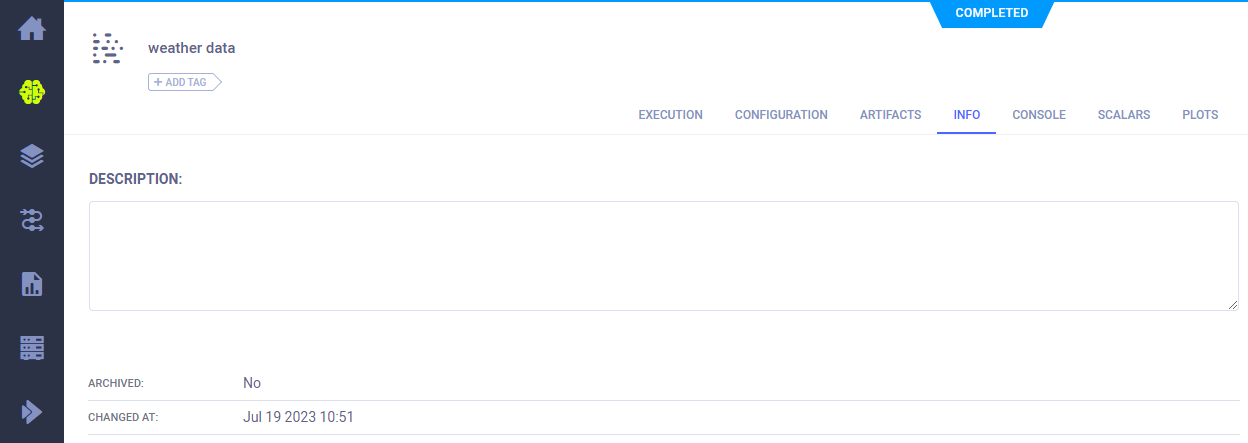
**Introduction**

In the rapidly evolving landscape of artificial intelligence, data exploration is the foundation upon which innovations are built. However, the journey of AI developers is often met with significant challenges such as in the area of ambiguous data tagging. For instance, some platforms consolidate metadata input into a single description box. This unstructured and confusing approach hampers efficient data tagging and exploration for AI developers. These obstacles not only hinder their progress but also impede the potential for discoveries that lie within the realm of available data.



