AI Sandbox: A Secure and Controllable Human-AI Interactive Platform for Cross-Disciplinary Research and Education through Large Language Models

Tara Troiano, Amaar Rehman, Jonathan Liu, Jesse Parron, Weitian Wang*
Montclair State University
troianot1, rehmana5, liuj6, parronj, wangw@montclair.edu

Abstract – **Recent advancements in Artificial Intelligence (AI) have made significant improvements in modeling existence and creating generalizations by learning natural patterns from the physical world and interacting with humans. It has become increasingly beneficial in many areas such as smart manufacturing, healthcare, intelligent transportation, and education. In this work, we present the design and development of an AI sandbox, a secure and controllable human-AI interactive platform, to advance cross-disciplinary research and education with large language models (LLMs). The developed AI sandbox features its integration and embodiment with generative AI, mainly including a multi-function graphical user interface (GUI), a reliable and security-oriented infrastructure, and a multi-LLM-enabled engine system. It can provide users with a scalable environment for developing, testing, and deploying AI applications without user-provided data being employed for public model training. With the incorporation of multiple LLMs, the friendly GUI allows users to easily access and switch between them. The system can effectively employ a variety of methods for balancing the security needs of future users while still providing rich learning opportunities that come with generative AI. Implementation results and analysis of the AI sandbox suggest that its human-centered design has promising potential in catalyzing AI-enabled creations and applications, advancing responsibility, explainability, trustworthiness, accessibility, and diversity in human-AI interaction and collaboration.**

Keywords - **Artificial intelligence, human-centered AI, sandbox, security, cross-disciplinary, large language models, generative AI**

I. INTRODUCTION

In recent years, Artificial Intelligence (AI) has seen important and rapid evolutions [1]. These newfound developments have been the cornerstone of growth across various sectors such as manufacturing, healthcare, intelligent transportation, and education [2-5]. A study conducted in the pharmaceutical industry showcased how generative AI can enhance productivity, but also impact and further existing technologies [6]. These advancements now allow for new approaches and opportunities in drug discovery and research. In the area of education, a study found that generative AI can provide a personalized learning experience for students [7]. This will greatly enhance the cultivation of student learning and skills. However, an important question can be raised: how can we ensure generative AI is used correctly and safely? A promising solution to it is the creation of an AI sandbox.

An AI sandbox is a controlled environment employed to access large language models (LLMs) that prevent the transfer of users' data, which may be used for training public models [8]. With an AI sandbox, users can benefit from the power that an LLM provides, such as text prompting, image generation, speech-to-text, and document analysis, while ensuring the privacy of their data. Privacy and security are the most critical aspects of the AI sandbox; thus, the system implements industry-standard methodologies to protect users from data leaks, re-identification, and man-in-the-middle attacks.

This platform is needed for improved data security and user privacy. Prompting could contain users' confidential information [9], which may be stored permanently in an LLM database. Storage of any form of personal data is at risk of data leakage, which is unintentional exposure of information to any party outside of a generative AI system. Once data is leaked, it may be increasingly difficult to find and remove. In addition, limiting data encryption to one company doesn't allow for cross-model comparison between the output of different LLMs, which adversely impacts the sustainable use of AI technologies for both research and education, especially in cross-disciplinary areas [10-13]. Different companies that provide sophisticated, publicly available LLMs may use different datasets to train their models. This, in turn, creates differences in output among different companies' models, as well as unique strengths and weaknesses for each LLM. For example, Antropic's Claude model has some of its training focused on developing a 'character' [14]. This potentially makes Claude a stronger candidate than others for paraphrasing books, generating creative prompts, or interpreting images, but might hinder it regarding very granular types of rule-following exercises, such as applying grammar or syntax rules to a piece of writing. Learning how to utilize the different strengths in LLM outputs is a very important aspect of learning with AI. Therefore, an AI sandbox, integrating with multiple LLMs and engines, is an ideal environment for users to learn and harness AI technologies.

