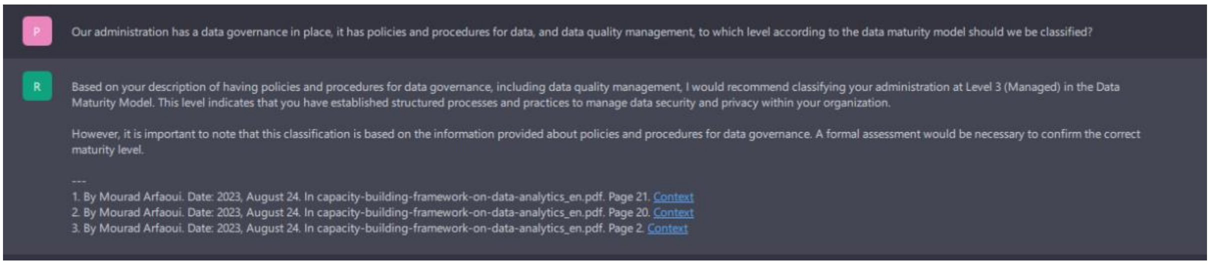
How to create a personal, secure Chatbot assistant on Customs Data

Are you interested in evaluating your organization's data analytics maturity based on the WCO's proposed model, and keen to enhance your capabilities? As demonstrated, the chatbot trained on WCO data can offer these insights and more, tailored specifically to your domain and requirements. For instance, it can conduct diagnostic assessments for you based on your specific queries.



Our administration has a data governance in place, it has policies and procedures for data, and data quality management, to which level according to the data maturity model should we be classified?

Based on your description of having policies and procedures for data governance, including data quality management, I would recommend classifying your administration at Level 3 (Managed) in the Data Maturity Model. This level indicates that you have established structured processes and practices to manage data security and privacy within your organization.

However, it is important to note that this classification is based on the information provided about policies and procedures for data governance. A formal assessment would be necessary to confirm the correct maturity level.

---

1. By Mourad Arfaoui. Date: 2023, August 24. In capacity-building-framework-on-data-analytics_en.pdf. Page 21. Context
2. By Mourad Arfaoui. Date: 2023, August 24. In capacity-building-framework-on-data-analytics_en.pdf. Page 20. Context
3. By Mourad Arfaoui. Date: 2023, August 24. In capacity-building-framework-on-data-analytics_en.pdf. Page 2. Context

ChatGPT has garnered significant attention for its wide range of functionalities, including answering questions, writing emails or essays, and serving as a personal or coding assistant. The tool rapidly gained popularity, attracting over a million users within a week of its launch, and has been used by over 100 million people globally. It has also amassed over 9 billion all-time page visits. The World Customs Organization (WCO) is interested in exploring the use of ChatGPT and other Large Language Models in the Customs sector, provided that data security and privacy can be ensured.

OpenAI has stated on their website that ChatGPT users now have the option to disable chat history, which limits the data retained for training the models. When this feature is active, conversations are kept for only 30 days and are reviewed solely for abuse monitoring before permanent deletion. However, there are reports suggesting that ChatGPT still collects a substantial amount of user data, including conversations, location details, and IP addresses, raising concerns over data security and privacy.

Given these concerns, sharing confidential information from the WCO or Customs Organizations with ChatGPT poses a notable risk. Does this mean that Customs agencies should avoid using such emerging technologies for domain-specific tasks? Not necessarily. The rise of ChatGPT has also catalyzed the development of many open-source alternatives. These open-source Large Language Models and chatbots offer the option to keep user data private and allow organizations to integrate pre-trained models with their internal resources, thereby balancing innovation with security.

Before delving into the methodologies, let's first clarify some of the technical terminology commonly encountered in the realm of chatbots.

Large Language Models (LLMs): LLMs are sophisticated machine learning algorithms capable of understanding and generating text that closely mimics human language. Trained on extensive datasets often sourced from the internet, these models analyze linguistic patterns and structures to carry out tasks like summarization, translation, text prediction, and content generation.

