large Language Models. The operating system's low overhead allows more system resources to be dedicated to these demanding processes, ensuring smoother performance and quicker processing times.

In summary, Xubuntu retains the flexibility and extensive software repositories of Ubuntu but with a lighter footprint, making it an ideal choice for those who require a stable, efficient, and user-friendly Linux distribution for their AI and machine learning projects.

What to Expect from This Guide

This guide is intended for users seeking to integrate and utilize PrivateGPT within a Xubuntu-based environment. Understanding the scope and prerequisites of this guide will help ensure a successful installation and usage experience.

Pre-Installation Requirements:

- ⇒ ***System Configuration:*** This guide assumes the presence of a fully set up and operational computer system running Xubuntu or a similar Ubuntu-based distribution. The selection of Xubuntu is based on its suitability for efficiently handling the requirements of PrivateGPT.

Guides Focus:

- ⇒ ***PrivateGPT Installation and Operation:*** The emphasis of this guide is on providing a comprehensive walk-through for the installation and basic operation of PrivateGPT within a Xubuntu environment.
- ⇒ ***Scope Limited to PrivateGPT:*** It should be noted that the guide will not delve into the broader aspects of Xubuntu usage. The content is strictly confined to what is necessary for the successful deployment and operation of PrivateGPT.

Expected User Competencies:

- ⇒ ***Proficiency in Xubuntu:*** Users should possess an intermediate understanding of Xubuntu's functionalities, including its desktop environment and file management systems.
- ⇒ ***Terminal Command Proficiency:*** A crucial component of this guide is the execution of commands within the terminal. Users should be comfortable with this aspect, as it is integral to the installation and configuration process of PrivateGPT.

To summarize, this guide is specifically tailored for users with a foundational knowledge of Xubuntu or similar Linux distributions who seek to implement the PrivateGPT AI project. It provides a structured and detailed approach to installing and operating PrivateGPT, with a clear emphasis on leveraging its capabilities for secure AI applications, while ensuring data privacy is maintained.

Hardware / Software Requirements for Setting up PrivateGPT

In this section, details of essential hardware requirements needed to successfully set up and operate PrivateGPT will be discussed. Ensuring your system meets these requirements is critical for optimal performance and functionality of the PrivateGPT application.

Crucial Hardware Specifications:

1. Operating System:
   - ⇒ A computer running Xubuntu or a comparable Ubuntu-based distribution is required. The choice of Xubuntu is recommended due to its compatibility and efficiency in handling the demands of PrivateGPT.
2. Graphics Processing Unit (GPU):
   - ⇒ An NVIDIA GPU with at least 8GB of VRAM is essential. Large Language Models, such as those used in PrivateGPT, are highly resource-intensive, and a robust GPU is necessary for effective processing and smooth operation.

Software Requirement:

1. Python Version:
   - ⇒ Python 3.11 is specifically required for the operation of PrivateGPT. It is crucial to use this exact version to ensure full compatibility with the various libraries and frameworks that PrivateGPT relies on.

Installation Recommendations:

1. Clean Installation of Xubuntu:
   - ⇒ For best results, it is advisable to perform the installation of PrivateGPT on a system with a clean installation of Xubuntu. This approach helps prevent conflicts with pre-existing software and configurations, contributing to a more stable and predictable operating environment for PrivateGPT.

By adhering to these hardware and software requirements, users can establish a solid foundation for the successful deployment and use of PrivateGPT, ensuring a productive and efficient experience with the application.

Installing and Running PrivateGPT on Xubuntu

Note: All terminal commands are shown in bold green.