One of the AI sandbox platforms was piloted by Harvard University through its Information Technology department in September 2023 [15]. Harvard University states that its sandbox can ensure the safety of data up to level L3 of its data classification system [16]. It defines L3 data as medium-risk confidential information for those with a 'business need to know'; examples of this type of data include "non-public legal work and litigation information", "non-public financial statements", and even "most Harvard source code". Harvard

AI sandbox has the capacity for data visualization and image generation. Other current market deployments of AI sandboxes focus on different functionalities, such as Clemson University's AI sandbox, which is designed more for training AI to handle business problems [17]. Another instance is Meta, which focuses on creating a space for advertisers to test marketing strategies [18]. Additionally, a company called "the Institute of the Future of Work" is developing a sandbox with applications specifically for managing the workplace [19]. Based on these existing platforms, we conduct a survey where prevalent limitations are identified, such as non-generative AI features and few functions for user-system interaction.

To alleviate these gaps, we design and develop an interactive and functional AI sandbox with generative AI integration and embodiment to facilitate cross-disciplinary research and education for both academic and industrial communities. As one of the pioneering systems, the presented AI sandbox in this work is able to provide users with a controllable environment for developing, testing, and deploying AI applications without user-provided data being employed for public model training. This AI sandbox platform is mainly built with a multi-function graphical user interface (GUI), a reliable and security-oriented infrastructure, and a multi-LLM-enabled engine system. The environment is established through robust security measures via different layers of the system and is supported by separating local and cloud-based data processing in the back-end code. The system also flags empty queries and ensures that users are given appropriate responses to their questions. The friendly GUI allows users to easily access different models and switch between them. This study highlights an adaptive way to handle data security and user privacy as those needs become more complex and more integrated into an AI architecture. The developed platform is tested and assessed in a variety of different tasks. This work aims to advance the capacity of AI sandboxes and facilitate human-AI interaction that will catalyze new AI-enabled creations and applications, shape positive user experiences in AI, and benefit society.

II. System Development

A. System Overview

Fig. 1 presents the functional workflow of the developed AI sandbox system. A new user can initiate the system by making a prompt using the GUI, and backend modules receive and handle the request. Dedicated modules, such as those for file uploads, text prompts, or chat-based interactions, handle files or text input based on the user's request. After that, these inputs are sent to an AI layer, which uses selected large language models and APIs to carry out operations like text analysis, image creation, and document parsing. The outputs, such as summaries or pictures, are produced and sent to the user after the system successfully processes the request. An error-handling mechanism is triggered when the system cannot perform a request, whether due to processing issues, invalid inputs, or unsupported formats. This feature ensures transparency and enhances the system's overall usability by giving the user reasons for not fulfilling the request. In addition, users are able to interact with AI in a way that doesn't jeopardize data security and user privacy, which highly ensures the trustworthiness of AI use.

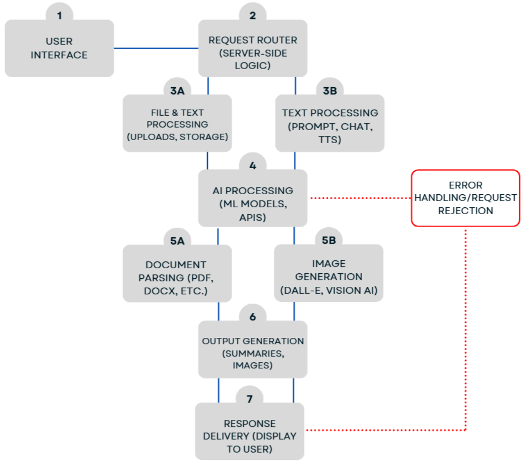

1
USER INTERFACE
2
REQUEST ROUTER (SERVER-SIDE LOGIC)
3A
FILE & TEXT PROCESSING (UPLOADS, STORAGE)
3B
TEXT PROCESSING (PROMPT, CHAT, TTS)
4
AI PROCESSING (ML MODELS, APIS)
ERROR HANDLING/REQUEST REJECTION
5A
DOCUMENT PARSING (PDF, DOCX, ETC.)
5B
IMAGE GENERATION (DALL-E, VISION AI)
6
OUTPUT GENERATION (SUMMARIES, IMAGES)
7
RESPONSE DELIVERY (DISPLAY TO USER)

Fig. 1. Functional workflow of the AI sandbox system.

