CSI Hackathon 2023

(Theme 2: Generative & Natural Language)

Team Name: A.I.chemists

Team members

Tanvir Ahmed (tanvir@uwm.edu)

Nayan Banik (nbanik@uwm.edu)

Md Tanzil Shahria (mshahria@uwm.edu)

Anisha Tasnim (tasnim@uwm.edu)

Table of Contents

Table of Figures

1 Project Overview

The initiative undertaken by our team, A.I.chemists, focuses on bridging the gap between the intricate world of advanced manufacturing and those unfamiliar with its technical nuances. Our project centers around the development of a sophisticated intelligent chatbot, tailored to serve as a virtual guide through the complex labyrinth of the UWM CSI advanced manufacturing test bed.

1.1 Objectives

Our primary objective is to unravel the intricacies of the advanced manufacturing test bed, translating its complex operations and advanced technologies into a language that is accessible to all. This endeavor is particularly challenging given the technical depth and specialized knowledge that the industry entails. To address this, our chatbot employs advanced AI technologies, leveraging Azure's powerful cloud capabilities and the specialized CSI's Factory GPT model. This model is specifically trained in manufacturing vernacular, enabling it to understand and articulate complex concepts in simple terms. The chatbot is designed to extract and process key information from detailed operating manuals and technical handbooks, transforming this data into an easily digestible format. This process is pivotal in making the chatbot a reliable, informative, and accessible virtual guide.

1.2 Potential Impact

The anticipated impact of this project, centered around the development of an AI-powered chatbot for the UWM CSI advanced manufacturing test bed, is both profound and multifaceted. Primarily, this chatbot is envisioned to revolutionize the way diverse groups engage with and comprehend the complexities inherent in advanced manufacturing processes. Its deployment marks a significant stride in educational enhancement, particularly in demystifying the sophisticated operations of the manufacturing sector.

- Educational Enhancement and Inclusive Participation: The chatbot is designed to break down complex industrial procedures into understandable segments, thereby serving as a potent educational tool. It aims to bridge the gap between advanced manufacturing experts and those with limited or no technical background. By rendering intricate manufacturing concepts more accessible, the chatbot fosters a conducive environment for learning and curiosity. This is particularly crucial in an era characterized by rapid technological advancements in manufacturing. The chatbot's role in enhancing understanding is not limited to students and educators; it extends to encompass industry professionals and policymakers. This broad spectrum of beneficiaries underscores the chatbot's potential in facilitating a deeper, more nuanced understanding of the field.
- Catalyst for Broadening Discourse: Beyond its educational utility, the chatbot acts as a catalyst for broader participation in discussions surrounding industrial innovation. It creates a platform where individuals from various backgrounds can engage in meaningful dialogue about advanced manufacturing. This inclusivity is vital for fostering a diverse pool of ideas and perspectives, enriching the discourse in this field. In doing so, the chatbot plays a critical role in promoting collaborative exploration and innovation.
- Transformative Role of Artificial Intelligence: The project also stands as a testament to the transformative power of artificial intelligence in bridging knowledge gaps. By harnessing AI to translate technical knowledge into an approachable format, the chatbot challenges the traditional barriers of complexity and specialized language that often alienate non-experts. It represents a paradigm shift towards a more democratized approach to learning and interaction

with complex technologies. This democratization of knowledge is not just a technological leap; it signifies a fundamental change in how we interact with and perceive the world of advanced manufacturing.

- Extending Impact Beyond Educational Boundaries: The chatbot's impact transcends the realms of education and industry discourse. It sets a precedent in utilizing AI to make specialized knowledge universally accessible, thereby inspiring innovation and participation across various audiences. This goes beyond merely presenting information; it involves engaging users in an interactive learning experience, thereby deepening their understanding and interest in the field.

2 Technical Documentation

The technical foundation of our project leverages Azure's robust cloud capabilities and OpenAI's advanced language models. Our chatbot is built on the Retrieval-Augmented Generation (RAG) pattern, which combines the strengths of Azure AI Search and Azure OpenAI. Azure AI Search plays a pivotal role in data indexing and retrieval. The chatbot can rapidly access a wide array of information by organizing and indexing detailed user manuals and other technical documents related to the manufacturing test bed. This setup ensures the chatbot retrieves the most relevant and accurate data in response to user inquiries. The code is based on azure-search-openai-demo.