GPT4All: GPT4All is an open-source platform designed to facilitate the integration of Large Language Models into various applications without the need for costly subscriptions to specialized hardware or platforms. It enables individuals and organizations to train, customize, and deploy highly capable language models on conventional computing hardware. Its ultimate aim is to offer an instruction-optimized, assistant-style language model that is accessible, extendable, and free for both individual users and enterprises.

LangChain: LangChain is a Python-based library that serves as a comprehensive toolkit for building applications that leverage language models. Whether you aim to develop a chatbot, translation service, or automated content generator, LangChain provides the essential components to simplify the development process.

API: An Application Programming Interface (API) provides a list of things that a software program can do, like sending an email, looking up a location, or saving a file. When you, or another software, make a specific request from that list, the system carries out the task and returns the result. This allows different software programs to work together easily.

In the following, two methods with distinct advantages and drawbacks on open-source tools application for customized safe to use Chatbots will be introduced.

1. Using Python libraries to manually use open Source LLM

The first approach involves using open-source libraries, which provide convenient access to specialized Python libraries.

Python libraries are collections of pre-written code that streamline programming tasks. They enable users to perform complex operations without having to write the code from the ground up, thereby increasing efficiency.

Key libraries employed in this scenario include LangChain, GPT4All, PyPDF, and ChromaDB.

LangChain allows for easy integration with an open-source Large Language Model (LLM) named GPT4All. This model can be downloaded and run locally on your own computer, eliminating the need to connect to external servers via an API, as is required with paid models like GPT-4. Given the data security concerns mentioned earlier, running the model locally guarantees that data will not be exposed to third parties.

For those interested in the technical details, the source code is available for review. However, for the purpose of this discussion, we will focus solely on the outcomes of the implementation.

In our case, we used a publicly available PDF of the Data Analytics Capacity Building Framework to train the LLM. The model was trained to understand the contents of the document. After the training phase, we tasked the model with summarizing the document.

The model produced an accurate summary, although it is worth noting that the execution time was relatively slow.
The reason for this can be attributed to multiple factors, here are some examples:

- **Computation Power Limitations:** The code is executed on the free tier of Google Colaboratory, in alignment with the goal of making the project accessible to a wide range of users. It's important to note that this environment provides a limited 13GB of RAM. Large Language Models typically demand more computational power, which may contribute to delays in response time.
- **Document Size Limitations:** The size of the PDF used for training can also impact processing time. In this case, the Data Analytics Capacity Building Framework document spans 66 pages. Parsing and training on such a document using the PyPDF library may result in extended delays.
- **Library Efficiency:** The performance can also be influenced by the efficiency of the LangChain and GPT4All libraries themselves. These open-source solutions may not be as finely tuned for specific tasks as their commercial counterparts, possibly affecting overall performance.

Alongside the performance issue we can find that if we want to increase performance by utilizing Open AI API, we would be facing the security issues.

Still, there are advantages to be had by using this model, if proper optimization can be had:

As we are writing our own source code, there is the ability to customize the code to own needs, and flexibility to fine-tune parameters to improve precision and performance. Also we can find lots of documentation online on the libraries used and alternate methods of implementation and optimization.

In addition to performance limitations, shifting to OpenAI's API to improve speed would introduce data security concerns.

However, there are distinct advantages to using this approach if proper optimizations can be implemented:

