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

(by Joel Choi, WCO Data Expert, Capacity Building Directorate)

Are you interested in evaluating your organization's data analytics maturity based on the WCO's proposed model, and keen to enhance your capabilities, but lack the expertise to do so? As demonstrated in the picture, 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.

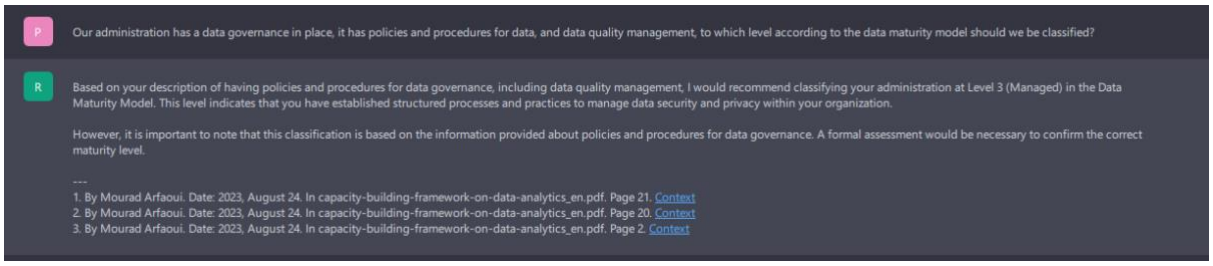


Figure 1: Chatbot recommending Level of Analytics Maturity as proposed by the WCO Maturity Model, based on organizations capabilities

Rapidly gaining traction since its launch, **ChatGPT** has captivated a global user base with its multifaceted capabilities—ranging from answering queries to drafting emails and essays, as well as functioning as a personal or coding assistant. Within just one week of its debut, the platform attracted over a million users, ultimately serving more than 100 million people worldwide and accumulating over 9 billion total page visits. These impressive metrics have piqued the interest of the World Customs Organization (WCO), which is exploring the potential integration of ChatGPT and other Large Language Models into Customs operations, conditional on meeting data security and privacy standards.

While OpenAI has addressed some **privacy concerns** by allowing users to disable chat history—a move that restricts data retention to a 30-day window for abuse monitoring—doubts still linger. Several reports indicate that despite these measures, ChatGPT continues to collect a broad spectrum of user data, including conversations, geolocation, and IP addresses. This presents a significant security risk for organizations like the WCO that handle confidential information.

However, these challenges don't necessarily mean Customs agencies should eschew emerging technologies like ChatGPT for domain-specific tasks. The meteoric rise of ChatGPT has simultaneously fueled the growth of **open-source alternatives**. These platforms empower organizations to safeguard user data while leveraging pre-trained models, thus striking a crucial balance between innovation and security. Before delving into the methodologies, let's first clarify some of the technical terminology commonly encountered in the realm of chatbots.

Before moving to the concrete open source methods used, we will introduce some key technical terms used a lot in the area of Chatbots and Large Language Models.

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. Here you can find 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.