2.1 Frontend Technologies

- **React**: React is a JavaScript library for building user interfaces, especially single-page applications. It's known for its efficiency in rendering dynamic content and managing state across components. In this project, React is utilized to create a seamless and interactive user interface. It manages the chatbot's state, handles user interactions, and updates the UI in real time as users engage with the chatbot. React's component-based architecture makes the chatbot's frontend modular and scalable. It enables the development team to reuse components and manage the UI's state more efficiently, leading to a faster and smoother user experience.

- **Fluent UI**: Fluent UI is a collection of UX frameworks for creating web and mobile applications that share Microsoft's design language. It provides a set of reusable, accessible components that conform to Microsoft's design standards. Fluent UI components are integrated into the React application to ensure a consistent and familiar look and feel, aligning with Microsoft's design principles. These components are used to build various elements of the chatbot's interface, such as buttons, input fields, and layout structures. By using Fluent UI, the chatbot inherits a professional and polished appearance that is visually appealing and user-friendly. The consistency in design helps in providing a smooth user experience and ensures accessibility standards are met.

2.2 Backend Technologies

- **Quart**: Quart is an asynchronous Python web framework. It allows developers to build web applications that can handle large numbers of simultaneous connections and data requests efficiently. Quart serves as the backbone of the chatbot's backend. It manages asynchronous tasks, processes user requests, and handles the interactions between the frontend, AI components, and the data storage system. Quart's asynchronous capabilities are essential for a chatbot application where real-time data processing and response generation are crucial. It

improves the chatbot's performance by enabling faster handling of concurrent user interactions.

2.3 AI and Data Handling Technologies

- **Azure AI Search**: Azure AI Search is a cloud search service with built-in artificial intelligence capabilities. It allows for the indexing and querying of large volumes of data efficiently. Azure AI Search is used to index and retrieve information from uploaded PDF documents. It powers the chatbot's ability to search through extensive data sets and find relevant information in response to user queries. The use of Azure AI Search enhances the chatbot's functionality in terms of data retrieval and processing. Its AI capabilities allow for more accurate and relevant search results, improving the overall effectiveness of the chatbot.

- **Azure Blob Storage**: Azure Blob Storage is Microsoft's object storage solution for the cloud. It is optimized for storing massive amounts of unstructured data, such as text or binary data. Azure Blob Storage is utilized to store the chatbot's data, primarily the PDF documents that serve as the knowledge base. It ensures that the data is securely and efficiently managed. By using Azure Blob Storage, the chatbot benefits from a high degree of scalability and security in data management. It supports the chatbot's need for a robust and reliable storage solution that can handle large volumes of data.

- **OpenAI ChatCompletion API**: The OpenAI ChatCompletion API is part of OpenAI's suite of tools that offer advanced natural language processing capabilities. It is designed to generate human-like text based on the input provided. This API is crucial in transforming user queries into enhanced search queries and generating human-like, contextually relevant responses based on the data retrieved from Azure AI Search. The integration of the OpenAI ChatCompletion API brings a high level of sophistication to the chatbot's conversational abilities. It allows the chatbot to understand and respond to queries in a way that is natural and intuitive for users.

- **Azure OpenAI Service**: The chatbot's intelligence is powered by Azure OpenAI Service, which accesses the gpt models such as gpt-3.5-turbo and gpt-4-turbo. This model is adept at understanding and generating human-like text, enabling the chatbot to provide accurate, contextually relevant, and engaging responses.

- **System Architecture**: The architecture supports real-time interactions, minimizing latency between query input and response output. Azure AI Search and OpenAI are seamlessly integrated, ensuring a smooth data flow and consistent performance. The chatbot offers a ChatGPT-style interaction, maintaining a natural, engaging conversation flow, with features to track source content and evaluate the trustworthiness of responses. The following diagram

(Figure 1) is adapted from azure-search-openai-demo illustrates the interaction between the user queries, Azure AI Search, OpenAI's response generation, and the chat interface.