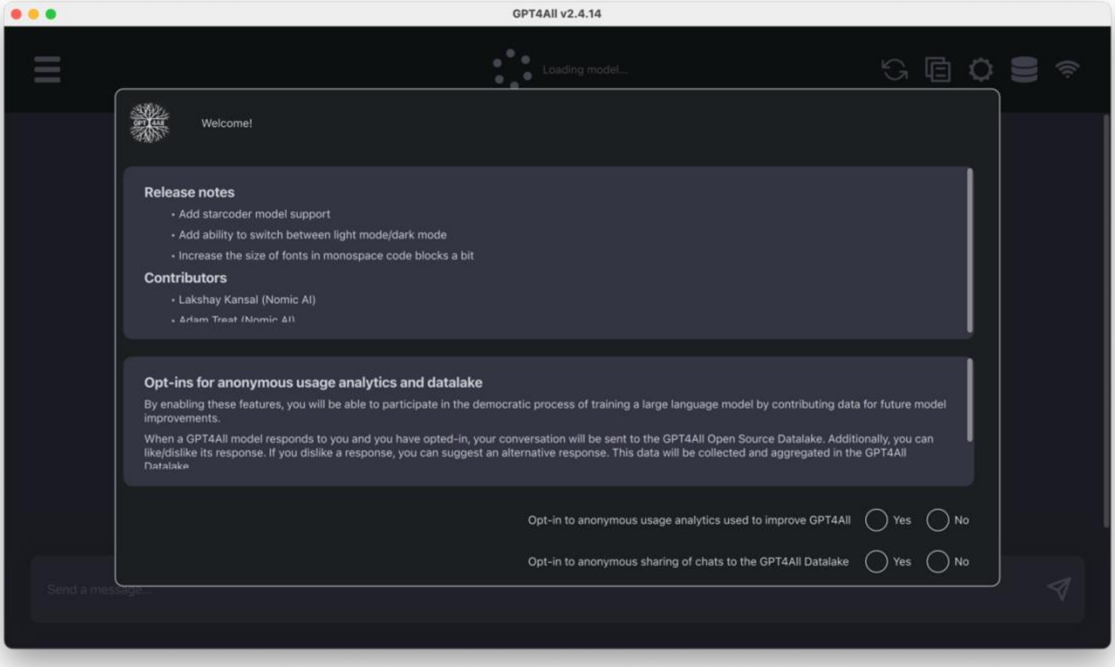
By writing our own source code, we gain the flexibility to tailor the code to meet specific requirements and fine-tune various parameters to enhance both precision and performance. Furthermore, an abundance of online documentation is available for the libraries in use, offering alternative strategies for implementation and optimization.

2. Using GPT4All Application

We already introduced the GPT4All LLM, but there is also a tool with the same name, that is a free-to-use, locally running, privacy-aware chatbot. No GPU or internet required. It can be installed on windows, MAC and linux. Much like the previous approach this can be downloaded and used locally to bypass any security and privacy concerns. Another Advantage is that it is a No-Code solution, meaning everyone can just follow some simple steps to make use of this tool:

Here is a short tutorial on how to use this tool to train models on your data and ask questions about your files:

1. Install GPT4All from their website for your Operating System.
2. If you prefer not to share any data, make sure to select 'No' for both opt-in options that appear on the pop-up screen.



GPT4All v2.4.14

Welcome!

Release notes

- Add starcoder model support
- Add ability to switch between light mode/dark mode
- Increase the size of fonts in monospace code blocks a bit

Contributors

- Lakshay Kansal (Nomic AI)
- Adam Treat (Nomic AI)