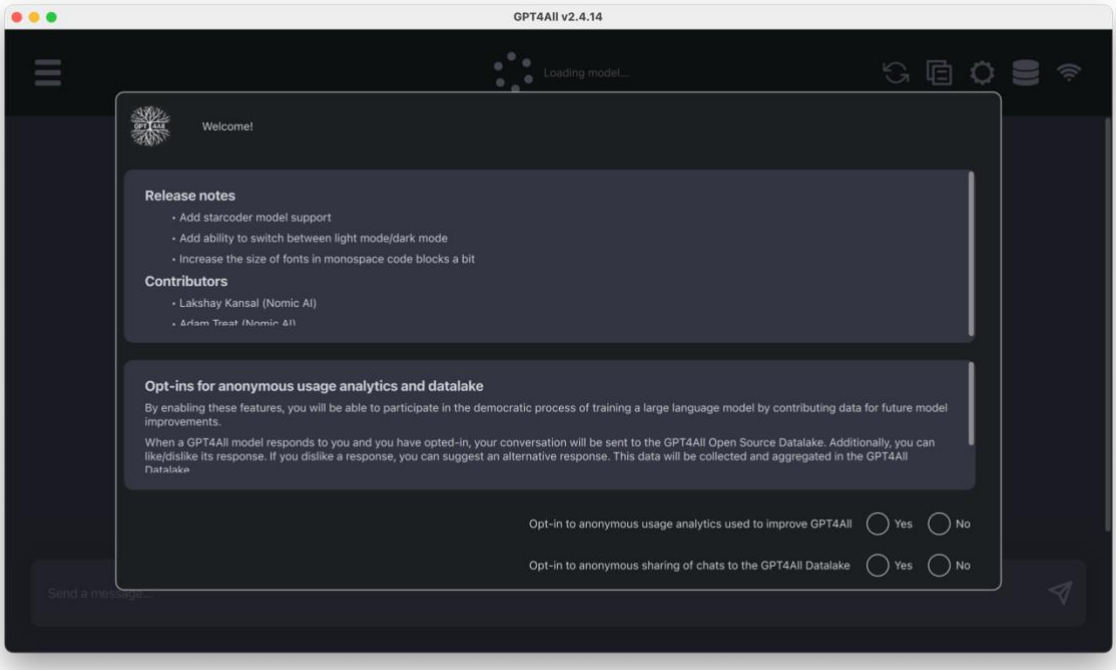
2. Using GPT4All Application

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:

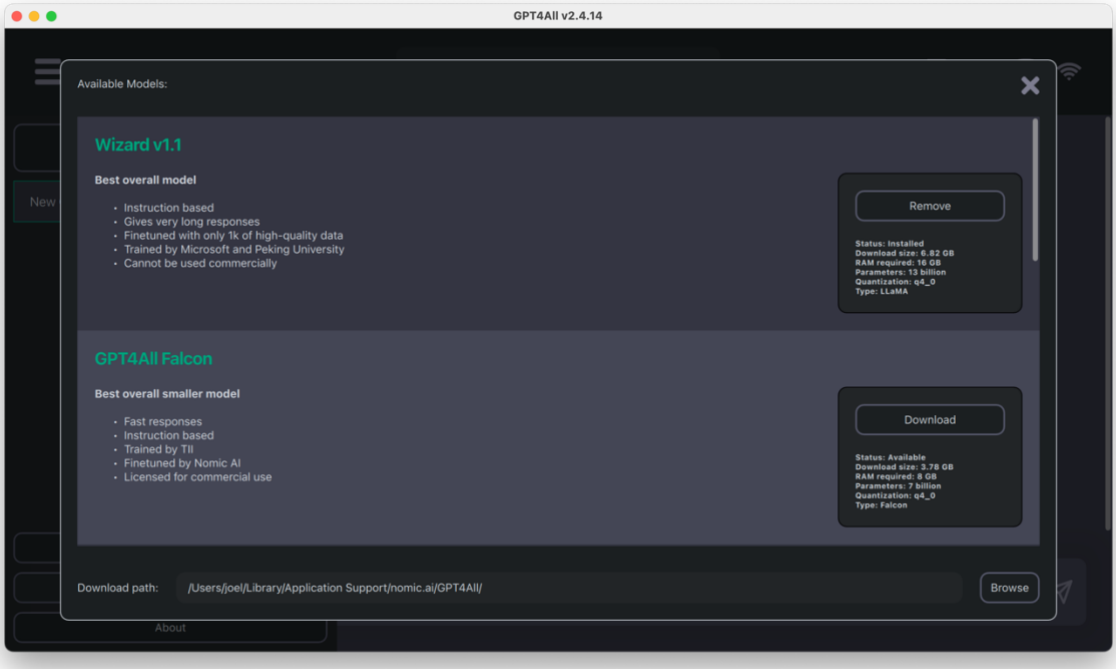
Here is a short tutorial on how to use this tool to train models on your data and ask questions about your files:

1. Install GPT4All from their website for your Operating System.

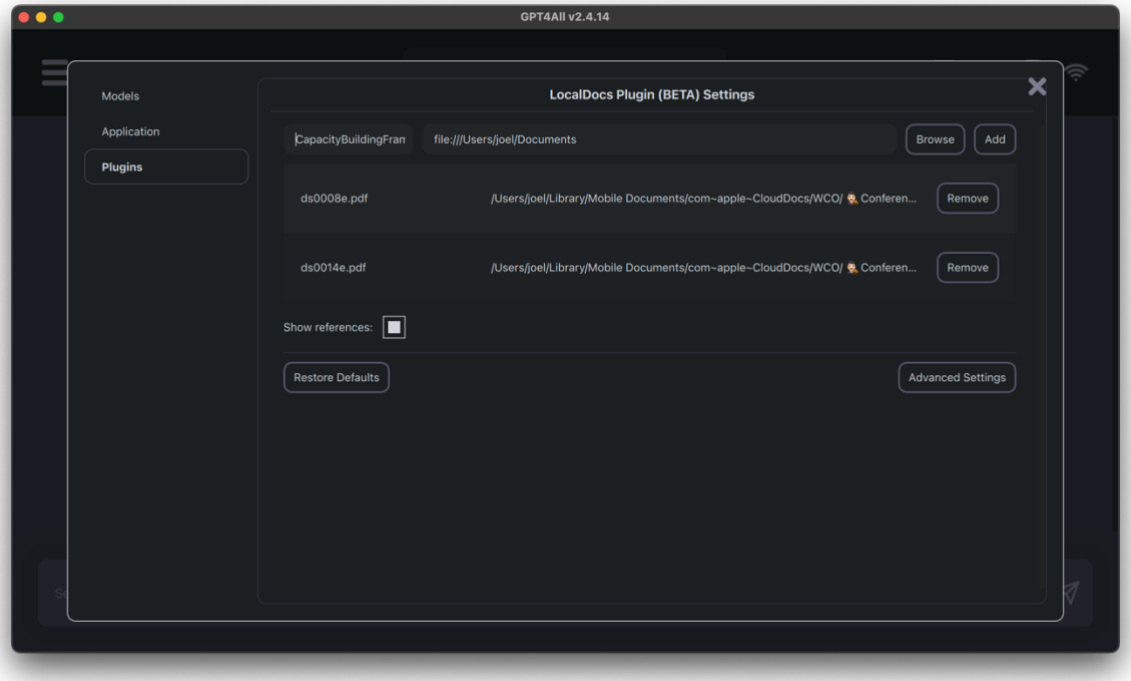
2. If you prefer not to share any data, make sure to select '**No**' for both opt-in options that appear on the pop-up screen. In this case we are actively trying to preserve security and privacy, so not sharing the data would be recommendable.