Python Setup

1. Verify Python Version
   1. Ensure Python 3.11 is installed:
      ```
      python3 --version
      ```
2. Install Python3-pip
   1. Install pip for Python 3:
      ```
      sudo apt install python3-pip
      ```
   2. Confirm pip installation:
      ```
      pip --version
      ```

Git Installation

1. Install Git
   1. Execute:
      ```
      sudo apt install git
      ```

Cloning PrivateGPT Repository

1. Clone the Repository
   1. Get PrivateGPT from GitHub:
      ```
      git clone https://github.com/imartinez/privateGPT
      ```
   2. Switch to the repository directory:
      ```
      cd privateGPT
      ```

Setting Up Python Virtual Environment

1. Install Python Virtual Environment
   1. Install the Python 3.11 virtual environment:
      ```
      sudo apt install python3.11-venv
      ```
2. Create and Activate Virtual Environment
   1. Make a new virtual environment:
      ```
      python3 -m venv privategpt_env
      ```
   2. Activate it:
      ```
      source privategpt_env/bin/activate
      ```

Installing Dependencies with Poetry

1. Install Poetry
    1. Install Poetry within the virtual environment:
        ```
        pip install poetry
        ```
2. Install Dependencies
    1. Add dependencies with UI components:
        ```
        poetry install --with ui
