Solution Architecture for Gen AI

Ram N Sangwan

- **Solution Guidelines - Well-Architected principles**
- **Chat Session management**
- **Architectures for Various Use Cases**
- **Manage Token limitation**
- **Deployment Standards – Cloud or On-Prem**
- **Private GPT**

Solution Guidelines - Well-Architected principles

Typically focus on key areas to ensure the technology is used effectively and responsibly.

Some insights:

Choice of Foundation Model: depends domain, data, and the specific use case.

Accessibility and Control: on Prem for full control or using them as a managed cloud service for speed and simplicity.

The decision will impact factors like infrastructure management, cost predictability, and complexity.

Solution Guidelines - Well-Architected principles

Architecture Components: data processing layer, a generative model layer, a feedback & improvement layer, and a deployment & integration layer.

Operational and Model Risks: Models must be managed to control privacy and ensure security.

Data and Infrastructure Strategy: Many organizations struggle with siloed, rapidly changing, or low-quality data, which impacts the performance of AI models. It’s essential to have a solid data and infrastructure strategy to support the demands of Gen AI.

Regulatory Compliance and Ethical Considerations

Chat Session Management

Guidelines that could be useful

1. **Ensure User Privacy and Data Security**.
2. **Maintain Transparency**.
3. **Promote Ethical Interactions**.
4. **Manage Expectations**: cut-off knowledge date and its inability to access or retrieve real-time information without specific tools.
5. **Feedback Mechanism**.

Standard Architectures for various use cases

Generative Adversarial Networks (GANs)

Use Case

- Excellent for generating photorealistic images, art, and even enhancing low-resolution images.

Strengths

- Capable of producing high-quality and highly realistic outputs.

Weaknesses

- Training can be unstable and often requires careful tuning of hyperparameters.

Compared to other models, GANs can generate the most visually compelling results but are often harder to train.

Variational Autoencoders (VAEs)

Use Case

- Used for image generation, but with a focus on encoding an input into a latent space and then reconstructing it.

Strengths

- More stable training than GANs and can learn to represent complex probability distributions.

Weaknesses

- Tend to produce less sharp images compared to GANs.

VAEs are more stable and easier to train than GANs but do not match the output quality of GANs in terms of image crispness.

Long Short-Term Memory Networks (LSTM)

Use Case

- Suitable for time-series prediction, music composition, and text generation with a focus on sequences.

Strengths

- Good at capturing time-dependent properties in sequential data.

Weaknesses

- Struggles with long-range dependencies and is less effective than transformer models for longer sequences.

LSTMs are more efficient than transformers but lack the same performance on complex tasks requiring understanding of long contexts.

Auto-Regressive Models (e.g., PixelRNN, WaveNet)

Use Case

- PixelRNN is used for image generation, while WaveNet is used for generating audio, such as human-like speech or music.

Strengths

- Can produce high-quality outputs by modelling the probability distribution of a sequence one element at a time.

Weaknesses

- The sequential nature of prediction can be slow and computationally expensive

They are excellent in their respective domains (images and audio) but are generally slower than parallelizable models due to their sequential processing nature.

Comparative Summary

- **GANs** are the best for generating realistic images but are difficult to train.

- **LSTMs** are a go-to for simpler sequential tasks where computational efficiency is a concern, though they may falter with very long sequences.

- **VAEs** offer a balance between image quality and training stability and are excellent for tasks that require understanding the underlying data distribution.

- **Transformers** are the gold standard for text-related tasks requiring the understanding of complex contexts but come with high computational costs.

- **Auto-Regressive Models** like PixelRNN and WaveNet excel in their specific domains of image and audio generation, respectively, but are not as fast as parallel processing models.

Manage Token limitations

Manage Token limitations

Summarize Content

Before inputting content into LLM, summarize it to reduce the number of tokens.

Chunking

Process each chunk individually and then synthesize the outputs.

Focused Prompts

Craft your prompts to be as focused as possible.

Iterative Refinement

Iteratively refine the output by asking follow-up questions.

Use Other Models

Consider using smaller models wherever possible, that are more token-efficient.

Increase Efficiency with Commands

Use the model's understanding of commands to execute complex tasks in a token-efficient manner.

Manage Token limitations

Pre-Processing

Use **external tools** for pre-processing tasks like extracting text from images or simplifying sentences to make them more concise.

Post-Processing

After getting the output, use post-processing to stitch together and make sense of the information if you had to break it into parts.

Pipeline Approaches

Create a pipeline that uses different models for different tasks, reserving the large language model for the most complex parts of the task.

Cache Responses

If you are likely to ask the same or similar questions, cache the responses so you don't have to use tokens for the same query again.

Optimize API Usage

If you are using an API, ensure that your API calls are optimized to send and receive as much relevant information as possible within the token limits.

Deployment Standards – Cloud or On-Prem

Deployment Standards – Cloud or On-Prem

- According to 451 Research, 90% of companies use cloud services in some form.
- However, the same report found that 60% of workloads are still run on-premise, indicating a balance between the two.
- In terms of costs, **OCI** can range from a few hundred dollars per month to tens of thousands of dollars per month depending on size of projects.
- However, on-premise solutions can have an upfront cost of several thousand dollars, with additional ongoing costs for maintenance and updates.

Cloud vs. On-Premise Hosting for AI Applications

- These are often compared to renting vs. buying a home.
- Cloud hosting is a lot like renting; the stay of AI applications is as long as the contract terms dictate.
- The maintenance of the hardware is the responsibility of the hosting provider.
- On-premise hosting, on the other hand, is like buying a home; the application can stay on the hardware as long as business requires it.