Opt-ins for anonymous usage analytics and datalake

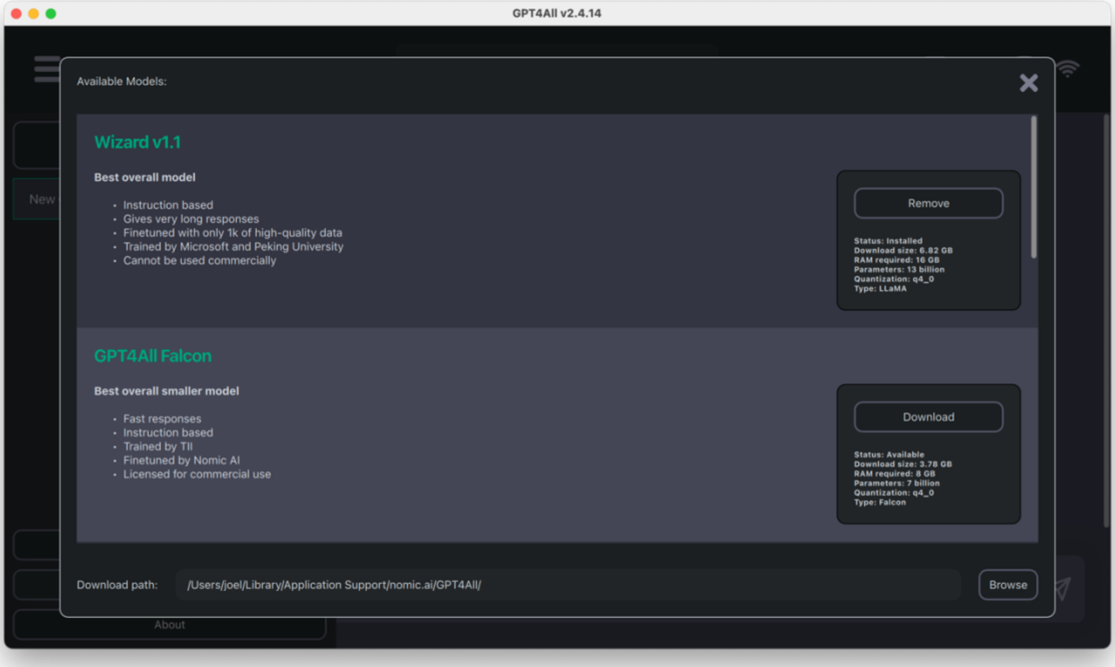
By enabling these features, you will be able to participate in the democratic process of training a large language model by contributing data for future model improvements.

When a GPT4All model responds to you and you have opted-in, your conversation will be sent to the GPT4All Open Source Datalake. Additionally, you can like/dislike its response. If you dislike a response, you can suggest an alternative response. This data will be collected and aggregated in the GPT4All Datalake

Opt-in to anonymous usage analytics used to improve GPT4All ○ Yes ○ No

Opt-in to anonymous sharing of chats to the GPT4All Datalake ○ Yes ○ No

3. Next, you'll need to download a Large Language Model (LLM) to use. Free options include the Wizard or the GPT4All Falcon Model. You can also connect to OpenAI's GPT Model, though this would require data sharing via the API. Each model comes with detailed descriptions, so take a moment to read through these to find the model that best suits your needs. If you have ample storage space on your computer, I recommend downloading Wizard v1.1 by clicking the 'Download' button.



Available Models:

Wizard v1.1

Best overall model

- Instruction based
- Gives very long responses
- Finetuned with only 1k of high-quality data
- Trained by Microsoft and Peking University
- Cannot be used commercially

Status: Installed
Download size: 6.82 GB
RAM required: 16 GB
Parameters: 13 billion
Quantization: q4_0
Type: LLaMA