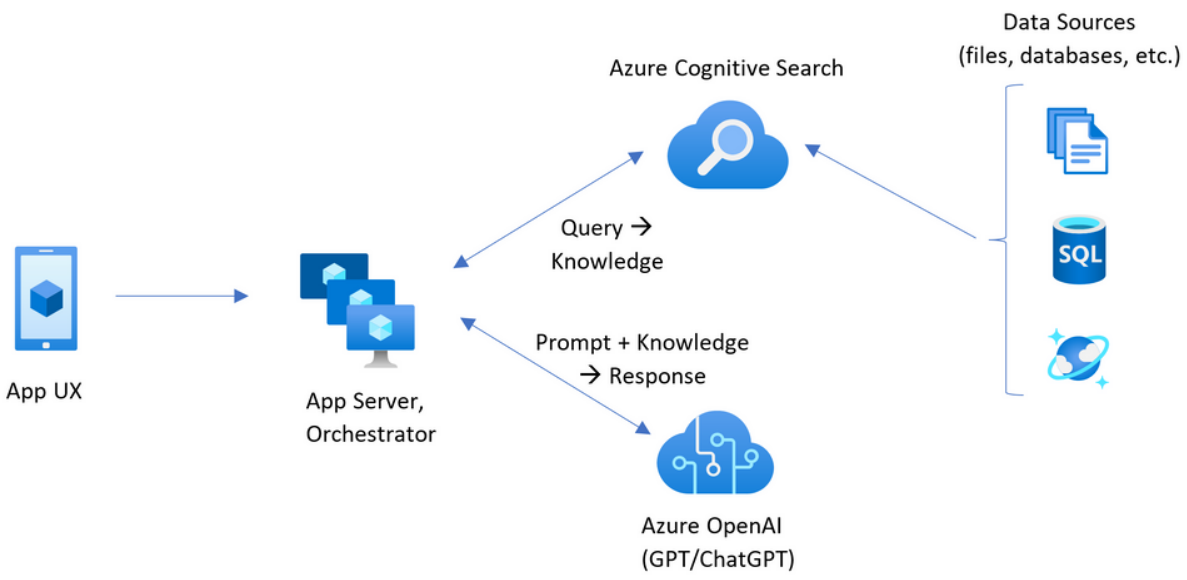

Figure 1: System Architecture Diagram Illustrating the Chatbot Workflow (taken from azure-search-openai-demo)

3 Data Preprocessing Steps

The data preprocessing involves several steps to ensure the chatbot can efficiently process and retrieve information from the provided PDF documents. The following diagram (Figure 2) adapted from azure-search-openai-demo illustrates the preprocessing pipelines.

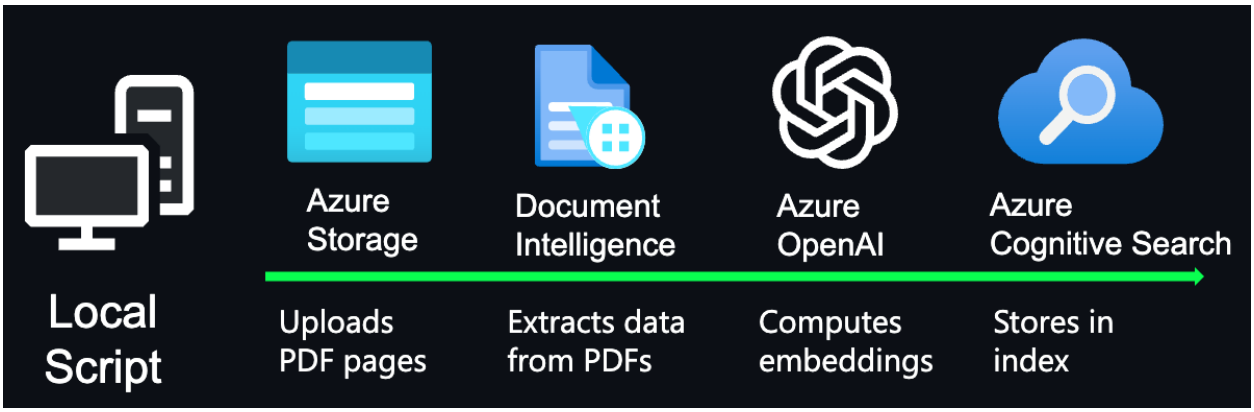

Figure 2: Data Preprocessing Workflow for Chatbot Information Retrieval (taken from azure-search-openai-demo)

- **Local Script:** The preprocessing pipeline begins with a Local Script, which is a program run on a user's computer that selects and uploads PDF files to be processed. This script is the point of initiation for the entire data flow, responsible for interacting with cloud-based services and ensuring the correct documents are queued for extraction and analysis.