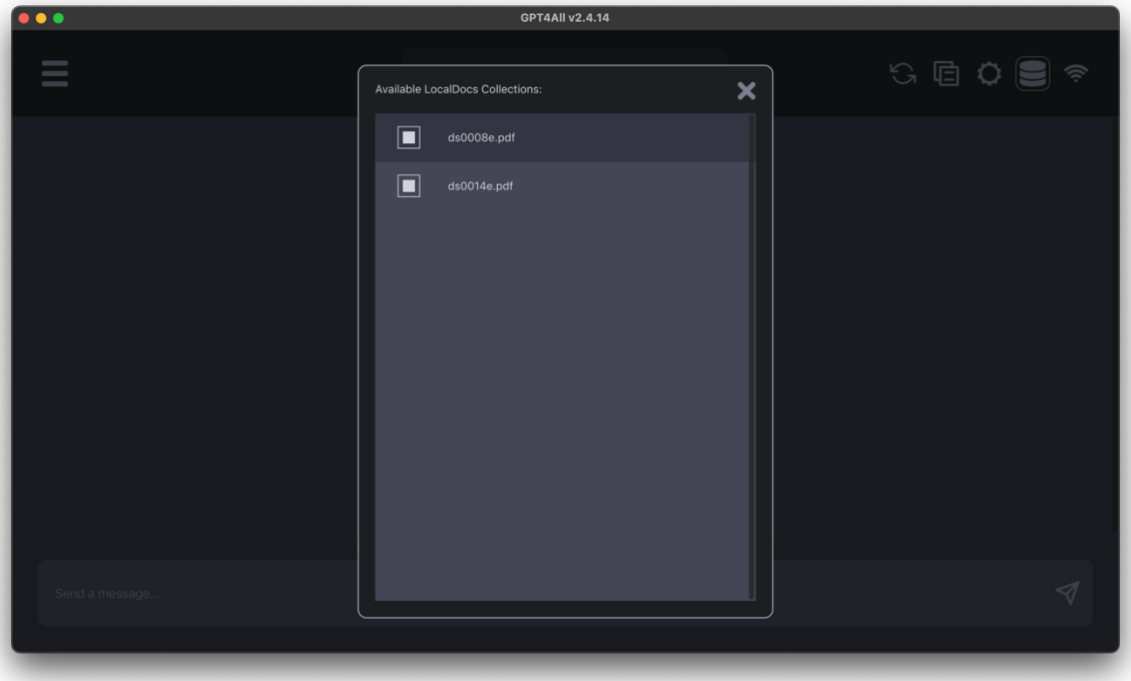
3. Next, you'll need to download a Large Language Model (LLM) to use. Free options include the Wizard or the GPT4All Falcon Model. You can also connect to OpenAI's GPT Model, though this would require data sharing via the API. Each model comes with detailed descriptions, so take a moment to read through 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.



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.



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.



You're all set to begin! To start asking questions, simply close this pop-up and enter your query in the message box. You're now ready to explore!

This tool offers a wealth of features that can significantly aid you. For instance, it has helped me assess my organization's maturity level and has provided targeted recommendations for enhancement. Unlike ChatGPT, this tool also includes the added benefit of providing precise citations for the sources of its information.

Examples of Using the GPT4All Chatbot on Customs Data Analytics Maturity

Here we offer sample use-cases illustrating how the chatbot can be employed for diagnostics and maturity level assessments. However, the potential applications within the realm of Customs work are bound only by your creativity. You can harness the full capabilities of the chatbot by inputting the relevant documents.