Unified Metadata Field: Represented in a Single Description Box

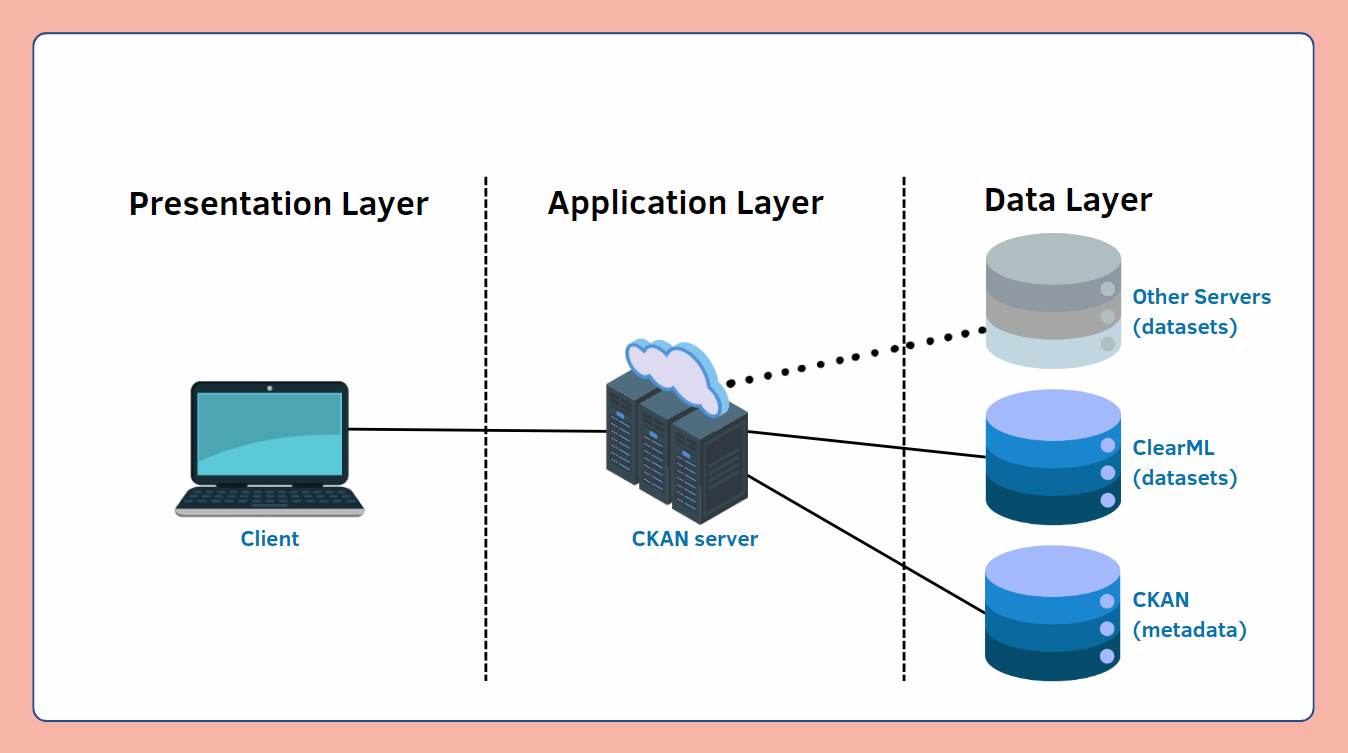
**Background and Objectives**

To address these issues, our focus was on developing a Data Catalog with robust dataset management capabilities, following the principles of CRUD (Create, Read, Update, Delete). Our primary objective was to construct a Web Application that would grant users with simplified dataset creation, efficient manipulation, and seamless access control.

**Project Description**

Introducing DataVerse, our internship project that aims to create a data management solution. Leveraging CKAN and ClearML. DataVerse provides a unified platform for seamless data organization and exploration, granting engineers easy access to a comprehensive data catalog with detailed metadata for enhanced data understanding. The application will also enable data version tree tracking, empowering engineers to analyze dataset changes, leading to improved model training and overall data comprehension. With ClearML integration, we ensure a smooth transition from data exploration to ML experiment tracking, optimizing workflow efficiency.

Looking towards the future, our vision extends beyond the integration of ClearML alone. We aspire to create a versatile platform that can be expanded to cater to other data servers and management tools, thus providing a flexible and extensible data management ecosystem.