- **Azure Storage:** Once the documents are ready, they are uploaded to Azure Storage, a cloud service provided by Microsoft that offers secure, scalable, and accessible storage for large quantities of unstructured data. In this step, the PDF pages are stored as binary data, ready to be accessed by subsequent services for further processing.
- **Document Intelligence:** To address token limits in OpenAI API and enhance context sensitivity, PDFs are split into smaller chunks using a sliding window approach. This is handled by the scripts/prepdocs.py script. The next stage involves Document Intelligence, a sophisticated service that utilizes advanced machine learning algorithms to automatically extract text and other relevant information from the stored PDF documents. This process is key for converting the unstructured data within the PDFs into structured data that can be more easily manipulated and analyzed.
- **Azure OpenAI:** After extraction, the structured data is passed to Azure OpenAI, where it is transformed into embeddings. These embeddings are dense vector representations of the documents, designed to capture the nuanced semantic meaning of the text. This step is crucial for enabling complex search functionalities that go beyond simple keyword matching, allowing for more intelligent and contextual data retrieval.
- **Azure Cognitive Search:** Finally, the embeddings are indexed by Azure Cognitive Search, a powerful search-as-a-service solution that allows for the rapid and sophisticated querying of large datasets. By storing the embeddings in an index, Azure Cognitive Search can facilitate fast, efficient, and highly relevant search results based on the semantic content of the documents, rather than just surface-level text matches. This concludes the pipeline, providing end-users with an advanced search capability over the processed documents.

3.1 Document Management Protocol

Base Dataset Application: The core documents provided by CSI, designated as the "Base Dataset," specifically "CSI Phase 1 rev 3 capital removed.pdf," "Operation and Maintenance Manual.pdf," and "Operations Manual.pdf," were directly utilized in their unaltered state to facilitate the training and benchmarking processes for both the gpt-3.5-turbo and gpt-4-turbo models.

Hypothesis

- Updated Dataset Development: For the "Updated Dataset," we engaged Adobe's Optical Character Recognition (OCR) capabilities to supplement Azure AI Search's OCR functions. This decision was informed by our assessment of Azure AI services' current limitations in directly interpreting images, a significant feature in user manuals which often integrate visual components for better understanding. Language models require textual input for effective training; hence, transforming imagery into descriptive text is crucial for providing additional insights into the testbed's functionalities, robotics details, and particularly the Human-Machine Interface (HMI) content.
- Image Description Integration: Believing in the superior OCR capabilities of Adobe, we proceeded to annotate the bulk of images found in the documents with descriptive text. This enhancement process involved encircling the images with narrative descriptions within the OCR-processed PDFs, thereby enriching the documents with valuable context. The files were then renamed to "Updated_CSI Phase 1 rev 3 capital removed.pdf," "Updated_Operation and Maintenance Manual.pdf," and "Updated_Operations Manual.pdf" to reflect their enriched

status. These revised documents were then employed in the continued training and testing of both the GPT-3.5 Turbo and GPT-4 Turbo models.

4 The Chatbot

The uploaded images showcase the user interface of a chatbot developed for the CSI Advanced Manufacturing Test Bed. Figure 3 depicts the initial state of the chatbot interface, powered by GPT-3.5 with an updated dataset, inviting users to inquire about various aspects of the CSI testbed. Figure 4 illustrates the chatbot's interaction interface, where it provides a detailed step-by-step response on how to reset the tray 2 count in Station 1, demonstrating the chatbot's ability to deliver specific operational guidance. The screenshots highlight the chatbot's user-friendly design and its capacity to facilitate real-time assistance and information retrieval through a conversational AI platform.