B. AI Sandbox Engines

Multiple LLMs are integrated into the developed AI sandbox as its engines for users when performing tasks. With different models provided by OpenAI, users will be able to select models ranging from the latest model GPT, GPT-4.0, or the more cost-effective model, GPT-3.5 Turbo [20]. GPT, or a Generative Pre-trained Transformer, is a language model developed by OpenAI designed to understand and generate human-like text. Using a deep learning architecture called a transformer, GPT can process languages with high accuracy. GPT models have been trained on a vast amount of text data, which allows them to perform a wide range of functions. In addition, DALL-E [21] is also utilized in our system. It is OpenAI's generative model designed to create high-quality images from text descriptions. The model is built with an advanced transformer-based algorithm that can translate prompts into visual outputs ranging from realistic images to cartoon images. DALL-E blends creativity and precision, allowing users to easily produce visual image content.

Currently, there are four core functions in the developed AI sandbox including text generation, image generation, speech-to-text generation, and image analysis. The text generation leverages the GPT Model to provide insightful responses, content creation, and support academic work. With this function, LLMs will answer users' questions in detail, summarize text, or propose new ideas based on the given prompt. Users can receive valuable assistance in various tasks ranging from everyday questions to solving research problems. The image generation leverages the DALL-E model to create custom images based on the user's text prompt. It enables users to generate visuals that align with specific descriptions, allowing them to turn their ideas into visual images. This feature is useful for producing artwork for projects, creating visuals for educational materials, or adding custom images to enhance academic presentations. The speech-to-text generation converts spoken English into written text. Similar to the text generation function, this function uses the GPT model to create a response. This function allows users to generate text hands-free, take notes, and capture ideas. The image analysis feature provides detailed descriptions of images and highlights key elements using the GPT model. This powerful tool allows users to gain a deeper understanding of visual content, which makes it easier to interpret an image and extract data.

As OpenAI continues to update its LLMs, the capabilities of each core function in our AI sandbox become increasingly more accurate, interactive, and effective in their respective fields. With each update, each model grows and becomes more accurate, precise, and versatile. Whether for academic work or idea creation, these functions offer powerful tools that allow users to achieve greater potential in innovative education, research, and everyday tasks.

C. GUI Design

The user-centered GUI of our AI sandbox, as shown in Fig. 2, is primarily designed with React using Typescript [22]. The Yarn package manager is used to spin the web-facing application of the software. Google Fonts APIs [23] are included in the base HTML file in the entire package. Optimizations are made between iterative designs of the GUI to improve the user experience, such as having settings be more clearly visually labeled, and separating human and AI chat bubbles by color. The settings are also labeled in plain English and kept permanently visible. This feature is important for novel technology users, who might not be familiar with new terminologies associated with LLMs or AI.

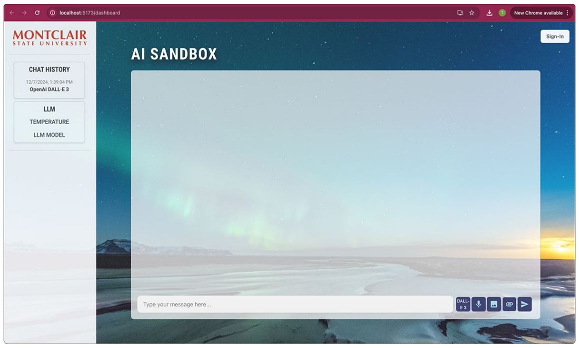


Fig. 2. The AI sandbox GUI.

The different types of inputs are organized in such a way that each can be accessed independently without requiring a submenu to access certain ones, as shown in Fig. 3. Since the developed system aims to welcome users with varying levels of familiarity with AI and technology in general, each function is built visibly digestible and available for use. We believe that this is an important aspect of accessibility that should not be overlooked in designing GUIs for the general population, especially on campuses with individuals with diverse life experiences and abilities. The GUI functions and application results are presented in Section III.