GPT4All Falcon

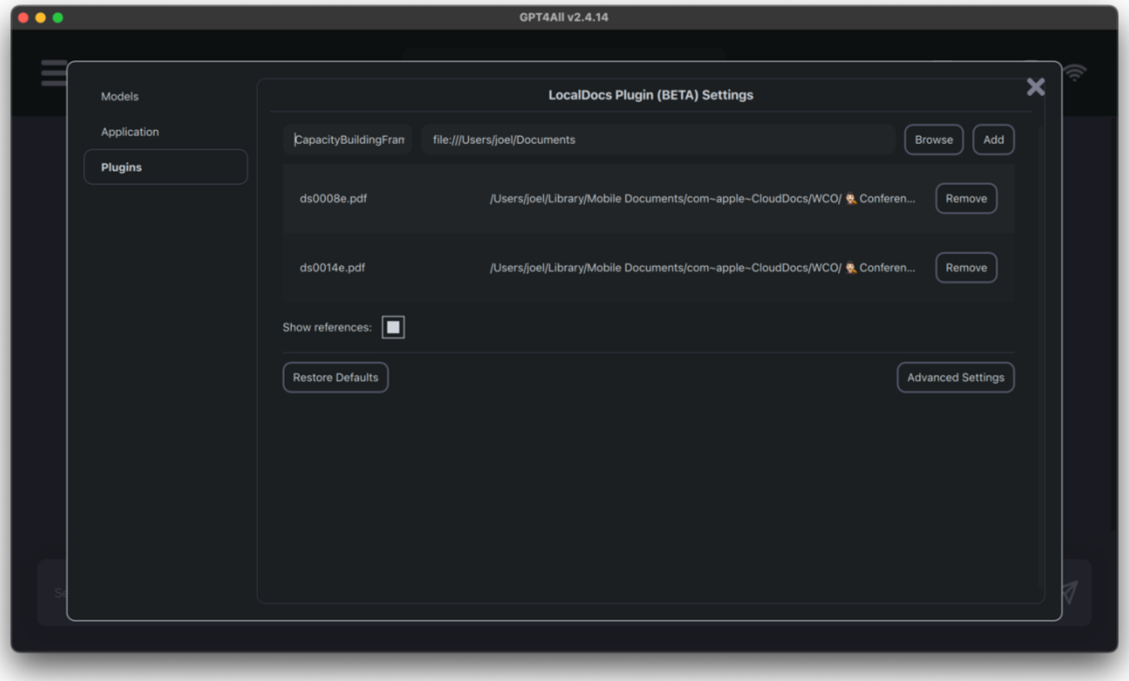
Best overall smaller model

- Fast responses
- Instruction based
- Trained by TII
- Finetuned by Nomic AI
- Licensed for commercial use

Status: Available
Download size: 3.78 GB
RAM required: 8 GB
Parameters: 7 billion
Quantization: q4_0
Type: Falcon

Download path: /Users/joel/Library/Application Support/nomic.ai/GPT4All/

4. Next, click on the settings icon and navigate to the 'Plug-ins' section. Here, you can browse and add various types of files, including PDFs, Word documents, and text files, among others. The process is slightly intricate: first, click 'Browse' to navigate to the folder where your desired document is stored. Then, enter the name of the document in the 'Collection name' field.



LocalDocs Plugin (BETA) Settings

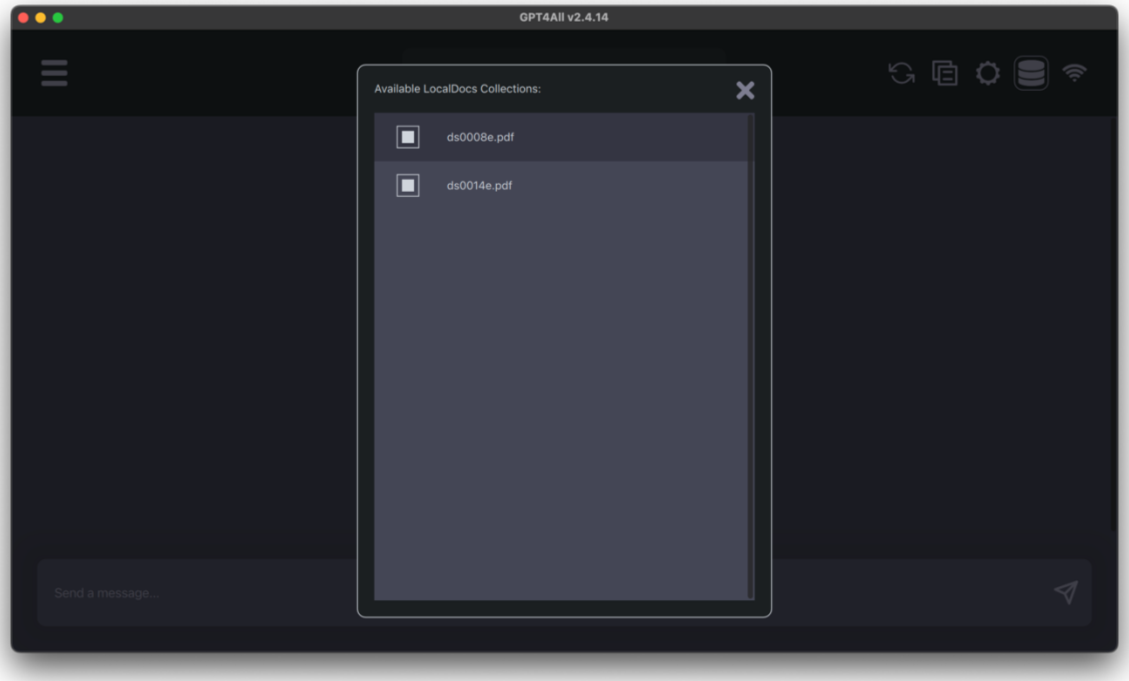
CapacityBuildingFran | file:///Users/joel/Documents