Our project's are hosted in GitHub: https://github.com/nynbnk/A.I.chemists_CSI_Hackathon

The deployed applications can be viewed on UWM CSI Chat Engine (GPT-4-turbo-updated-dataset) and UWM CSI Chat Engine (GPT-3.5-turbo-updated-dataset)

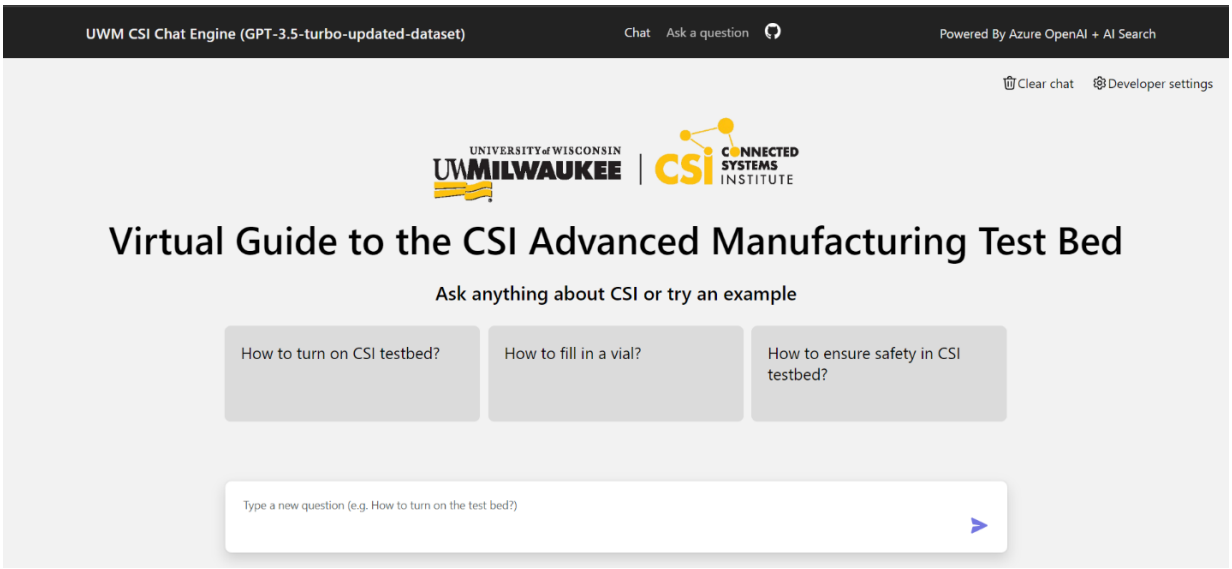

Figure 3: The developed chat bot interface - initial state (gpt-3.5-turbo with updated dataset)