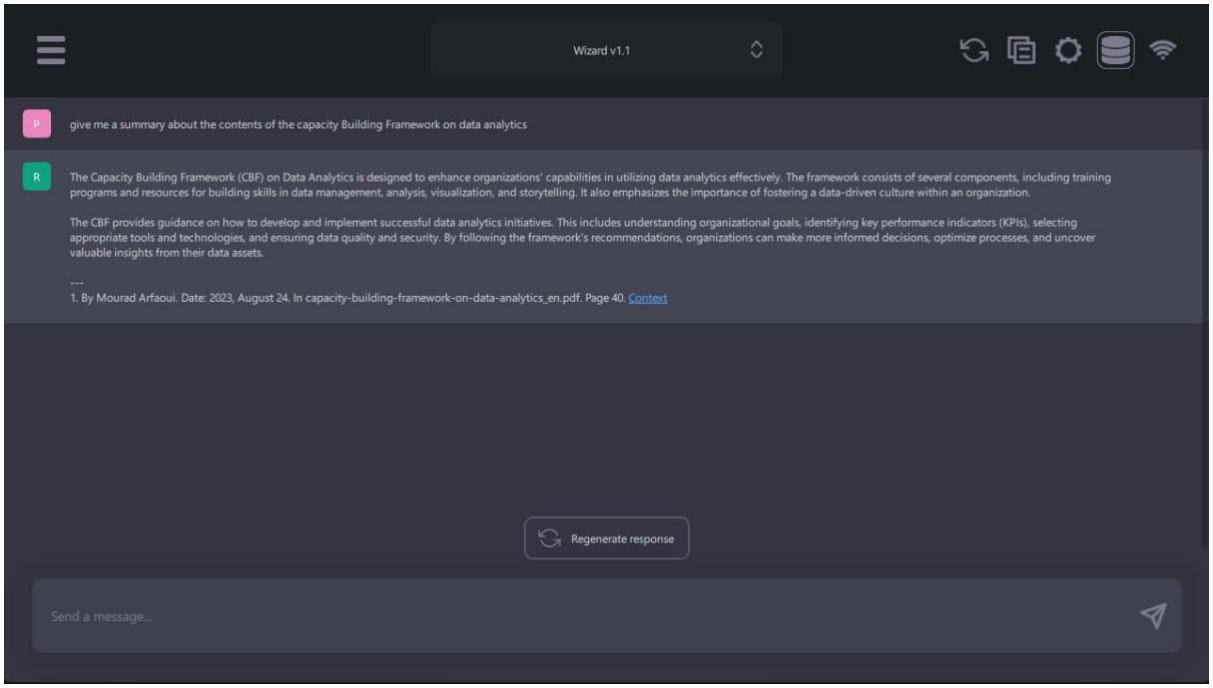


Figure 2: Chatbot gives a summary on the Capacity Building Framework Document in only a few seconds

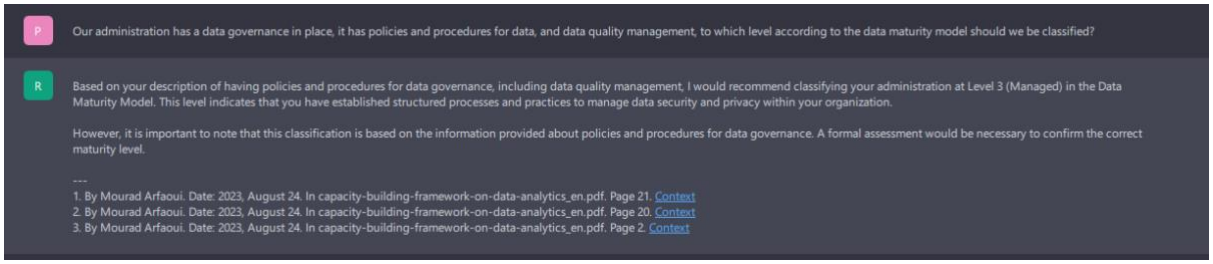


Figure 3: In this query, we outline the scope of our organization's efforts in Data Analytics Strategy and the tool identifies the maturity level to which we currently belong.