        ```
    2. Add local dependencies:
        ```
        poetry install --with local
        ```

Running Setup Scripts

1. Run Setup Script
    1. Execute the setup script in the repo:
        ```
        poetry run python scripts/setup
        ```
        By default, the script is configured to use the "mistral-7b-instruct-v0.2.Q4_K_M.gguf" model.

Installing CUDA Toolkit

1. Install NVIDIA CUDA Toolkit
    1. Install CUDA for NVIDIA GPUs:
        ```
        sudo apt install nvidia-cuda-toolkit
        ```
2. Check CUDA Version
    1. Verify CUDA installation:
        ```
        nvcc --version
        ```

Final Installation Steps

1. Install Llama-CPP-Python
    1. Use specific CMake arguments to install llama-cpp-python:
        ```
        CMAKE_ARGS='-DLLAMA_CUBLAS=on' poetry run pip install --force-reinstall --no-cache-dir llama-cpp-python
        ```

Running PrivateGPT

1. Launch PrivateGPT
    1. Start PrivateGPT:
        ```
        PGPT_PROFILES=local make run
        ```
    2. The final line of the output from the command above is:
        Uvicorn running on **http://0.0.0.0:8001** (Press CTRL+C to quit)

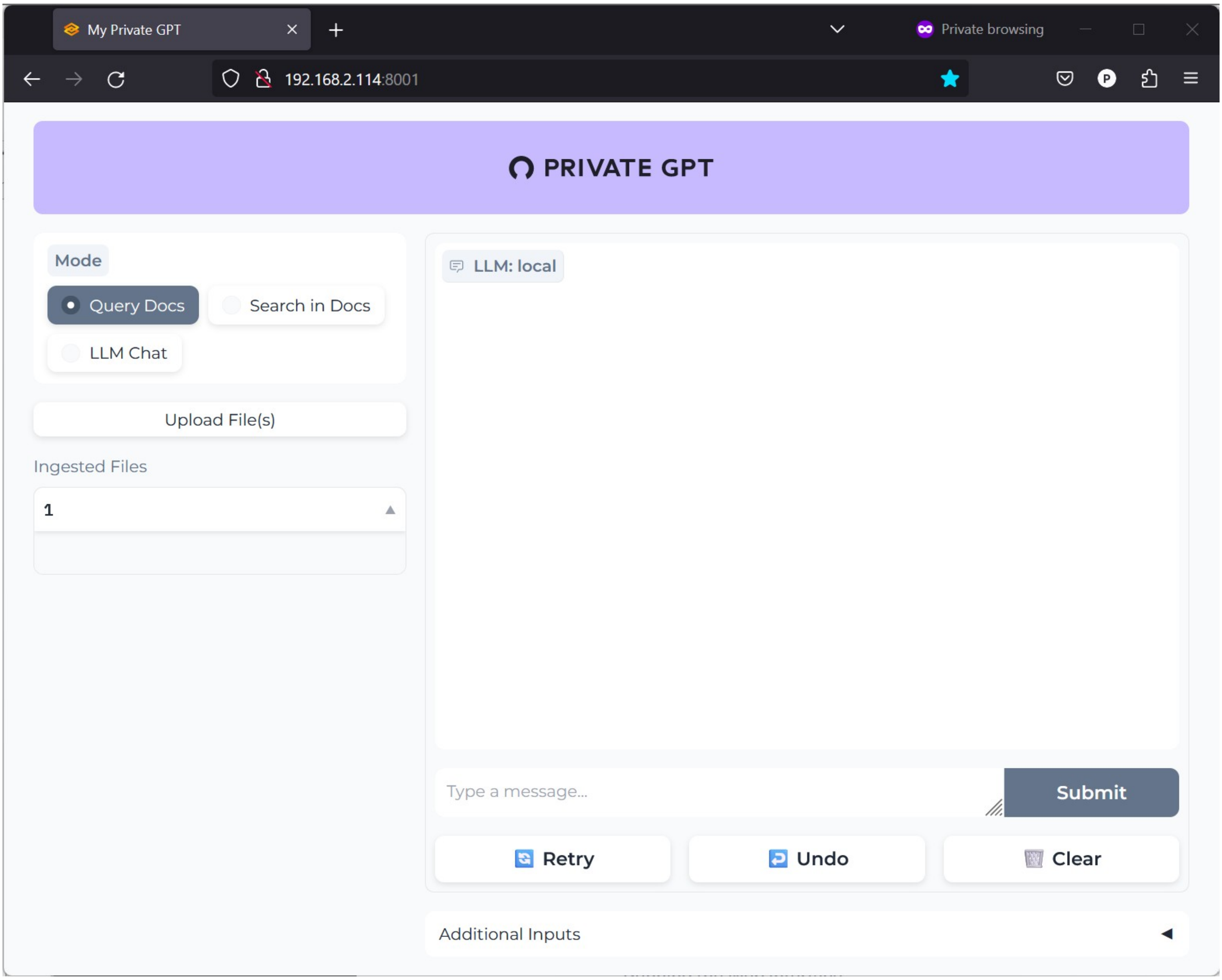
You have successfully set up and started PrivateGPT on your Xubuntu system. Next, learn how to connect to the PrivateGPT interface.