Cloud vs. On-Premise

Scalability

- On-premise hosting offers complete control over the hardware, which means that the administrators of a company can tightly control updates.
- But on-premise hosting does require advanced planning to scale hardware.
- This is because it requires time to gather the necessary data for updating it.
- Cloud resources can be rapidly adjusted to accommodate specific demands and increase the scalability of hardware.

Cloud vs. On-Premise

Security

- Full control over data stored on enterprise premises. Hosting providers must keep their systems updated and data encrypted to avoid breaches.
- Still, your company can’t be sure where your data is stored and how often it is backed up; data is also accessible by third parties.

Cloud vs. On-Premise

Data Gravity

“*The ability of data to attract applications, services, and other data towards itself*”

It is among the most important factors to be considered while choosing between cloud and on-premise platforms.

- Considering the costs of training neural networks, companies may want to deploy their AI applications on-premises.
- If the data required to build AI applications resides on the cloud, then it’s best to deploy applications there.
- The location of the largest source of data for an enterprise determines the location of its most critical applications.
- Oracle cloud offering **“Cloud@Customer”** can proved to be best here.

The Case for Cloud-Based AI Applications

- Instead of building out a massive data centre, you can use the infrastructure someone else already maintains.
- One reason why AI has become so pervasive is cloud providers offering plug-and-play AI cloud services, with enough compute power and pre-trained models to launch AI applications.
- In many cases the pre-trained models or storage requirements of the cloud can be cost-prohibitive; higher GPU counts get expensive fast and training large datasets on the public cloud can be too slow.
- Still, the cloud can often be the best option in terms of “testing the waters” of AI and experimenting with which AI initiatives work best for an organization.

The Case for On-Premise AI

- There’s a whole ecosystem of tools built for on-premise infrastructure that can work with mass amounts of compute power–which can be very expensive in the cloud.
- Thus, it is more economical to do this on-premise or prefer a capital expense to an operational expense model.
- If your organization wants to get more involved in this or roll-out AI at scale, then it may make more sense to invest in on-premises infrastructure instead of consuming cloud-based services.

Private GPT

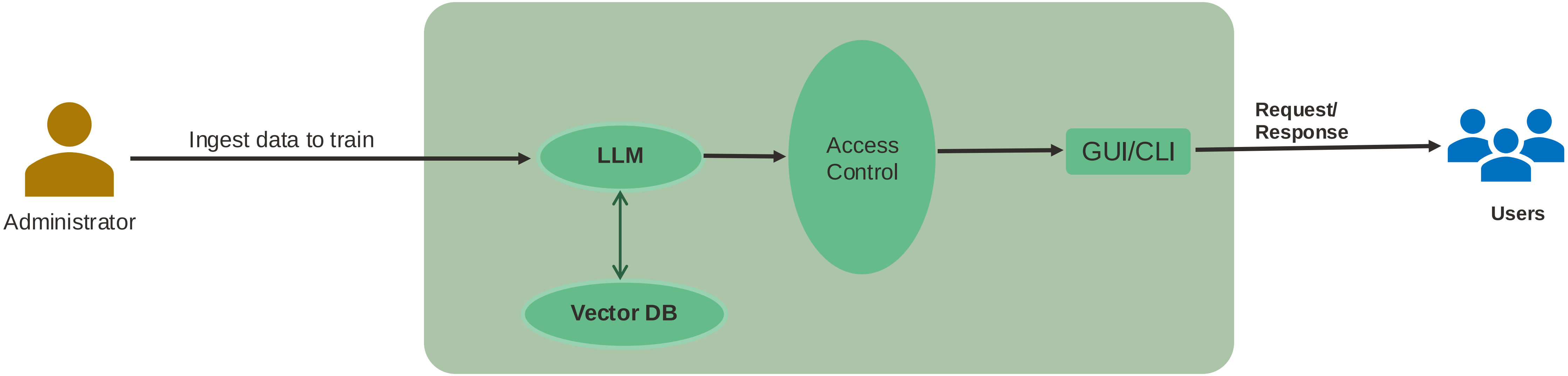
What is Private GPT?

- Private GPT allows enterprises to tap into the remarkable capabilities of large language models while prioritizing privacy and security.
- Private GPT operates on cloud or on-premises within an organization’s own servers and data centres.
- Use Cases
  - **Knowledge Management**.
  - **Customer Support**.
  - **Content Creation**.
  - **Data Analysis**.
  - **Automate Workflows**

Components of Private GPT and How Does it Work?

Private GPT is requires multiple components to work together to function.