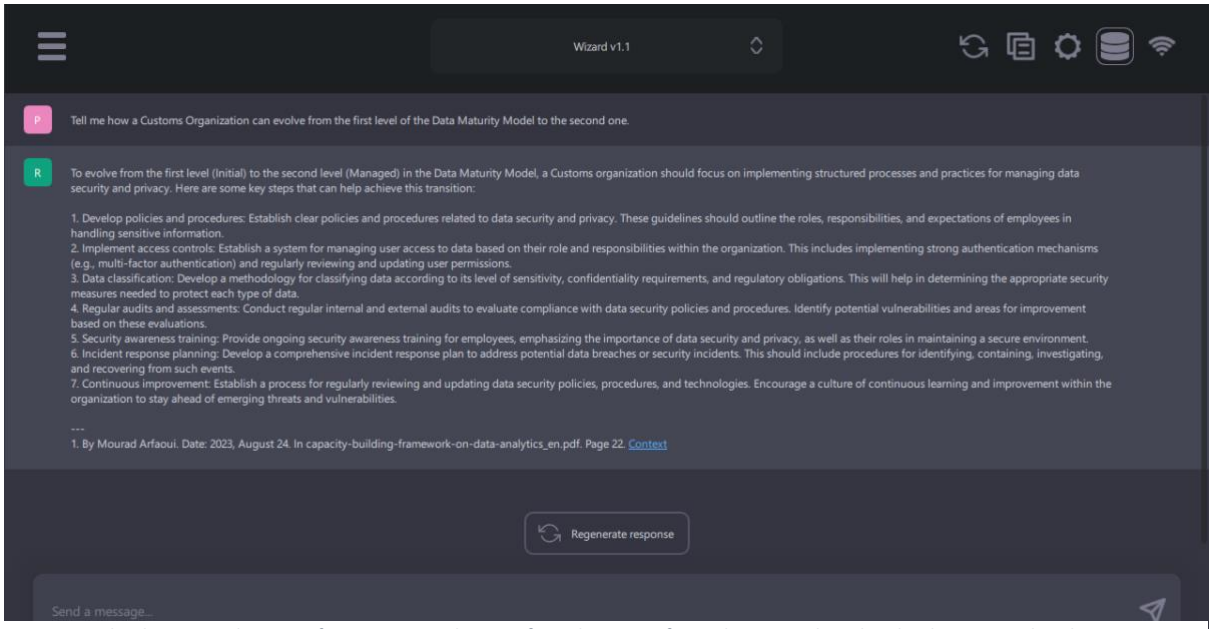


Figure 4: Chatbot provides specific recommendations for advancing from the second to the third maturity level.

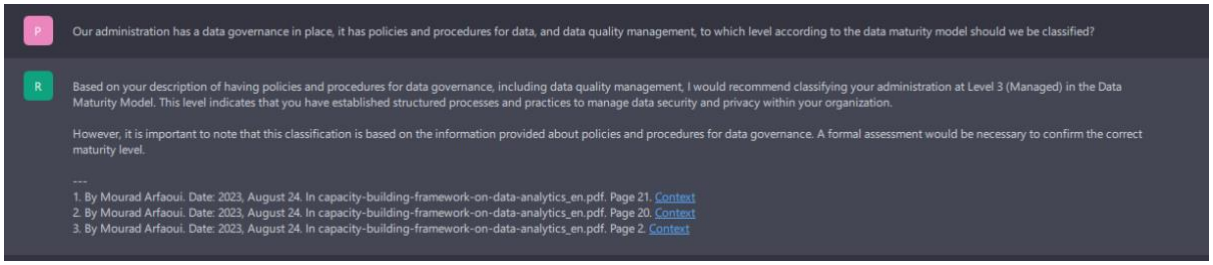


Figure 6: Chatbot assesses the Organizations Analytics Maturity according to the data it was provided.