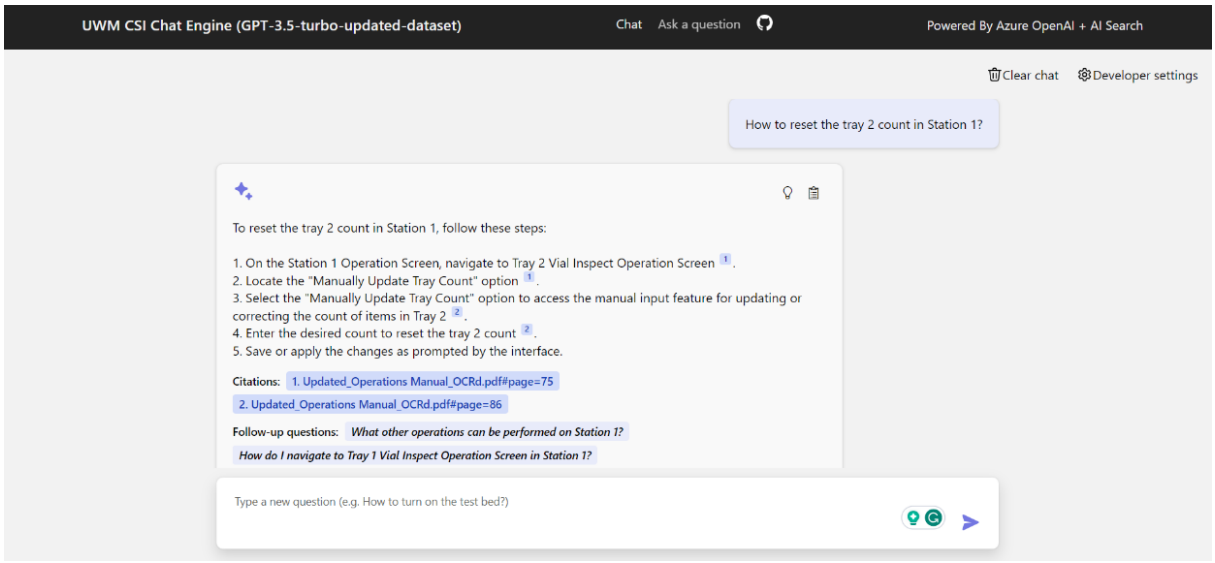

Figure 4: Interaction interface (gpt-3.5-turbo with updated dataset)

5 Team Contribution

The team's collaborative effort was pivotal in the successful development of the chatbot. Tanvir, as the team leader, orchestrated the project's overall direction and strategy, focusing on team management, data preprocessing, and contributing significantly to the model training process alongside Nayan. Nayan, as the key person in data training and model preparation, was instrumental in harnessing the capabilities of Azure OpenAI for the chatbot's language processing features. Tanzil played a crucial role in the frontend development, ensuring the chatbot's interface was user-friendly and intuitive, while also managing the project documentation. Anisha, with her focus on Q&A, was central in evaluating and refining the chatbot's performance, ensuring its responses were accurate, clear, and easily understandable. Her contributions were essential in enhancing the chatbot's usability, making it accessible and beneficial for users regardless of their background in manufacturing.

6 Challenges and Solutions

In addressing the core challenges of our project, we implemented targeted solutions to enhance the chatbot's functionality. Converting unstructured data into a structured format, spanning various sources such as images and overlapping text, proved to be a multifaceted challenge. There were around 80 figures, and we manually included figure descriptions and what the figures represent as a part of data pre-processing. Our strategy harnessed the power of cutting-edge AI tools, notably OpenAI, Azure AI, complemented by a conscientious infusion of manual intervention. It's worth noting, however, that the Azure OpenAI Service introduces a constraint on tokens, thereby shaping the landscape of our interactions and influencing the extent of our engagement within this innovative realm. The pricing dynamics between GPT-3.5 Turbo and its successor, GPT-4 Turbo, reveal intriguing nuances in language models. For the former, the cost stands at $0.0010 per 1,000 tokens for inputs and a slightly higher $0.0020 per 1,000 tokens for outputs, with the understanding that 1,000 tokens encapsulate approximately 750 words. Transitioning to the latter, GPT-4 Turbo shifts the economic landscape, demanding $0.01 per 1,000 tokens for inputs and a more substantial $0.03 per 1,000 tokens for outputs.

7 Performance

In our comprehensive evaluation of the five chatbot models, each team member played a crucial role in manually cross-checking the responses against the provided manuals. This meticulous process involved posing the same set of questions to each model and documenting their respective answers. The team then rigorously compared these responses with the actual content of the manuals to ensure accuracy and relevance. The questions and answers by each model can be found in Appendix-A. Our assessment criteria included accuracy, completeness, clarity, relevance, and documentation reference. For accuracy, we checked how closely the model's responses aligned with the information in the manuals. Completeness was evaluated based on the depth and thoroughness of each response. Clarity was determined by the ease of understanding the responses, while relevance was judged by how well the answers addressed the specific questions asked. Finally, the criterion of documentation reference assessed the model's ability to correctly cite specific sections or pages from the manuals in support of its responses. This hands-on approach by the team ensured that our evaluation was not only based on theoretical metrics but also grounded in practical verification, allowing us to conduct an unbiased and thorough comparison of each model's performance. Such a detailed and methodical evaluation was vital in determining the best-performing chatbot model, ensuring it met the high standards of accuracy, utility, and user-friendliness required for our project. Upon a meticulous subjective evaluation of the chatbot models against the defined criteria of *accuracy, completeness, clarity, relevance, and documentation reference*, we observe the following:

GPT4 bot_Base_Dataset excels in offering precise and pertinent information, often directly referencing the operations manual. Its responses are rich in detail, signifying a deep integration with the available data sources. However, this model tends to produce verbose responses that could potentially overwhelm users seeking succinct information. In contrast, the GPT3.5 bot_Base_Dataset shines in its ability to provide responses that are both clear and concise, which is advantageous for users requiring rapid and straightforward answers. Despite its relevance, this model occasionally sacrifices the depth of information, which may compromise the user's understanding in more complex queries.