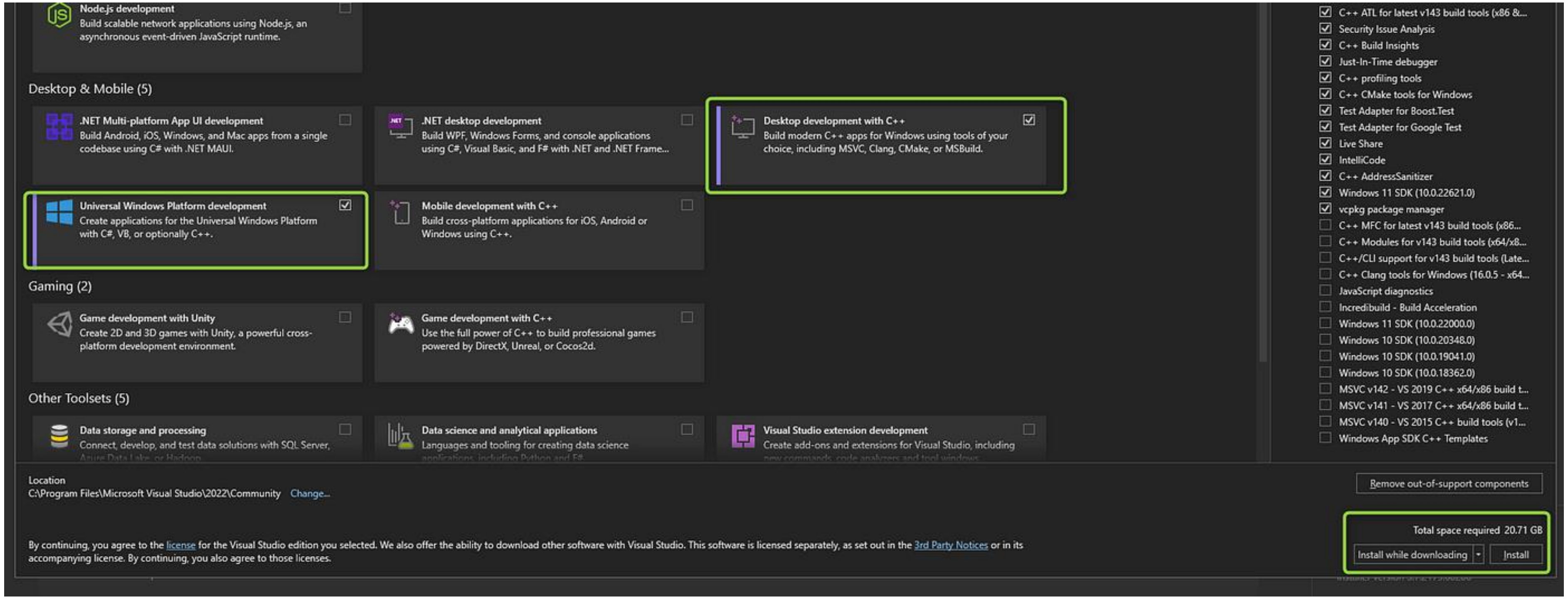
- **Private LLMs**: Private GPT supports proprietary models like GPT4ALL and LLAMA.



Private GPT - For Windows

Install Visual Studio 2022 Components

1. Universal Windows Platform development
2. Desktop Development with C++



☐ **Node.js development**
Build scalable network applications using Node.js, an asynchronous event-driven JavaScript runtime.

Desktop & Mobile (5)

☐ **.NET Multi-platform App UI development**
Build Android, iOS, Windows, and Mac apps from a single codebase using C# with .NET MAUI.

☐ **.NET desktop development**
Build WPF, Windows Forms, and console applications using C#, Visual Basic, and F# with .NET and .NET Frame...

☑ **Desktop development with C++**
Build modern C++ apps for Windows using tools of your choice, including MSVC, Clang, CMake, or MSBuild.

☑ **Universal Windows Platform development**
Create applications for the Universal Windows Platform with C#, VB, or optionally C++.

☐ **Mobile development with C++**
Build cross-platform applications for iOS, Android or Windows using C++.

Gaming (2)

☐ **Game development with Unity**
Create 2D and 3D games with Unity, a powerful cross-platform development environment.

☐ **Game development with C++**
Use the full power of C++ to build professional games powered by DirectX, Unreal, or Cocos2d.

Other Toolsets (5)

☐ **Data storage and processing**
Connect, develop, and test data solutions with SQL Server, Azure Data Lake, or Hadoop.

☐ **Data science and analytical applications**
Languages and tooling for creating data science applications, including Python and F#.

☐ **Visual Studio extension development**
Create add-ons and extensions for Visual Studio, including new commands, code analyzers and tool windows.

- ☑ C++ ATL for latest v143 build tools (x86 &...
- ☑ Security Issue Analysis
- ☑ C++ Build Insights
- ☑ Just-In-Time debugger
- ☑ C++ profiling tools
- ☑ C++ CMake tools for Windows
- ☑ Test Adapter for Boost.Test
- ☑ Test Adapter for Google Test
- ☑ Live Share
- ☑ IntelliCode
- ☑ C++ AddressSanitizer
- ☑ Windows 11 SDK (10.0.22621.0)
- ☑ vcpkg package manager
- ☐ C++ MFC for latest v143 build tools (x86...
- ☐ C++ Modules for v143 build tools (x64/x8...
- ☐ C++/CLI support for v143 build tools (Late...
- ☐ C++ Clang tools for Windows (16.0.5 - x64...
- ☐ JavaScript diagnostics
- ☐ Incredibuild - Build Acceleration
- ☐ Windows 11 SDK (10.0.22000.0)
- ☐ Windows 10 SDK (10.0.20348.0)
- ☐ Windows 10 SDK (10.0.19041.0)
- ☐ Windows 10 SDK (10.0.18362.0)
- ☐ MSVC v142 - VS 2019 C++ x64/x86 build t...
- ☐ MSVC v141 - VS 2017 C++ x64/x86 build t...
- ☐ MSVC v140 - VS 2015 C++ build tools (v1...
- ☐ Windows App SDK C++ Templates

Location
C:\Program Files\Microsoft Visual Studio\2022\Community Change...

By continuing, you agree to the license for the Visual Studio edition you selected. We also offer the ability to download other software with Visual Studio. This software is licensed separately, as set out in the 3rd Party Notices or in its accompanying license. By continuing, you also agree to those licenses.