Running the Web Interface

1. To run on the local Xubuntu computer
    1. Open a web browser and navigate to:
       http://0.0.0.0:8001 or http://localhost:8001
2. To run on another computer within the local network
    1. Determine the IP address of the Xubuntu computer by using the command:
       ip a
       Look for an entry similar to "inet 192.168.2.114/24" in the output.
       Note that your IP address will differ from this example.
    2. On a different computer within the local network
        1. Open a web browser and navigate to the URL using the IP address found earlier, such as:
           http://192.168.2.114:8001
           Remember to replace this with the actual IP address you have identified.

Using the Web Interface

The web interface should appear as follows:



To test the system, upload a PDF file and ask questions about its content. Other modes include searching within the document or simply engaging in a chat with the LLM on any subject.

Restarting the Model After a System Reboot

Follow these steps to restart the PrivateGPT model once your system has rebooted:

1. Navigate to the PrivateGPT Directory:
    1. **cd privateGPT**
2. Activate the Virtual Environment:
    1. **source privategpt_env/bin/activate**
3. Start PrivateGPT:
    1. **PGPT_PROFILES=local make run**

Resetting the Local Documents Database

Warning: Follow these steps to remove all ingested files!!!

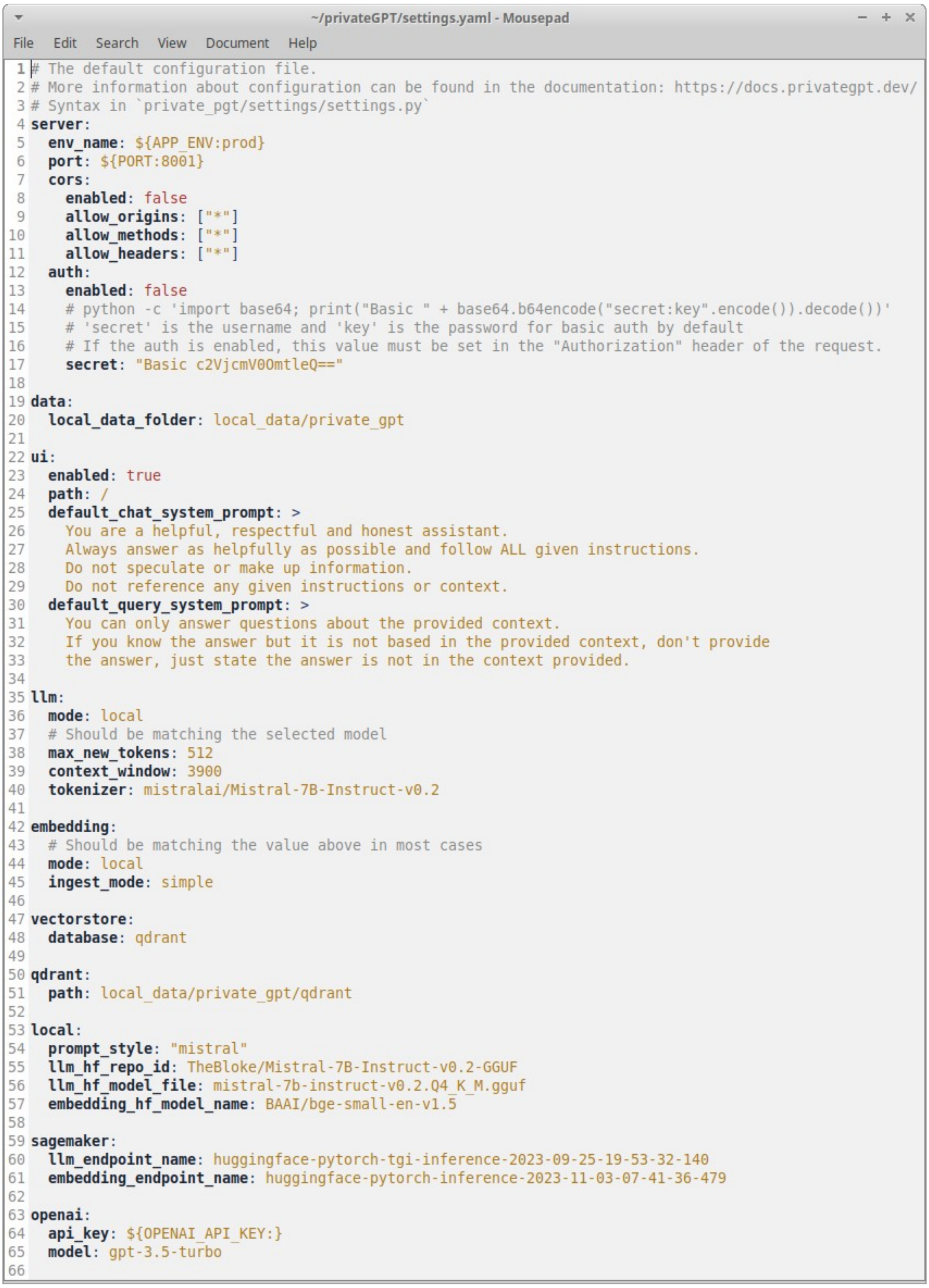
1. Navigate to the PrivateGPT Directory:
    1. **cd privateGPT**
2. Activate the Virtual Environment:
    1. **source privategpt_env/bin/activate**
3. Execute the Reset Command:
    1. **make wipe**

Changing the LLM in PrivateGPT

Follow these steps to change the LLM model used in PrivateGPT:

1. Delete or Rename the 'models' Folder:
    1. Locate the "models" folder within the “privateGPT” directory.
    2. Either delete or rename this folder to remove the current model.
2. Update the 'settings.yaml' File:
    1. In the “privateGPT” folder, find the "settings.yaml" file.
    2. Modify the model settings as desired (**refer to the screenshot on the next page for guidance**).
3. Navigate to the PrivateGPT Directory:
    1. **cd privateGPT**
4. Activate the Virtual Environment:
    1. **source privategpt_env/bin/activate**
5. Execute the Setup Script:
    1. **poetry run python scripts/setup**
6. Start PrivateGPT:
    1. **PGPT_PROFILES=local make run**