****

DataVerse’s Architecture

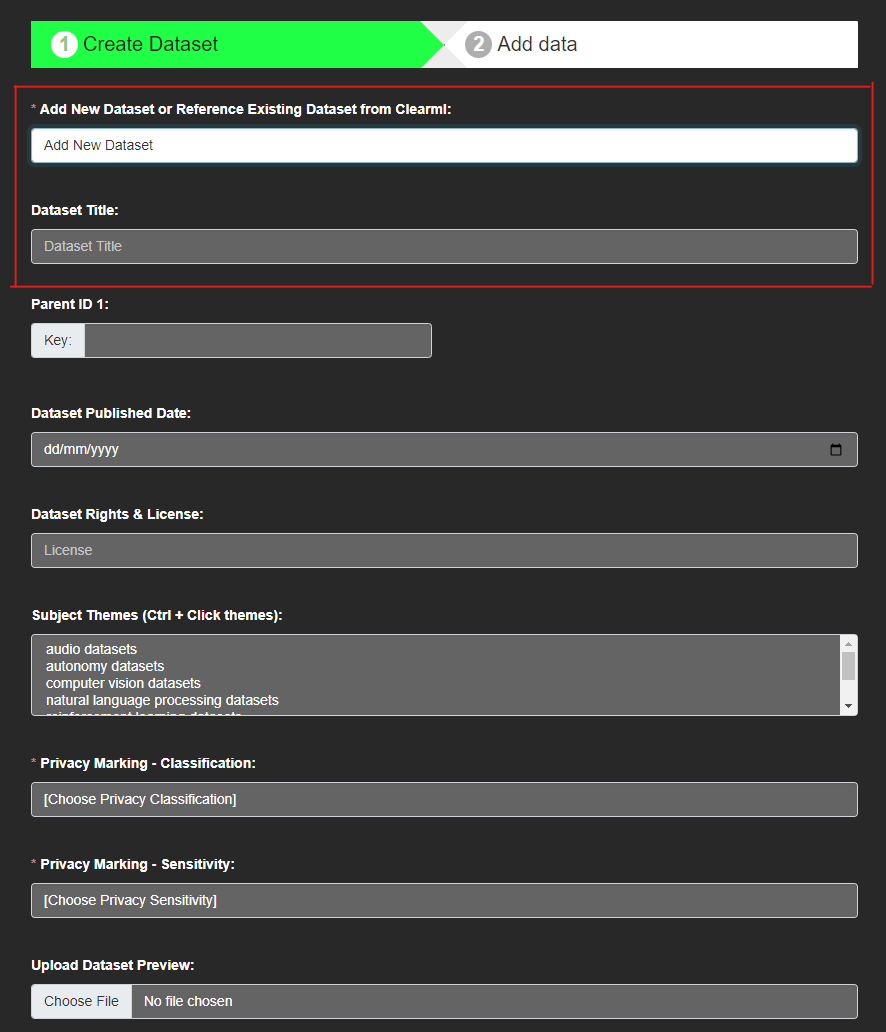
**Results and Achievements**

During our short internship stint, we have implemented several functionalities on DataVerse, catering to both data editors and data explorers alike. Below, we will delve into the details of some features, offering a comprehensive view of the platform's capabilities.

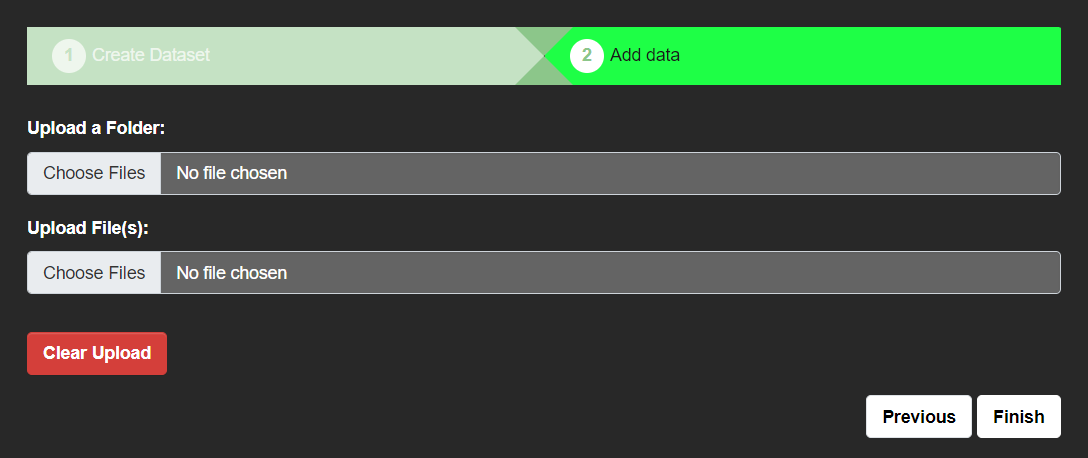
**For Data Editors**

**1. Upload Datasets and Their Metadata through an Intuitive UI**

DataVerse offers users a seamless experience to upload datasets and their metadata effortlessly. On the first page of the upload process, users can select “Add New Dataset” for the first input field before inputting all the essential metadata information.