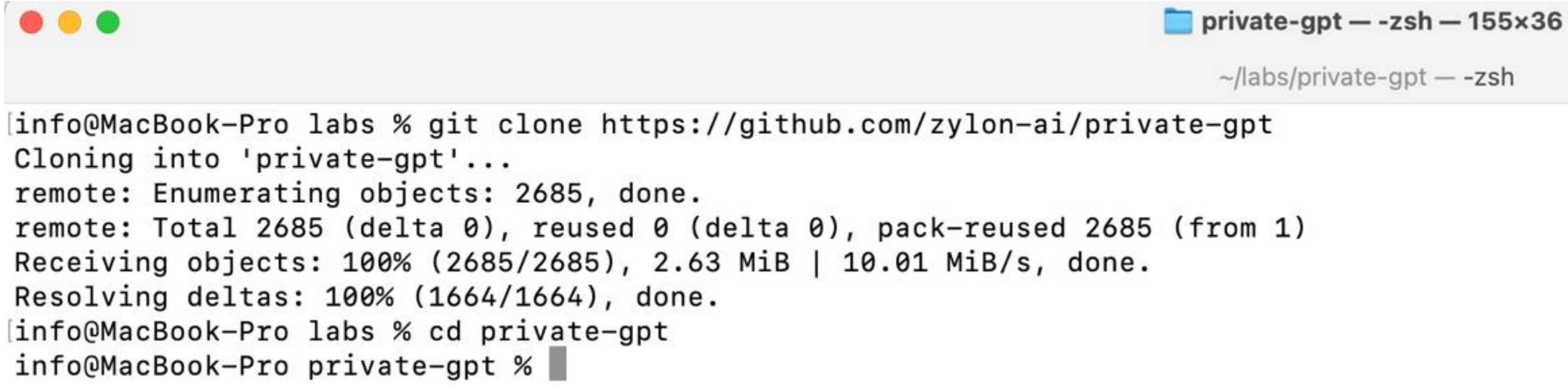
Total space required 20.71 GB

Private GPT – Common for Windows PC/Mac/Linux

Private GPT requires Python 3.11 or later.

- Clone the PrivateGPT Repository and change to the downloaded directory:

```
git clone https://github.com/zylon-ai/private-gpt
cd private-gpt
```



```
info@MacBook-Pro labs % git clone https://github.com/zylon-ai/private-gpt
Cloning into 'private-gpt'...
remote: Enumerating objects: 2685, done.
remote: Total 2685 (delta 0), reused 0 (delta 0), pack-reused 2685 (from 1)
Receiving objects: 100% (2685/2685), 2.63 MiB | 10.01 MiB/s, done.
Resolving deltas: 100% (1664/1664), done.
info@MacBook-Pro labs % cd private-gpt
info@MacBook-Pro private-gpt %
```

Private GPT for Windows PC/Mac/Linux

Import the PrivateGPT into an IDE if Needed

- Open terminal and Install Poetry for dependency management.

  ```
  pip install -qU poetry
  ```

- Install and Run Your Desired Setup
  - PrivateGPT allows customization of the setup, from fully local to cloud-based, by deciding the modules to use.
- ***For Windows**:*
  - *If you are using Windows, you'll need to set the env var, PGPT_PROFILES, in a different way, for example:*

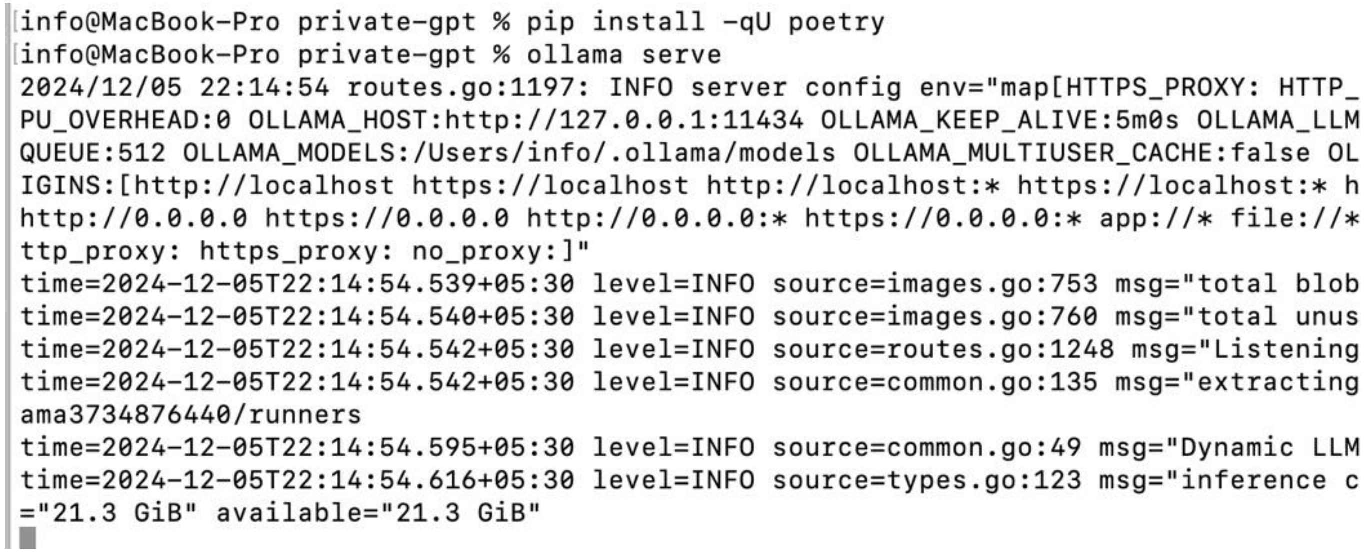
    ```
    # Powershell
    PGPT_PROFILES="ollama"
    make run
    ```