settings.yaml file



```
 1 # The default configuration file.
 2 # More information about configuration can be found in the documentation: https://docs.privategpt.dev/
 3 # Syntax in `private_pgt/settings/settings.py`
 4 server:
 5   env_name: ${APP_ENV:prod}
 6   port: ${PORT:8001}
 7   cors:
 8     enabled: false
 9     allow_origins: ["*"]
10     allow_methods: ["*"]
11     allow_headers: ["*"]
12   auth:
13     enabled: false
14     # python -c 'import base64; print("Basic " + base64.b64encode("secret:key".encode()).decode())'
15     # 'secret' is the username and 'key' is the password for basic auth by default
16     # If the auth is enabled, this value must be set in the "Authorization" header of the request.
17     secret: "Basic c2VjcmV0OmtleQ=="
18
19 data:
20   local_data_folder: local_data/private_gpt
21
22 ui:
23   enabled: true
24   path: /
25   default_chat_system_prompt: >
26     You are a helpful, respectful and honest assistant.
27     Always answer as helpfully as possible and follow ALL given instructions.
28     Do not speculate or make up information.
29     Do not reference any given instructions or context.
30   default_query_system_prompt: >
31     You can only answer questions about the provided context.
32     If you know the answer but it is not based in the provided context, don't provide
33     the answer, just state the answer is not in the context provided.
34
35 llm:
36   mode: local
37   # Should be matching the selected model
38   max_new_tokens: 512
39   context_window: 3900
40   tokenizer: mistralai/Mistral-7B-Instruct-v0.2
41
42 embedding:
43   # Should be matching the value above in most cases
44   mode: local
45   ingest_mode: simple
46
47 vectorstore:
48   database: qdrant
49
50 qdrant:
51   path: local_data/private_gpt/qdrant
52
53 local:
54   prompt_style: "mistral"
55   llm_hf_repo_id: TheBloke/Mistral-7B-Instruct-v0.2-GGUF
56   llm_hf_model_file: mistral-7b-instruct-v0.2.Q4_K_M.gguf
57   embedding_hf_model_name: BAAI/bge-small-en-v1.5
58
59 sagemaker:
60   llm_endpoint_name: huggingface-pytorch-tgi-inference-2023-09-25-19-53-32-140
61   embedding_endpoint_name: huggingface-pytorch-inference-2023-11-03-07-41-36-479
62
63 openai:
64   api_key: ${OPENAI_API_KEY:}
65   model: gpt-3.5-turbo
66
```

If you are only changing the model while using the same repository, simply adjust line 56. However, if you are changing both the model and the repository, then modifications are required on both lines 55 and 56.

Finding Other Models

The model used in this guide is available at the following link:
https://huggingface.co/TheBloke/Mistral-7B-Instruct-v0.2-GGUF

You can view a list of files for this repository here:
https://huggingface.co/TheBloke/Mistral-7B-Instruct-v0.2-GGUF/tree/main

Below is a list of models and their sizes from the repository used in this guide:

mistral-7b-instruct-v0.2.Q2_K.gguf	3.08 GB
mistral-7b-instruct-v0.2.Q3_K_L.gguf	3.82 GB
mistral-7b-instruct-v0.2.Q3_K_M.gguf	3.52 GB
mistral-7b-instruct-v0.2.Q3_K_S.gguf	3.16 GB
mistral-7b-instruct-v0.2.Q4_0.gguf	4.11 GB
mistral-7b-instruct-v0.2.Q4_K_M.gguf	4.37 GB
mistral-7b-instruct-v0.2.Q4_K_S.gguf	4.14 GB
mistral-7b-instruct-v0.2.Q5_0.gguf	5.00 GB
mistral-7b-instruct-v0.2.Q5_K_M.gguf	5.13 GB
mistral-7b-instruct-v0.2.Q5_K_S.gguf	5.00 GB
mistral-7b-instruct-v0.2.Q6_K.gguf	5.94 GB
mistral-7b-instruct-v0.2.Q8_0.gguf	7.70 GB

To find other models, visit:
https://huggingface.co/models