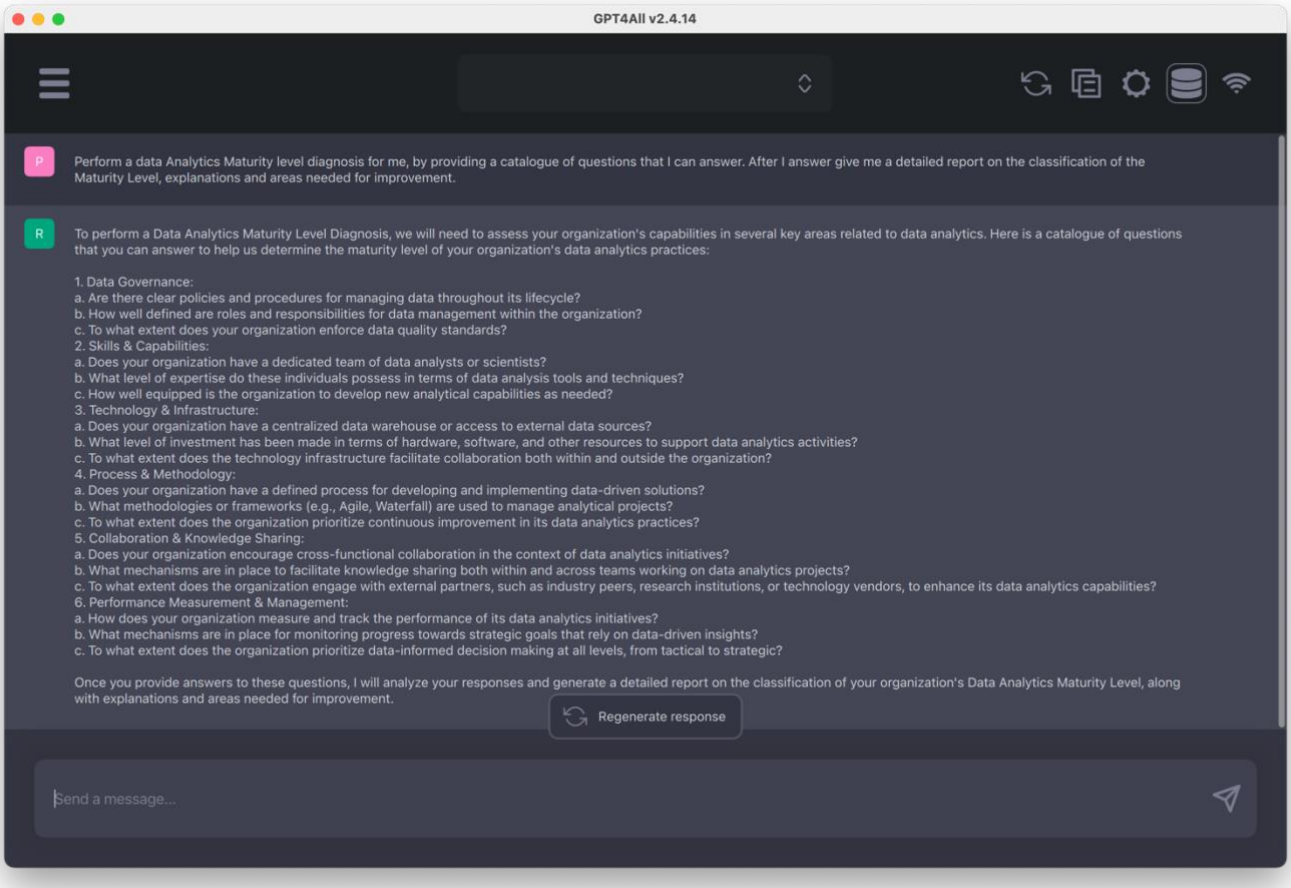


Figure 5: Going even further we can perform a full Diagnostic Process for Analytics Maturity, by asking it to create a catalogue of questions, which the Chatbot will evaluate and create a diagnostic report on.

This is the answer I provided to the chatbot with the intention of getting a grading on the second level of ‘Emerging’ according to the maturity model according to the WCO.

1. Data Governance:
a. Yes, we have an initial set of policies and procedures for managing data.
b. Moderate, there is a general understanding of roles but not in a structured form.
c. Partial, we have basic data quality checks but they are not uniformly enforced.

2. Skills & Capabilities:
a. Yes, we have a small team dedicated to data analytics.
b. Intermediate, they are proficient in basic tools like Excel and have some SQL knowledge.
c. Moderate, we have the ability to develop new capabilities but it's not a streamlined process.

3. Technology & Infrastructure:
a. Yes, we have a centralized data warehouse but it's not fully integrated.
b. Moderate, we have invested in essential hardware and basic analytics software.
c. Limited, internal collaboration is possible but not external.

4. Process & Methodology:
a. Yes, we have a basic process for developing data-driven solutions.
b. Mixed, we use a blend of Agile and Waterfall depending on the project.
c. Partial, we aim for continuous improvement but it's not systematically managed.

5. Collaboration & Knowledge Sharing:
a. Encouraged, but not a mandatory practice.
b. Ad-hoc, some teams document their work for internal sharing.
c. Limited, we occasionally engage with industry peers but not with research institutions.

6. Performance Measurement & Management:
a. Basic KPIs, we have some KPIs but they are not comprehensive.
b. Informal, monitoring is done but not rigorously linked to strategic goals.
c. Partial, data is used for decision-making but not deeply embedded in the culture.

The answer provided by the model is vague, because we provided it with vague answers, it is best to answer very specifically listing all your organizations efforts to get more precise answers. You do not have to worry about data leakage as this as mentioned previously only runs locally.