ds0008e.pdf /Users/joel/Library/Mobile Documents/com~apple~CloudDocs/WCO/ Conferen...

ds0014e.pdf /Users/joel/Library/Mobile Documents/com~apple~CloudDocs/WCO/ Conferen...

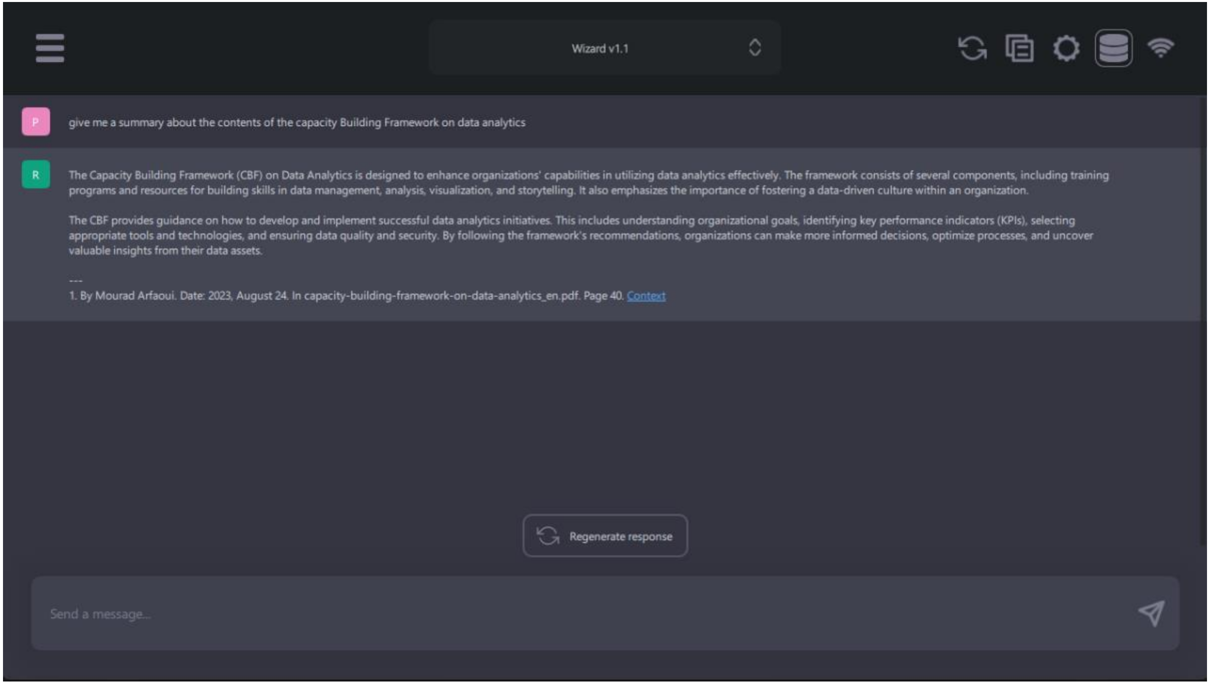
Show references: □

5. After completing these steps, you can close the pop-up window and click on the database icon. This will display the documents you've added, allowing you to select the specific files you wish to use.



Now you're all set! To start asking questions, simply close the pop-up window and type your query into the message box. You're ready to go!

This tool you will find to be very helpful, here are some examples of how the tool helped me with assessing the maturity of my organization as well as giving me recommendations for improving maturity.