Private GPT for Windows PC/Mac/Linux

Ollama-Powered Local setup

- Ollama provides local LLM and Embeddings super easy to install and use, abstracting the complexity of GPU support.
- Go to ollama.ai and follow the instructions to install Ollama on your machine.
- Start Ollama service. it will start a local inference server, serving both the LLM and the Embeddings):

```
ollama serve
```



```
[info@MacBook-Pro private-gpt % pip install -qU poetry
[info@MacBook-Pro private-gpt % ollama serve
2024/12/05 22:14:54 routes.go:1197: INFO server config env="map[HTTPS_PROXY: HTTP_
PU_OVERHEAD:0 OLLAMA_HOST:http://127.0.0.1:11434 OLLAMA_KEEP_ALIVE:5m0s OLLAMA_LLM
QUEUE:512 OLLAMA_MODELS:/Users/info/.ollama/models OLLAMA_MULTIUSER_CACHE:false OL
IGINS:[http://localhost https://localhost http://localhost:* https://localhost:* h
http://0.0.0.0 https://0.0.0.0 http://0.0.0.0:* https://0.0.0.0:* app://* file://*
ttp_proxy: https_proxy: no_proxy:]"
time=2024-12-05T22:14:54.539+05:30 level=INFO source=images.go:753 msg="total blob
time=2024-12-05T22:14:54.540+05:30 level=INFO source=images.go:760 msg="total unus
time=2024-12-05T22:14:54.542+05:30 level=INFO source=routes.go:1248 msg="Listening
time=2024-12-05T22:14:54.542+05:30 level=INFO source=common.go:135 msg="extracting
ama3734876440/runners
time=2024-12-05T22:14:54.595+05:30 level=INFO source=common.go:49 msg="Dynamic LLM
time=2024-12-05T22:14:54.616+05:30 level=INFO source=types.go:123 msg="inference c
="21.3 GiB" available="21.3 GiB"
```

Private GPT for Windows PC/Mac/Linux

Install the models to be used

- The default `settings-ollama.yaml` is configured to user llama3.1 8b LLM (~4GB) and nomic-embed-text Embeddings (~275MB)
- By default, PGPT will automatically pull models as needed.
- In any case, if you want to manually pull models, run:

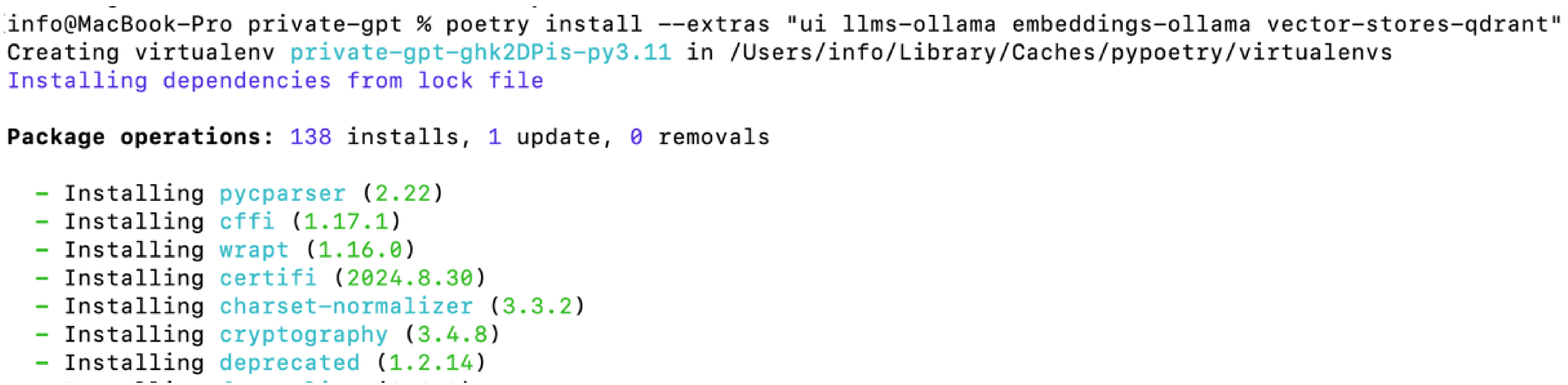
```
ollama pull llama3.1
ollama pull nomic-embed-text
```

Private GPT for Windows PC/Mac/Linux

Install PrivateGPT

- Once done, on a different terminal, install PrivateGPT:

```
poetry install --extras "ui llms-ollama embeddings-ollama vector-stores-qdrant”
```