First Page of Dataset Upload: Partial Metadata Shown

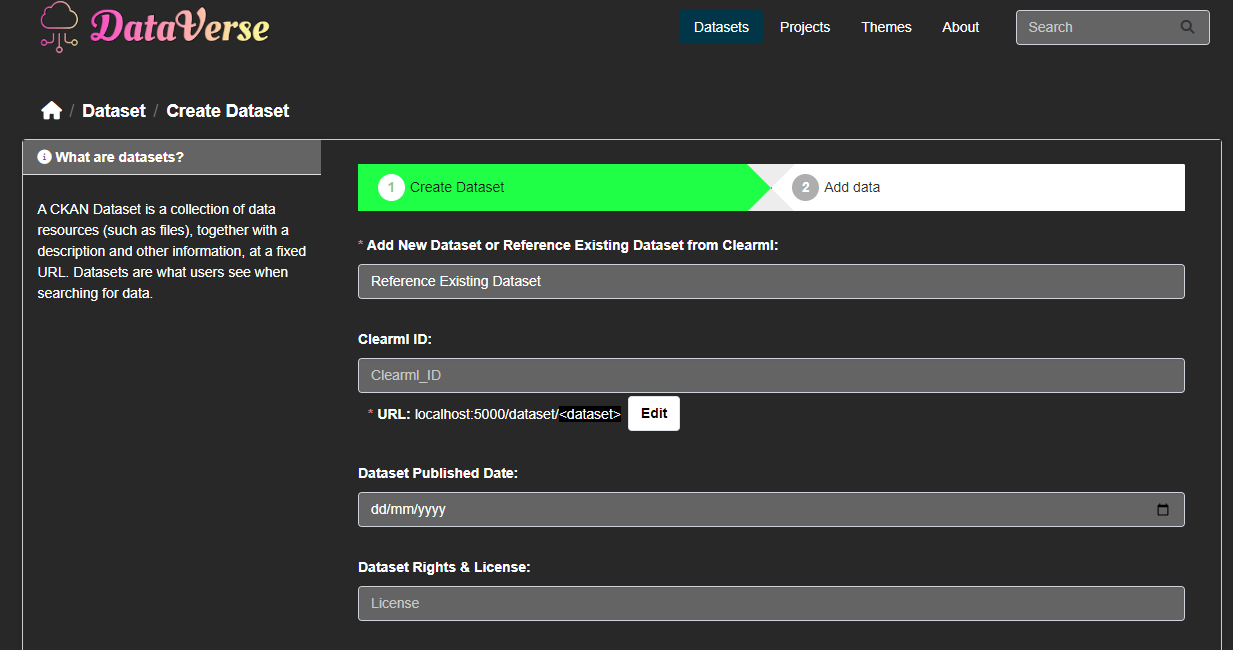
Once the metadata is provided, the user is directed to the second page, where they can conveniently choose to upload their dataset by either selecting a folder or individual files containing the data or both. Uploaded datasets will be sent to ClearML, whereas the metadata will be stored on the CKAN server.

****

Second Page of Dataset Upload

**2. Tag Metadata to Existing Datasets Stored in ClearML**

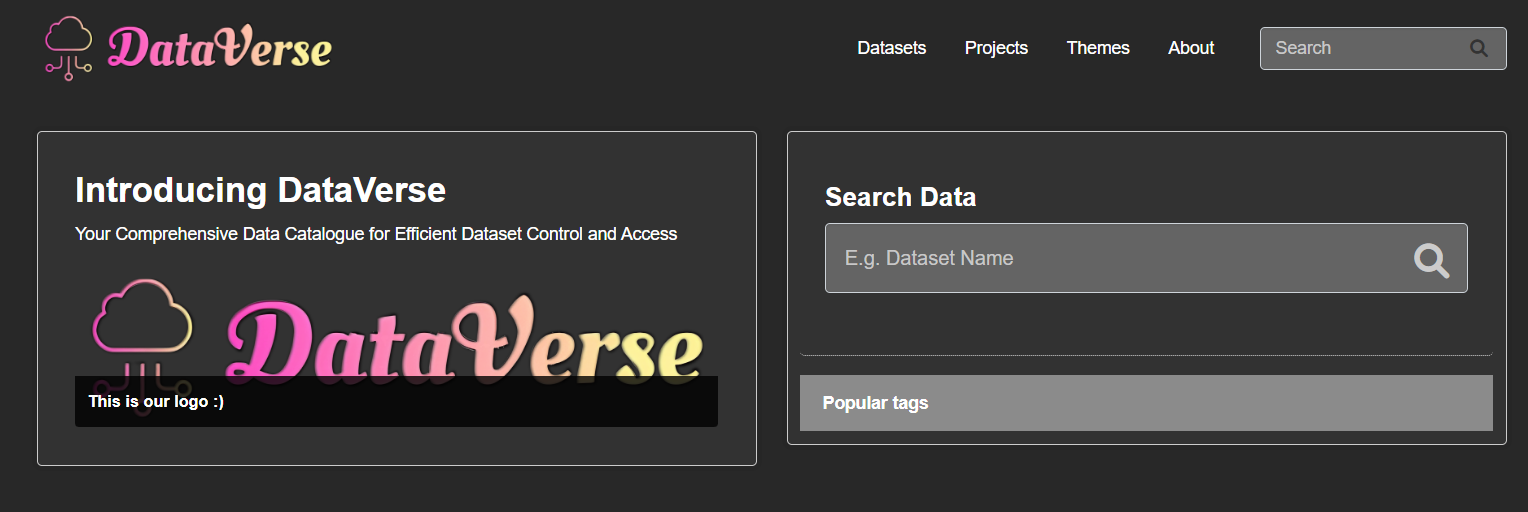
Tagging metadata to existing datasets stored in external servers, such as ClearML, can be done easily with our user-friendly process. During the dataset upload, users can simply select "Reference Existing Dataset from ClearML" as their first step. A prompt will appear, asking users to input the specific ClearML dataset ID they wish to tag the metadata to. Once the ID is entered, users just need to fill in all the necessary metadata fields before submitting the form. This streamlined approach ensures efficient data management and organization within the ClearML ecosystem