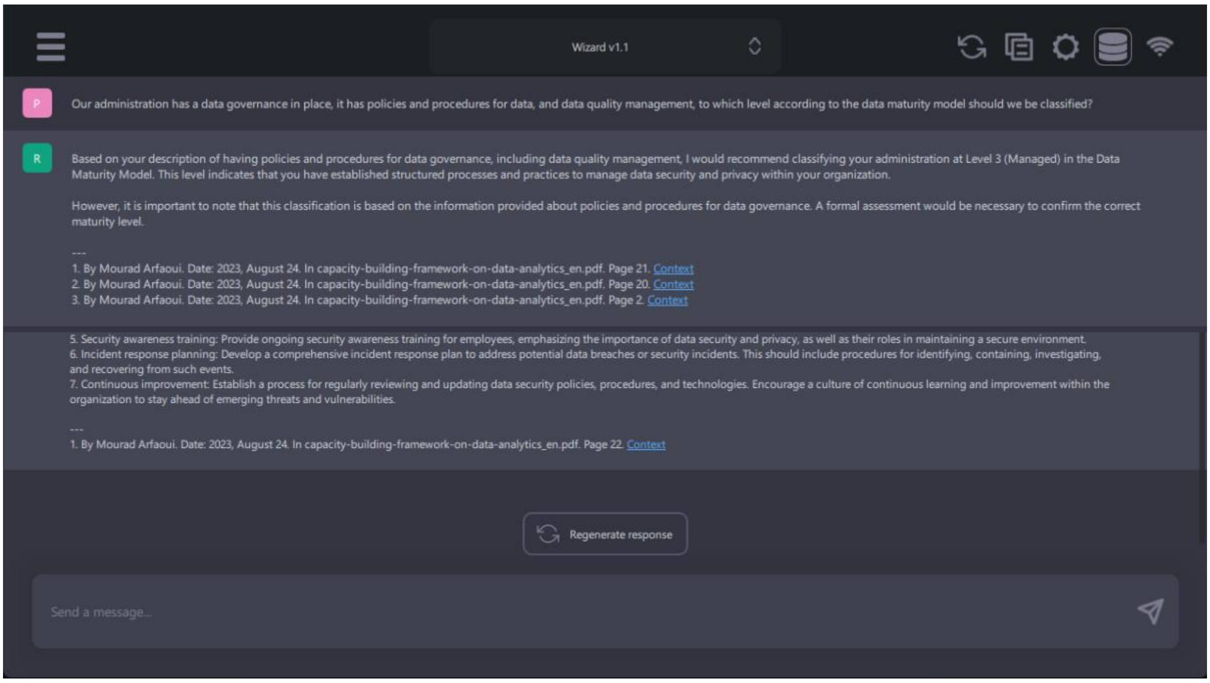
give me a summary about the contents of the capacity Building Framework on data analytics

The Capacity Building Framework (CBF) on Data Analytics is designed to enhance organizations' capabilities in utilizing data analytics effectively. The framework consists of several components, including training programs and resources for building skills in data management, analysis, visualization, and storytelling. It also emphasizes the importance of fostering a data-driven culture within an organization.

The CBF provides guidance on how to develop and implement successful data analytics initiatives. This includes understanding organizational goals, identifying key performance indicators (KPIs), selecting appropriate tools and technologies, and ensuring data quality and security. By following the framework's recommendations, organizations can make more informed decisions, optimize processes, and uncover valuable insights from their data assets.

---
1. By Mourad Arfaoui. Date: 2023, August 24. In capacity-building-framework-on-data-analytics_en.pdf. Page 40. Context

Here it gives me a summary on the Capacity Building Framework Document. Here it only takes a few seconds to produce a response.



P

Our administration has a data governance in place, it has policies and procedures for data, and data quality management, to which level according to the data maturity model should we be classified?

R

Based on your description of having policies and procedures for data governance, including data quality management, I would recommend classifying your administration at Level 3 (Managed) in the Data Maturity Model. This level indicates that you have established structured processes and practices to manage data security and privacy within your organization.