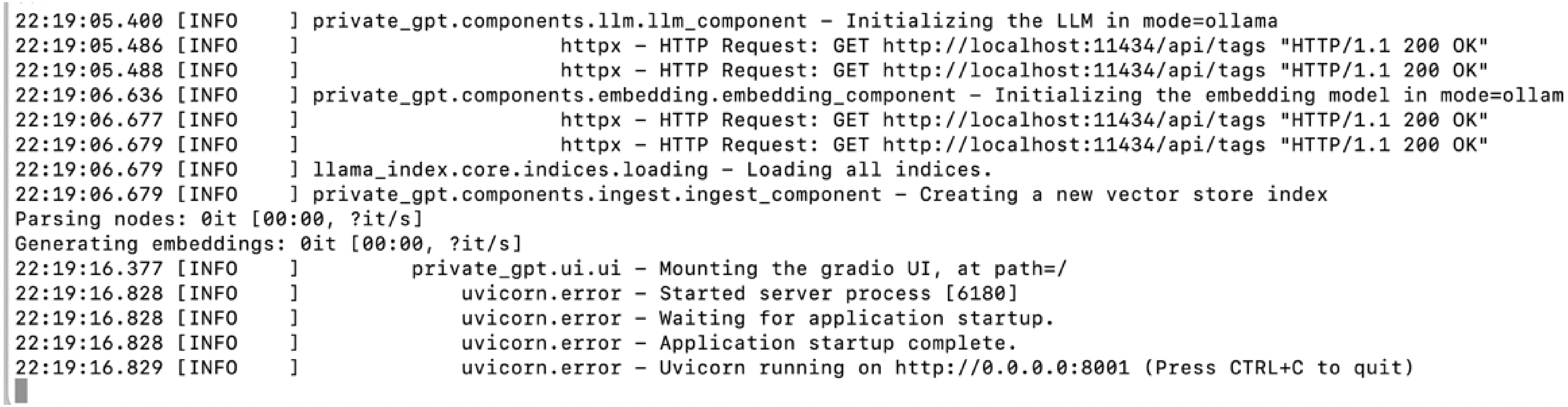
```
info@MacBook-Pro private-gpt % poetry install --extras "ui llms-ollama embeddings-ollama vector-stores-qdrant"
Creating virtualenv private-gpt-ghk2DPis-py3.11 in /Users/info/Library/Caches/pypoetry/virtualenvs
Installing dependencies from lock file

Package operations: 138 installs, 1 update, 0 removals

  - Installing pycparser (2.22)
  - Installing cffi (1.17.1)
  - Installing wrapt (1.16.0)
  - Installing certifi (2024.8.30)
  - Installing charset-normalizer (3.3.2)
  - Installing cryptography (3.4.8)
  - Installing deprecated (1.2.14)
```

Private GPT for Windows PC/Mac/Linux

Run PrivateGPT

- Make sure you have a working Ollama running locally before running the following command.

```
PGPT_PROFILES=ollama make run
```

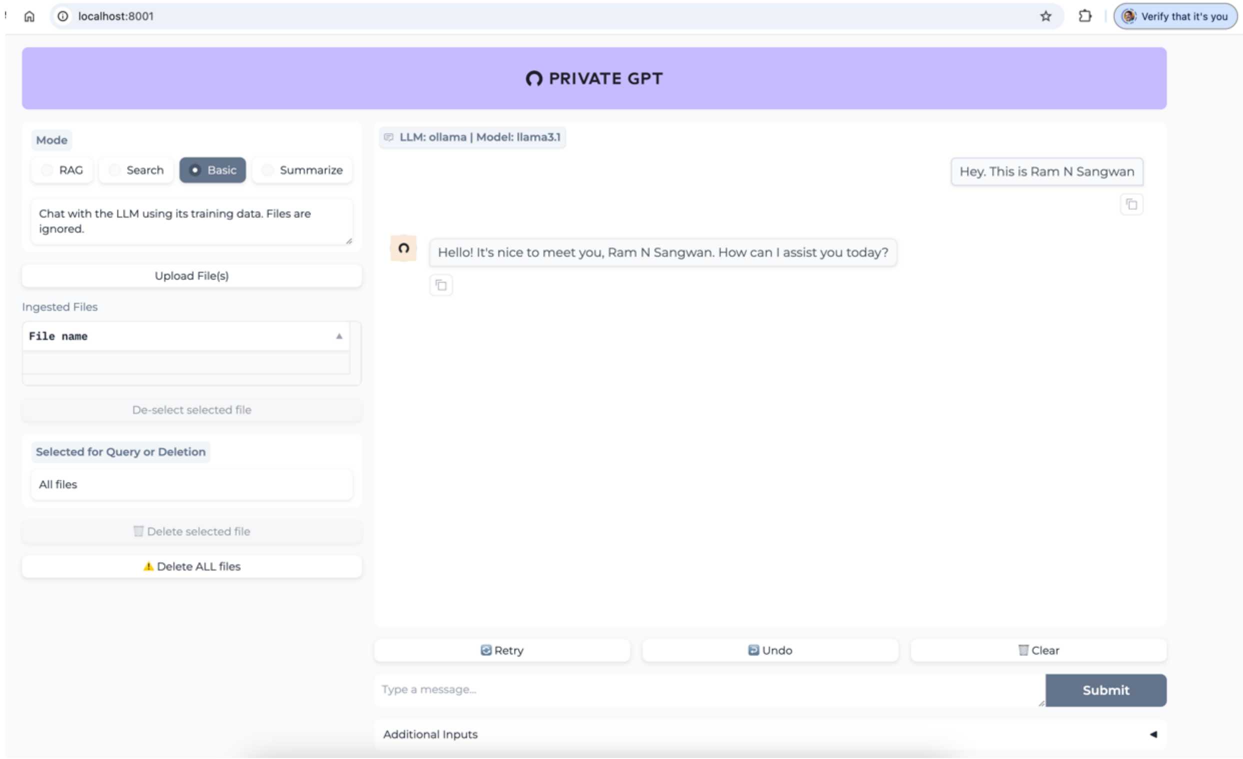


```
22:19:05.400 [INFO    ] private_gpt.components.llm.llm_component - Initializing the LLM in mode=ollama
22:19:05.486 [INFO    ]                     httpx - HTTP Request: GET http://localhost:11434/api/tags "HTTP/1.1 200 OK"
22:19:05.488 [INFO    ]                     httpx - HTTP Request: GET http://localhost:11434/api/tags "HTTP/1.1 200 OK"
22:19:06.636 [INFO    ] private_gpt.components.embedding.embedding_component - Initializing the embedding model in mode=ollam
22:19:06.677 [INFO    ]                     httpx - HTTP Request: GET http://localhost:11434/api/tags "HTTP/1.1 200 OK"
22:19:06.679 [INFO    ]                     httpx - HTTP Request: GET http://localhost:11434/api/tags "HTTP/1.1 200 OK"
22:19:06.679 [INFO    ] llama_index.core.indices.loading - Loading all indices.
22:19:06.679 [INFO    ] private_gpt.components.ingest.ingest_component - Creating a new vector store index
Parsing nodes: 0it [00:00, ?it/s]
Generating embeddings: 0it [00:00, ?it/s]
22:19:16.377 [INFO    ]         private_gpt.ui.ui - Mounting the gradio UI, at path=/
22:19:16.828 [INFO    ]             uvicorn.error - Started server process [6180]
22:19:16.828 [INFO    ]             uvicorn.error - Waiting for application startup.
22:19:16.828 [INFO    ]             uvicorn.error - Application startup complete.
22:19:16.829 [INFO    ]             uvicorn.error - Uvicorn running on http://0.0.0.0:8001 (Press CTRL+C to quit)
```