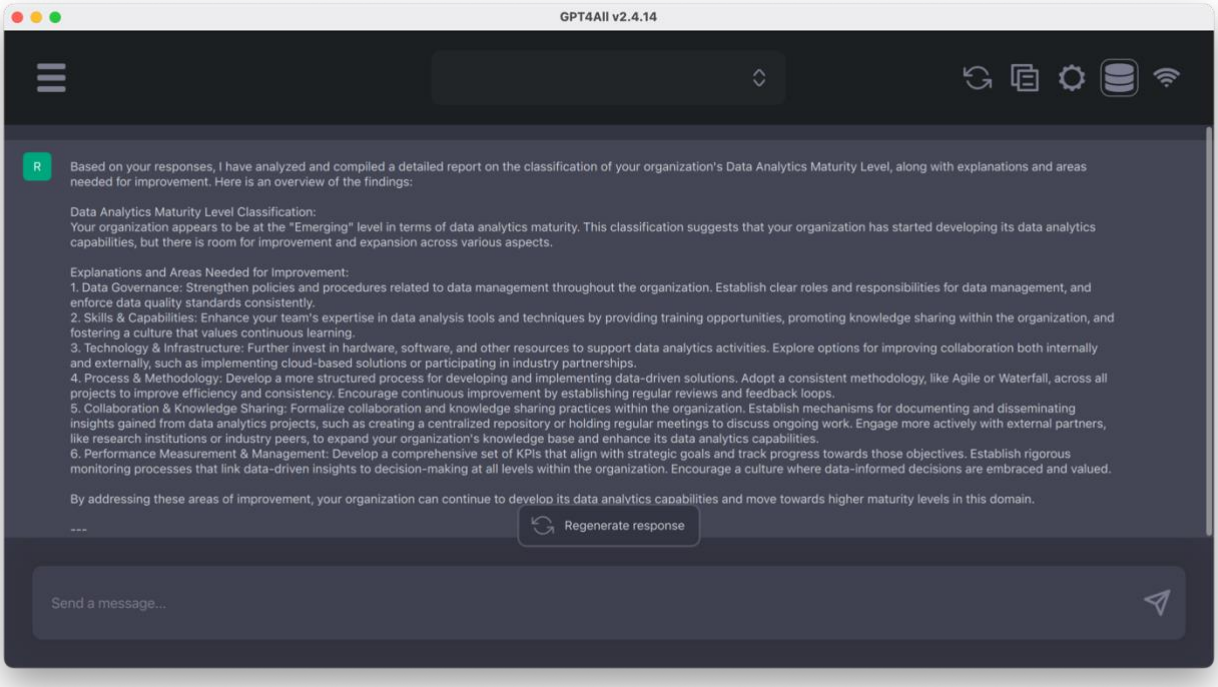


Figure 7: The Chatbot correctly assesses the maturity level to 'Emerging' it also provides explanations for its decision an areas of improvement.

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.

Summary and Conclusions

In conclusion, the utilization of chatbots, particularly those built on Large Language Models like GPT4All, holds significant promise for Customs Organizations and agencies working in similar high-security domains. These solutions offer an innovative approach to tasks ranging from assessing an organization's analytics maturity to providing real-time, actionable insights—all while minimizing the need for specialized technical skills.

Our exploration demonstrated two distinct pathways for implementing chatbots in a secure and private manner. On one hand, open-source libraries like LangChain offer a robust, though somewhat technically challenging, approach that provides a high degree of customizability and security. On the other, no-code solutions like the GPT4All application provide an accessible and secure way to leverage the capabilities of advanced Large Language Models, even for those with limited technical expertise.

While there are valid concerns about data security and privacy, these can be ameliorated through careful implementation and the selection of privacy-focused features, like local data processing and disabling chat history.

It is essential to recognize that the world of chatbots and Large Language Models is in a constant state of flux, with new technologies and methodologies emerging at a breakneck pace. This underlines the importance of remaining vigilant and updated on the latest advancements in the field to capitalize on new opportunities and address emergent challenges effectively.

The promise held by these technologies is immense, particularly for organizations that handle complex data analytics tasks. By making an informed and cautious approach towards adoption, Customs Organizations can substantially benefit in terms of operational efficiency, data analytics capabilities, and overall effectiveness. Therefore, as we move forward in this increasingly digitalized world, it becomes imperative for agencies to seriously consider the practical applications of these rapidly evolving technologies while diligently mitigating potential risks.