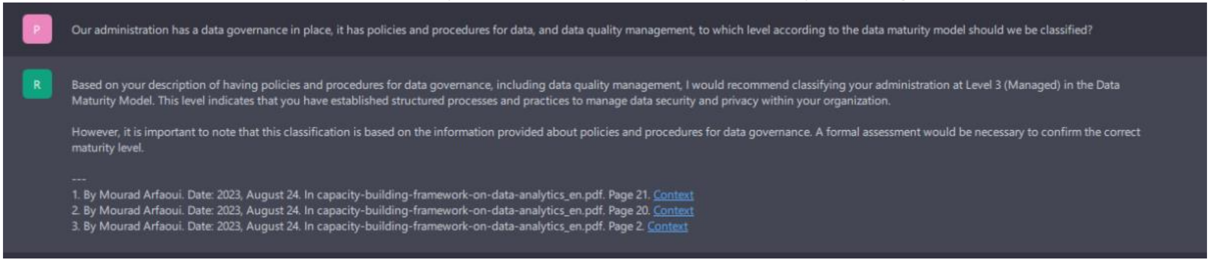
However, it is important to note that this classification is based on the information provided about policies and procedures for data governance. A formal assessment would be necessary to confirm the correct maturity level.

---
1. By Mourad Arfaoui. Date: 2023, August 24. In capacity-building-framework-on-data-analytics_en.pdf. Page 21. Context
2. By Mourad Arfaoui. Date: 2023, August 24. In capacity-building-framework-on-data-analytics_en.pdf. Page 20. Context
3. By Mourad Arfaoui. Date: 2023, August 24. In capacity-building-framework-on-data-analytics_en.pdf. Page 2. Context

5. Security awareness training: Provide ongoing security awareness training for employees, emphasizing the importance of data security and privacy, as well as their roles in maintaining a secure environment.
6. Incident response planning: Develop a comprehensive incident response plan to address potential data breaches or security incidents. This should include procedures for identifying, containing, investigating, and recovering from such events.
7. Continuous improvement: Establish a process for regularly reviewing and updating data security policies, procedures, and technologies. Encourage a culture of continuous learning and improvement within the organization to stay ahead of emerging threats and vulnerabilities.

---
1. By Mourad Arfaoui. Date: 2023, August 24. In capacity-building-framework-on-data-analytics_en.pdf. Page 22. Context

In this section, it provides specific recommendations for advancing from the second to the third maturity level.
In this query, we outline the scope of our organization's efforts in Data Analytics Strategy and the tool identifies the maturity level to which we currently belong.


P

Our administration has a data governance in place, it has policies and procedures for data, and data quality management, to which level according to the data maturity model should we be classified?

R

Based on your description of having policies and procedures for data governance, including data quality management, I would recommend classifying your administration at Level 3 (Managed) in the Data Maturity Model. This level indicates that you have established structured processes and practices to manage data security and privacy within your organization.

However, it is important to note that this classification is based on the information provided about policies and procedures for data governance. A formal assessment would be necessary to confirm the correct maturity level.

---
1. By Mourad Arfaoui. Date: 2023, August 24. In capacity-building-framework-on-data-analytics_en.pdf. Page 21. Context
2. By Mourad Arfaoui. Date: 2023, August 24. In capacity-building-framework-on-data-analytics_en.pdf. Page 20. Context
3. By Mourad Arfaoui. Date: 2023, August 24. In capacity-building-framework-on-data-analytics_en.pdf. Page 2. Context

While this tool has its limitations—namely, it doesn't allow users to fine-tune its implementation and restricts them to its built-in features—it's worth noting that this No-Code solution offers the most user-friendly and effective way for Customs Officers to experiment and gain practical benefits for their day-to-day tasks.

In conclusion, it's crucial to highlight that the landscape of Large Language Models and chatbots is continuously evolving. During the course of our study, new models and methods were being developed on a weekly basis. Given this rapid pace of innovation, the insights presented here could become outdated in just a few months. Therefore, it's advisable for those interested to stay updated on the latest trends in the field of LLMs and chatbots.