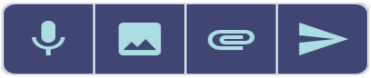


Fig. 3. Input buttons displaying the four main functions on the GUI.

D. Privacy and Security

When working with LLMs, it is imperative to take into account user privacy and data security. The AI sandbox follows a security-centric topology that focuses on securing personal information and user-prompted data across multiple layers of the system. The sandbox system is composed of three layers in which security must be enforced, including the user layer, network layer, and model layer.

At the user level, the AI sandbox enforces a walled-off environment where individual users are provided access to separate instances of the sandbox based on their two-factor authentication login. Once admitted, no personal information is stored within the sandbox environment itself. The ability to prompt the LLM on any information raises a concern for privacy, as it can perform data re-identification. According to [24], AI has the ability to recognize and infer from patterns, which may lead to the unintentional identification of an individual. The AI sandbox imposes pseudonymization techniques that recognize sensitive information and substitute it with realistic counterparts [25]. Additionally, as users work with LLMs, chat history is stored locally in a user's individual instance within the sandbox. Data stored in this chat is decrypted when the user accesses it, otherwise, it is tokenized and encrypted so no other individual can access it.

Once a user sends a prompt, we proceed to the network level of the sandbox. The security of the system's network is integral as we must protect outgoing and incoming data. OpenAI implements transport layer security, where they also incorporate encryption and decryption methods such as AES-256 and FIPS 140-2 [26]. These protocols not only scramble data as it is being sent over the network, but also require a decryption key to access the information. In addition to these security measures, the AI sandbox system will incorporate network segmentation, where data is segmented into smaller chunks to isolate information during network transmission. Lastly, the system will also incorporate techniques to identify and avoid DNS poisoning attacks.

In the model layer, different LLMs accept the information sent from the user and process the requests. This layer ensures the model is not saving any of the data sent from the user to train the model. The intent of avoiding information being used to further train the LLMs is due to data leakage. Although information being transmitted to the LLM should follow the levels of risk imposed by the administration, there is no guarantee that the model will not store small amounts of data. In this case, we try to minimize this in case of data leakage.

E. System Integration

The current integrated AI Sandbox is deployed locally on a laptop with an AMD Ryzen™ 7 7735HS processor, 16GB LPDDR5-6400 memory, and an NVIDIA GeForce RTX 3050 6GB GPU. The GUI is initially rendered on a 16-inch MacBook Pro (2019) featuring a 2.3 GHz 8-Core Intel Core i9 processor, Intel UHD Graphics 630 with 1536 MB memory, and 16GB 2667 MHz DDR4 memory. For future development of our AI Sandbox, we will transfer it to a GPU cluster, which offers significantly higher performance and computational power. The AI Sandbox's data management approach balances local processing and cloud storage to achieve optimal efficiency. Local storage handles sensitive documents and utilizes libraries such as NumPy for efficient computation. This method minimizes external exposure and improves data confidentiality via on-device storage. In contrast, cloud storage is used for the scalability and backup of extensive datasets, logs, and AI models, providing high availability and enabling collaborative access. Security protocols like data encryption in storage, access restrictions, and secure API endpoints ensure that confidential information is safeguarded throughout all system elements. The implementation of these elements is represented in Fig. 4, depicting the interaction among user devices, local processing nodes, and cloud services. This design guarantees that the system is secure and scalable, meeting the varied requirements of academic institutions and industries.

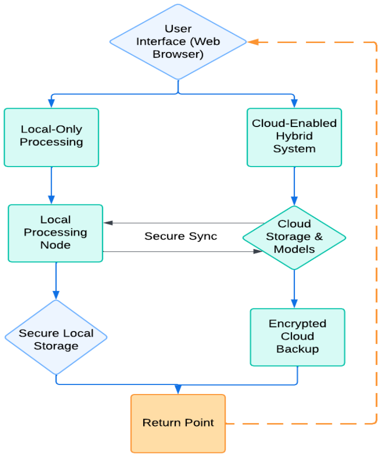


Fig. 4. Data management of the AI sandbox system.