Private GPT for Windows PC/Mac/Linux

Access PrivateGPT

- PrivateGPT will use the already existing `settings-ollama.yaml`, which is already configured to use Ollama LLM and Embeddings, and Qdrant.
- Review it and adapt it to your needs.
- The UI will be available at http://localhost:8001



localhost:8001

PRIVATE GPT

Mode: ○ RAG ○ Search ● Basic ○ Summarize

Chat with the LLM using its training data. Files are ignored.

Ingested Files

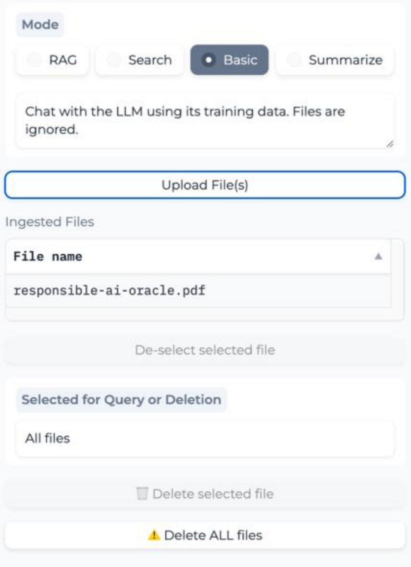
File name

Selected for Query or Deletion: All files

LLM: ollama | Model: llama3.1

Hey. This is Ram N Sangwan

Hello! It's nice to meet you, Ram N Sangwan. How can I assist you today?



Mode

○ RAG ○ Search ● Basic ○ Summarize

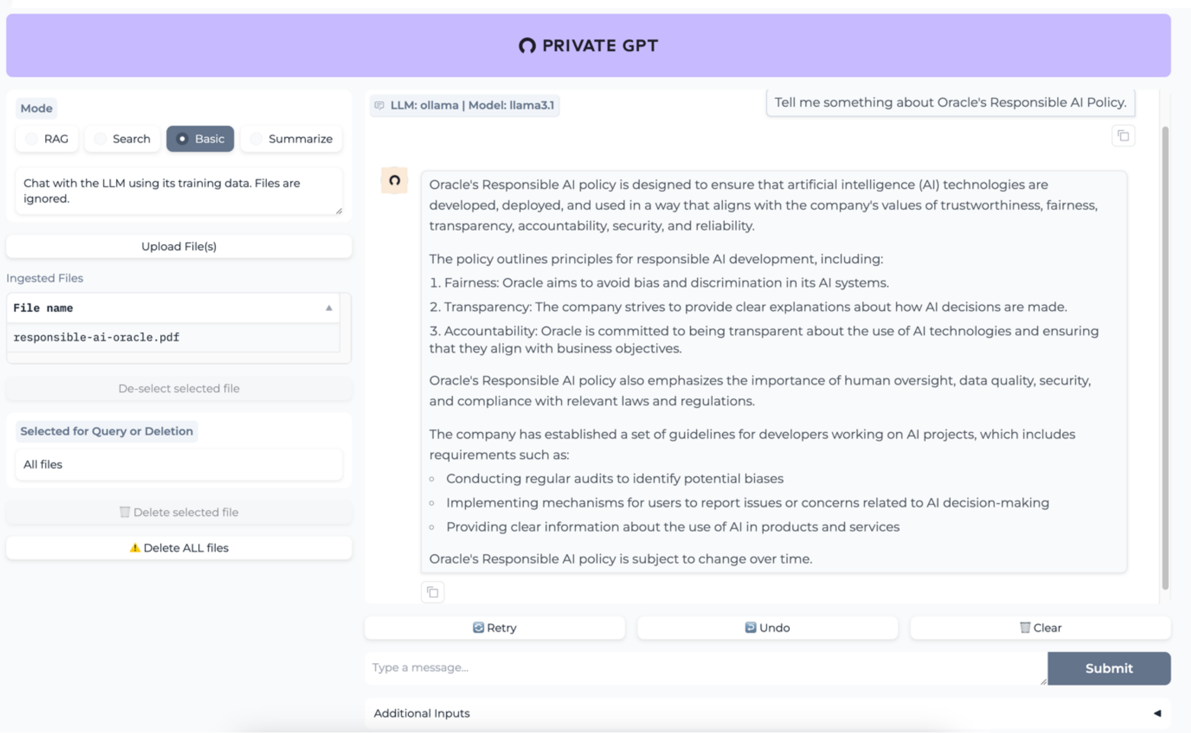
Chat with the LLM using its training data. Files are ignored.