The GPT4 bot_Updated_Dataset emerges as the most proficient, particularly after incorporating OCR enhancements to the dataset. This model's responses are not just accurate; they exhibit an elevated level of detail and are meticulously complete, often providing step-by-step guidance. The updated dataset appears to significantly enhance the model's capability to deliver comprehensive and nuanced explanations. On a similar note, the GPT3.5 bot_Updated_Dataset benefits from the enriched dataset, yielding responses that are accurate and relevant. The structure of its responses is logical and coherent, yet it sometimes delivers less granularity compared to the GPT4 model, which may leave users desiring more detailed explanations.

The CSI Azure AI model demonstrates a commendable performance with responses that are lucid and well-detailed, frequently citing specific sections of documentation. The relevance of the information provided is consistent with user queries, indicating a robust retrieval system. Nevertheless, this model occasionally acknowledges the absence of information, an admission that points to transparency but also suggests a potential gap in data completeness. Comparatively, while Azure AI provides clear and applicable responses, the GPT4 bot_Updated_Dataset stands out as the superior model. It adeptly balances the depth and breadth of information, thereby facilitating a user experience that is both educational and user-friendly. The model's sophisticated processing of

the updated information translates into detailed and informative answers, making it an invaluable resource for users engaging with the CSI testbed.

Therefore, the GPT4 bot_Updated_Dataset is adjudged the best-performing model overall. Its advanced capabilities in handling enriched data sources allow it to provide thorough explanations, address complex queries with precision, and offer actionable guidance, hence delivering an exceptional user experience. So, our hypothesis seems to be holding but extensive testing via objective parameters are required before finally concluding the model's capabilities.

8 Future Work

In our future developments, we are committed to significantly enhancing the chatbot's capabilities by integrating advanced artificial intelligence technologies and diverse data sources. A major focus will be on adopting more sophisticated GPT models, such as GPT-4-32k. This model's expanded knowledge base and superior processing abilities are expected to greatly improve the chatbot's understanding and articulation of complex manufacturing concepts. We also plan to introduce automated image recognition and analysis. This advancement will enable the chatbot to interact with and interpret not only text but also images and videos. This feature will be particularly transformative in analyzing visual data from manufacturing manuals, allowing for a more comprehensive and multi-dimensional user experience.

To further augment the chatbot's capabilities, we will implement real-time video feed processing. This will empower the chatbot to analyze and respond to live video inputs, making it an invaluable tool in environments where dynamic visual data is crucial, such as monitoring manufacturing processes or providing real-time assistance in navigating complex machinery. Another significant enhancement will be the development of coherent information structuring through multilevel cross-referencing. By analyzing additional referenced materials in the context of user queries, the chatbot will be able to offer more detailed and contextually rich responses. This will allow the chatbot to navigate complex layers of information and provide users with nuanced and highly accurate information. Inclusion of multi-language support is also on our agenda, aiming to break language barriers and extend the chatbot's reach to a global audience. This will make the chatbot more accessible and useful to a diverse range of users from different linguistic backgrounds.

Furthermore, we plan to make the chatbot accessible on mobile platforms, ensuring its availability across various devices and enhancing user convenience. These planned enhancements represent our dedication to continuous improvement and innovation in conversational AI. By integrating these advanced features, we aim to transform our chatbot into a more intelligent, versatile, and user-centric solution, setting new benchmarks in the application of AI in industrial and educational contexts.

9 Conclusion

In conclusion, our project has marked a significant advancement in conversational AI within the realm of advanced manufacturing. We have developed a sophisticated chatbot that leverages the power of advanced GPT models and innovative data processing techniques. This tool effectively bridges the gap between complex manufacturing processes and a diverse user base, demystifying technical intricacies with remarkable clarity and relevance.