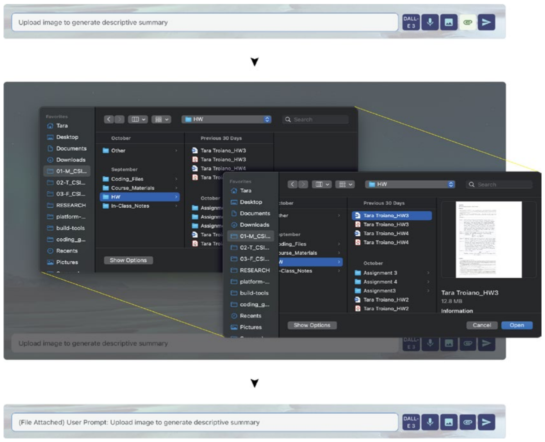
III. Application Results and Analysis

A. Application Results in Different Tasks

Currently, the functions of the AI sandbox interacting with LLMs to generate different outputs include text generation for general questions, description of uploaded images, image generation from text description, and conversion of speech into a text-based prompt. These functions can highly improve the accessibility and diversity of AI applications for users.

The first main function of the sandbox is interpreting uploaded images and generating a descriptive summary in reply. Fig. 5 shows how the user interacts with the image file upload. First, the user selects the OpenAI model that will be used to perform the task on the GUI. Then the user clicks the attachment button. The AI sandbox will direct the user to the local computer and let the user select the desired files. A prompt bar will indicate [File Attached] when successful.

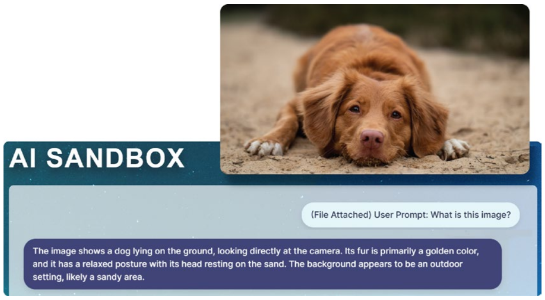
After the user performs text-generated confirmation with the AI sandbox, the selected LLM will execute the image interpretation request. The sandbox GUI will then give a written description of the image, as visibly displayed in Fig. 6. In this application, the image description reads: "The image shows a dog lying on the ground, looking directly at the camera. Its fur is primarily a golden color, and it has a relaxed posture with its head resting on the sand. The background appears to be an outdoor setting, likely a sandy area." This descriptive summary indicates that the AI sandbox accurately understands the uploaded image with the selected LLM.



Upload image to generate descriptive summary

(File Attached) User Prompt: Upload image to generate descriptive summary

Fig. 5. Selecting a file for the AI sandbox to interpret.



(File Attached) User Prompt: What is this image?

The image shows a dog lying on the ground, looking directly at the camera. Its fur is primarily a golden color, and it has a relaxed posture with its head resting on the sand. The background appears to be an outdoor setting, likely a sandy area.

Fig. 6. The descriptive summary of the uploaded image from the AI sandbox.

The second function of the AI sandbox is creating images from a written prompt. To test this function, the user inputs "golden retriever" as the text prompt with the selected DALL-E 3 model. The image that is created from the request is displayed in Fig. 7. The AI sandbox demonstrates that it is able to correctly and effectively generate the requested image with a rich context.

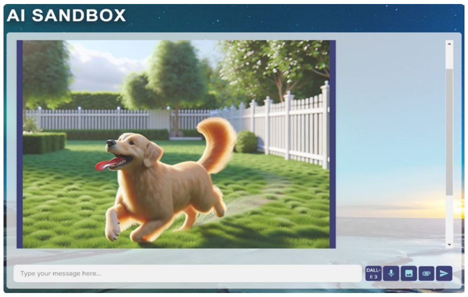
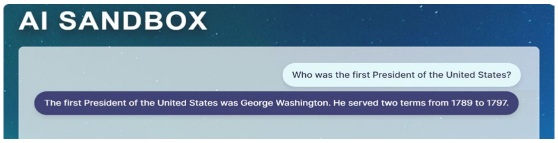


Fig. 7. A golden retriever image generated from the AI sandbox.