Ingested Files

File name
responsible-ai-oracle.pdf

Selected for Query or Deletion

All files



Mode

○ RAG ○ Search ● Basic ○ Summarize

Chat with the LLM using its training data. Files are ignored.

Ingested Files

File name
responsible-ai-oracle.pdf

Selected for Query or Deletion

All files

LLM: ollama | Model: llama3.1

Tell me something about Oracle's Responsible AI Policy.

Oracle's Responsible AI policy is designed to ensure that artificial intelligence (AI) technologies are developed, deployed, and used in a way that aligns with the company's values of trustworthiness, fairness, transparency, accountability, security, and reliability.

The policy outlines principles for responsible AI development, including:

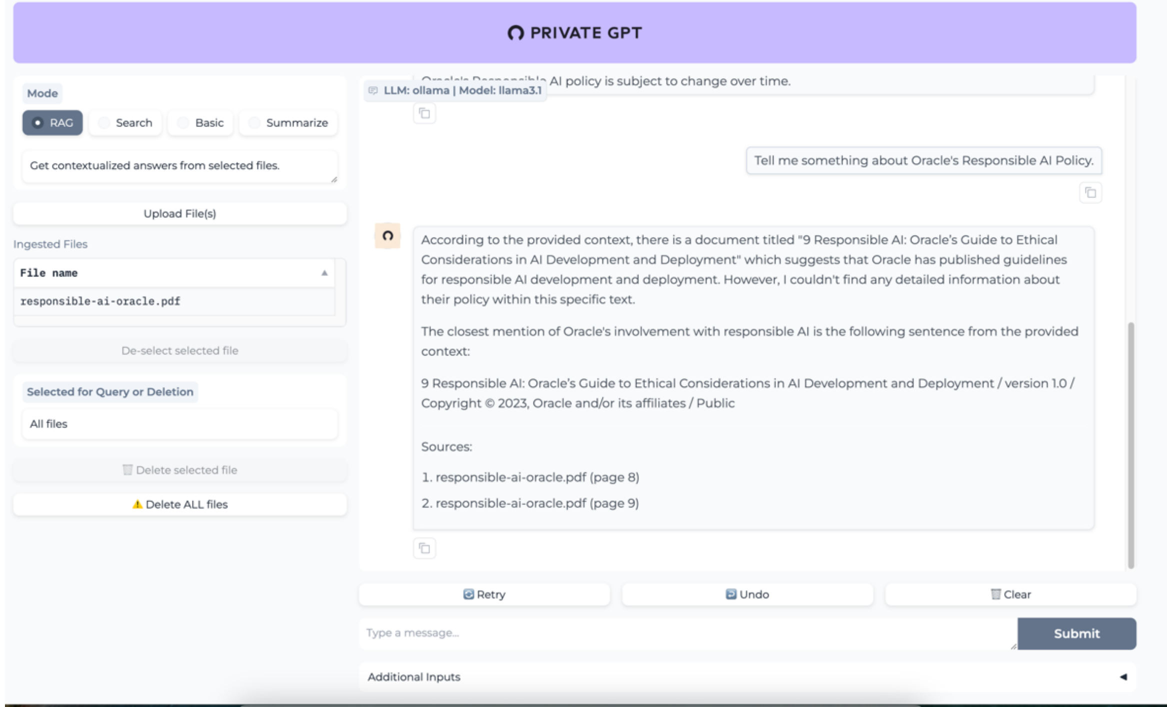
1. Fairness: Oracle aims to avoid bias and discrimination in its AI systems.
2. Transparency: The company strives to provide clear explanations about how AI decisions are made.
3. Accountability: Oracle is committed to being transparent about the use of AI technologies and ensuring that they align with business objectives.

Oracle's Responsible AI policy also emphasizes the importance of human oversight, data quality, security, and compliance with relevant laws and regulations.

The company has established a set of guidelines for developers working on AI projects, which includes requirements such as:

- Conducting regular audits to identify potential biases
- Implementing mechanisms for users to report issues or concerns related to AI decision-making
- Providing clear information about the use of AI in products and services

Oracle's Responsible AI policy is subject to change over time.



Oracle's Responsible AI policy is subject to change over time.

LLM: ollama | Model: llama3.1

Mode

● RAG ○ Search ○ Basic ○ Summarize

Get contextualized answers from selected files.

Ingested Files

File name
responsible-ai-oracle.pdf

Selected for Query or Deletion

All files

Tell me something about Oracle's Responsible AI Policy.

According to the provided context, there is a document titled "9 Responsible AI: Oracle's Guide to Ethical Considerations in AI Development and Deployment" which suggests that Oracle has published guidelines for responsible AI development and deployment. However, I couldn't find any detailed information about their policy within this specific text.

The closest mention of Oracle's involvement with responsible AI is the following sentence from the provided context:

9 Responsible AI: Oracle's Guide to Ethical Considerations in AI Development and Deployment / version 1.0 / Copyright © 2023, Oracle and/or its affiliates / Public

Sources:

1. responsible-ai-oracle.pdf (page 8)
2. responsible-ai-oracle.pdf (page 9)

Thank You