****Reference Existing Dataset in ClearML Page.

**For Data Explorers**

**1. Explore Uploaded Datasets Easily with a Robust Search Engine**

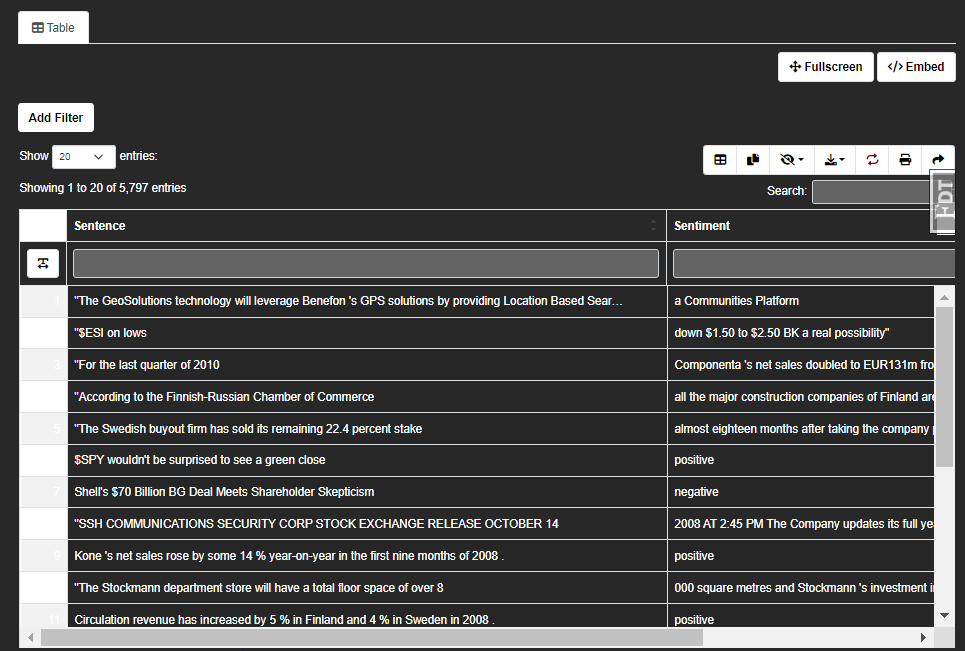
On the Dataverse homepage, users have access to a powerful search box, enabling them to easily discover relevant datasets. The search engine goes beyond just scanning titles; it intelligently considers all metadata fields and uploaded file types to filter out the available datasets. It utilizes fuzzy logic to map the entered keywords to metadata fields with similar content. With this comprehensive approach, users can efficiently find datasets tailored to their specific needs, making dataset exploration an easy and productive experience.



Snippet of DataVerse’s HomePage

**2. View Previews of Uploaded Datasets**

It is important to have the ability to preview datasets before downloading them to save valuable time and ensure relevance. In Dataverse, we offer a crucial functionality that allows data scientists to view dataset previews, offering a glimpse of what the datasets entail. Our preview feature supports various file types, including audio, CSV, image, text, and video files. Looking ahead, we intend to expand this feature to encompass even more file formats.



Finance Sentiment Analysis: CSV File Preview

**3. Track Data Versioning of Uploaded Datasets**

In Dataverse, uploaded datasets are version-tracked, providing data explorers and engineers with a comprehensive history of changes and updates over time. This feature ensures data integrity, transparency, and enables confident analysis and comparison of different dataset iterations, fostering impactful insights.