Another function is the ability to convert from speech into a text-based prompt. A general microphone can be used with the computer where the AI Sandbox is deployed. In this validation, as presented in Fig. 8, the user inputs a question using the microphone: "Who was the first President of the United States?". The system answers: "The first President of the United States was George Washington. He served two terms from 1789 to 1797. Fig. 9 shows how the buttons change on the GUI when the microphone is activated. The buttons are highlighted orange when the microphone is in use, and the microphone button is highlighted with a cyan-blue ring when the machine converts the sound into text.



Who was the first President of the United States?

The first President of the United States was George Washington. He served two terms from 1789 to 1797.

Fig. 8. Output from the microphone translated into text.

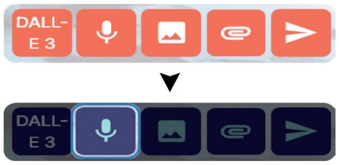
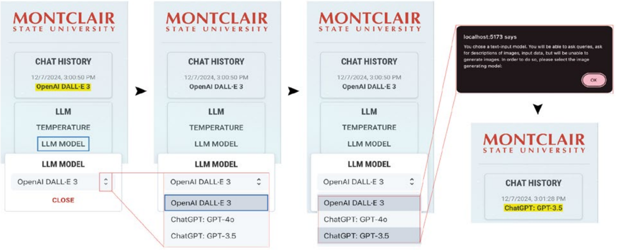


Fig. 9. Changes of the buttons' states when using the microphone.

As shown in the validation tasks, we have currently implemented functionality to explicitly call upon different LLMs from OpenAI for the AI sandbox, including ChatGPT-3.5 Turbo, ChatGPT-4o, and DALL-E 3. We have also designed a pop-up warning when switching between different types of output-generating models. For example, switching between a version of ChatGPT (text-driven output) and DALL-E (image output). As displayed in Fig. 10, the user accesses a dropdown to select an LLM. If the switching is between different types of LLMs, the user receives a prompt reminding them that they can access one type of media, but not another. These designs can highly shape positive user experiences and trust in AI applications.



localhost:5173 says

You chose a text-input model. You will be able to ask queries, ask for descriptions of images, input data, but will be unable to generate images. In order to do so, please select the image generating model.

Fig. 10. The functionality of dropdown when switching models.

B. User-Centered Design Analysis

Depending on the complexity of the request, the AI sandbox's response time can range from 2 to 30 seconds. Simple text-based prompts are answered more quickly in comparison to tasks like image generation or processing documents or images into a response. The AI sandbox platform is accessible across different operating systems on both Windows and macOS. This will enlarge its development and deployment potential and diversity for broader applications. Users are not allowed to submit empty queries into the search bar. This prevents issues that occur when the LLM is given null entries. Non-existent connections with the LLMs will also be flagged within a response bubble in the GUI. These user-centered error-handling designs in the AI Sandbox can largely guarantee stable and responsible system performance to improve user experience and trustworthiness in AI use.

In addition to guiding users from invalid prompts, the system can verify input forms and offer feedback. For example, it will advise users to reword their request or upload the appropriate file types. It also handles complicated or unsupported requests by providing the user with error messages, such as those alerting them to system restrictions or temporary unavailability of certain features. These error management measures are able to preserve the explainability and integrity of the AI sandbox system, allowing for smooth human-AI interaction with reliability and fairness.

C. Efficiency and Scalability Analysis

The AI sandbox's data management system is purposefully built for efficiency, security, and scalability by deciding between local and cloud-based processing mechanisms. To ensure data privacy, tasks that require higher levels of data security (like graphing and rendering sensitive data) are carried out locally. This method uses robust programming languages and libraries, like the NumPy library in Python, to compute and present data efficiently without requiring cloud connectivity. On the other hand, cloud-based solutions guarantee extendibility and accessibility while preserving safe data handling for particular activities and are only used for tasks that need substantial computational resources or collaborative access. These measures can positively enable the efficiency and scalability of the AI sandbox, offering compatible human-AI collaboration with responsibility and inclusivity.

IV. Discussions and Future Work

The application results and analysis in Section III suggest that the developed AI sandbox can provide users with a secure and controllable interactive platform for cross-disciplinary usage via different LLMs. Effective security techniques have been proposed in the developed system, such as using pseudonymization to swap identifying information with a substitute, using encryption to prevent access to data, and using network segmentation to break data into smaller chunks. The backend supports the internal software structure by keeping input in a modular organization. Error handling measures also ensure the transparency, responsibility, and trustworthiness of AI capabilities and handling of requests. The user-friendly GUI makes the system initiate and enter prompts clearly without unnecessary distractions.

These benchmarks are significant because they allow the users to access the benefits of AI in a way that's intuitive, approachable, and safer. In addition, being able to create walled-off environments for LLMs is a highly important aspect of cybersecurity and responsibility in the future deployment of LLMs on a larger scale. As AI becomes increasingly integrated into society, the safety requirements for using AI increase exponentially. The developed AI sandbox is crucial in further creating methods and technologies to provide multiple sectors of academia and industry access to LLMs in a way that meets users' needs and inspires confidence in utilizing AI.

To improve its performance for shaping better user experiences, we aim to address the following features for the AI sandbox in our next steps, including temperature control, chat history, downloading sessions, clearing information, and revisiting the last 10 searches. While they are implemented in the GUI, they need to be further developed and optimized.

The temperature control for LLMs is defined by the probability that a token gets selected when a model is generating output [27]. Lower temperatures are associated with putting more weight into selecting similar tokens to the prompt, so the output is more predictable and more likely to be factually accurate. This is a more sophisticated feature, and we will include it in the following versions. The features around chat history, downloading sessions, deleting/clearing data, and revisiting prior searches would also be included in our future work as they are all intrinsically connected to an instance of a user's account.

V. Conclusion

In this work, we have developed an AI sandbox, a human-AI interactive platform integrating with multiple LLMs for users in cross-disciplinary applications. The platform features its integration and embodiment with generative AI, mainly including a multi-function graphical user interface, a reliable and security-oriented infrastructure, and a multi-LLM-enabled engine system. It has a robust security topology that can allow users to employ LLMs with a highly reduced risk of data leaks. The developed AI sandbox is able to establish a walled-off and controllable environment for users to harness the power of LLMs to answer text-based prompts, generate images, interpret data, and translate voice-based input into tangible answers. The implementation of this environment is through robust security measures via different layers of the system and is supported by separating local and cloud-based data processing. The developed AI sandbox differs from previously established work as it offers user-centered designs and functions in its development process. Application results and analysis indicate that the developed system has promising potential to advance responsibility, explainability, fairness, trustworthiness, accessibility, and diversity in human-AI interaction and collaboration. We have also discussed the future work of this platform to improve its performance for shaping better user experiences in AI.

Acknowledgments

This work is supported in part by the National Science Foundation under Grants CNS-2117308 and CMMI-2338767.

References

[1] W. Ertel, *Introduction to Artificial Intelligence*. Springer Nature, 2024.

[2] H. Diamantopoulos and W. Wang, "Accommodating and Assisting Human Partners in Human-Robot Collaborative Tasks through Emotion Understanding," in *2021 International Conference on Mechanical and Aerospace Engineering (ICMAE)*, 2021: IEEE, pp. 523-528.

[3] W. Wang, C. Coutras, and M. Zhu, "Empowering computing students with proficiency in robotics via situated learning," *Smart Learning Environments,* vol. 8, no. 1, pp. 1-18, 2021, doi: 10.1186/s40561-021-00167-6.

[4] R. Li, W. Wang, Y. Chen, S. Srinivasan, and V. N. Krovi, "An End-to-End Fully Automatic Bay Parking Approach for Autonomous Vehicles," in *ASME 2018 Dynamic Systems and Control Conference*, 2018, V002T15A004.

[5] W. Wang, R. Li, L. Guo, Z. M. Diekel, and Y. Jia, "Hands-Free Maneuvers of Robotic Vehicles via Human Intentions Understanding Using Wearable Sensing," *Journal of Robotics,* vol. 2018, no. 1, p. 4546094, 2018.

[6] G. Doron, S. Genway, M. Roberts, and S. Jasti, "Generative AI: driving productivity and scientific breakthroughs in pharmaceutical R&D," *Drug Discovery Today,* p. 104272, 2024.

[7] J. Su and W. Yang, "Unlocking the power of ChatGPT: A framework for applying generative AI in education," *ECNU Review of Education,* vol. 6, no. 3, pp. 355-366, 2023.

[8] AI Sandbox, [Online]. Available: https://huit.harvard.edu/ai-sandbox.

[9] G. Feretzakis and V. S. Verykios, "Trustworthy AI: Securing sensitive data in large language models," *AI,* vol. 5, no. 4, pp. 2773-2800, 2024.

[10] W. Wang, Z. Przedworska, J. Parron, M. Lyons, M. Zhu, and A. Tuininga, "MCROS: A Multimodal Collaborative Robot System for Human-Centered Tasks," in *2024 International Conference on Networking, Sensing and Control (ICNSC)*, 2024: IEEE, pp. 1-6.

[11] J. Parron, T. T. Nguyen, and W. Wang, "Development of A Multimodal Trust Database in Human-Robot Collaborative Contexts," in *2023 IEEE 14th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 2023: IEEE, pp. 0601-0605.

[12] C. Hannum, R. Li, and W. Wang, "A Trust-Assist Framework for Human-Robot Co-Carry Tasks," *Robotics,* vol. 12, no. 2, pp. 1-19, 2023.

[13] W. Wang, R. Li, Y. Chen, Z. M. Diekel, and Y. Jia, "Facilitating Human–Robot Collaborative Tasks by Teaching-Learning-Collaboration From Human Demonstrations," *IEEE Transactions on Automation Science and Engineering,* vol. 16, no. 2, pp. 640-653, 2018.

[14] Claude's Character, [Online]. Available: https://www.anthropic.com/news/claude-character.

[15] AI Sandbox Pilot Launches, [Online]. Available: https://huit.harvard.edu/news/ai-sandbox-pilot.

[16] Harvard Data Classification, [Online]. Available: https://privsec.harvard.edu/data-classification-table.

[17] Clemson AI Sandbox, [Online]. Available: https://www.clemson.edu/centers-institutes/launchpad/about/ai-center.html.

[18] Meta AI Sandbox, [Online]. Available: https://www.facebook.com/business/news/introducing-ai-sandbox-and-expanding-meta-advantage-suite.

[19] IFOW AI Sandbox, [Online]. Available: https://www.ifow.org/landing-page/sandbox.

[20] OpenAI Models, [Online]. Available: https://platform.openai.com/docs/models.

[21] DALL-E 3, [Online]. Available: https://openai.com/index/dall-e-3/.

[22] React, [Online]. Available: https://react.dev/learn/typescript.

[23] Google Fonts, [Online]. Available: https://developers.google.com/fonts.

[24] F. Descalzo, "Designing Artificial Intelligence with Privacy at the Center," in *2024 IEEE Biennial Congress of Argentina (ARGENCON)*, 2024: IEEE, pp. 1-4.

[25] O. Yermilov, V. Raheja, and A. Chernodub, "Privacy-and utility-preserving nlp with anonymized data: A case study of pseudonymization," *arXiv preprint arXiv:2306.05561,* 2023.

[26] Federal Information Processing Standard, [Online]. Available: https://learn.microsoft.com/en-us/azure/complian ce/offerings /offering-fips-140-2.

[27] Temperature control for LLM, [Online]. Available: https://www.promptingguide.ai/introduction/settings.