**Challenges**

Developing DataVerse presented its fair share of challenges, particularly due to our limited prior experience. These challenges pushed us to think outside the box. However, with perseverance and collaboration, we were able to overcome these obstacles and find effective solutions.

One of the major hurdles was establishing communication with external servers to retrieve and store datasets efficiently. As newcomers to server-side scripting, we initially struggled to navigate the intricacies of this process. For example, we were sure our credential initialization to establish a connection with ClearML was set up correctly, but we could not connect to the server. With guidance from our mentor and extensive research, we were able to successfully implement the communication, which was due to the wrong file path that the credentials were stored in. By leveraging on their robust server APIs, we ensured a reliable connection with the external servers.

Additionally, we had a limited timeframe as our internship had a relatively short duration, which posed an additional challenge in terms of project completion. To mitigate this, we adopted an agile approach to project management, prioritizing tasks and focusing on essential functionalities. By maintaining open lines of communication within the team, we successfully delivered an alpha version of DataVerse within the given timeframe.

**Internship Experience**

**Jin Rui**

The beginning of the internship posed a steep learning curve. Joining three weeks after Hao Yi, I had much ground to cover in understanding CKAN’s operations. Luckily, Hao Yi demonstrated remarkable competence and guided me through the complexities of working with CKAN's features. Once I grasped the framework, Hao Yi and I collaborated perfectly to meet our objectives. Despite encountering numerous challenges along the way, our tenacity and determination helped us overcome each hurdle successfully. As a student specializing in Data Science and Artificial Intelligence, the software development aspect of the project posed difficulties for me to comprehend, sometimes leading to frustration. Nevertheless, the thrill and exhilaration of witnessing my logical reasoning materialize in the application far outweighed any negative emotions I had experienced.

I am immensely grateful to my mentor, Kah Siong, whose guidance and expertise were invaluable throughout the entire duration of the internship.

Furthermore, I would like to extend my heartfelt appreciation to my colleague, Mira, for her invaluable assistance in deciphering the intricacies of CKAN's documentation whenever we encountered obstacles. Her insights and collaborative spirit were instrumental in helping us navigate through complex concepts.

Overall, this internship has been rewarding, providing invaluable insights into the practical implementation of my field.

**Hao Yi**

My internship experience provided me with numerous invaluable learning opportunities that went beyond what I could have acquired through traditional classroom learning. I delved into Docker to host my web application, ensuring seamless deployment. This experience not only expanded my technical skills, but also exposed me to the world of containerization and its importance in modern software development.

Additionally, I immersed myself in understanding CKAN’s framework and how to effectively communicate with external servers through server-side scripting. This allowed me to grasp the intricacies of CKAN and explore its various features for dataset management. This could not have been possible without the help of my colleague Mira, who took time out of her own schedule to answer my endless questions. Moreover, I had the privilege of utilizing ChatGPT to explain complex concepts, which immensely helped me and my project partner Jin Rui break down difficult problems.

As with any learning journey, challenges were inevitable. As a relatively inexperienced intern, I faced difficulties in establishing communication with external servers and adapting fixed databases to suit our dataset structure. Furthermore, I realized that my limited experience with frameworks led me to over complicate problems at times. However, with the unwavering support and guidance of my mentor Kah Siong and Mira, and the collaborative efforts with Jin Rui, we were able to overcome these challenges and make significant progress in the development of the Data Catalog, and released a beta product of our Data Catalog, DataVerse.

I would like to thank my mentor, Kah Siong, for his endless patience to aid me in understanding difficult concepts, and to my colleague Mira, who had great insights as to how we could solve our problems. Lastly, I would like to give a big shoutout to my intern partner Jin Rui. Without him, this project would have been an immense hurdle to cross. This internship has been an enriching and eye-opening experience and I am very grateful for the opportunity to join DSTA for my summer.

**Conclusion**

In conclusion, DataVerse represents a powerful data management solution crafted during our enriching internship. It effectively tackles challenges encountered by AI developers in data exploration, offering seamless dataset creation, manipulation, and version tracking. Its user-friendly interface ensures effortless dataset tagging and exploration. Despite obstacles, collaborative efforts and perseverance enabled us to deliver an alpha version successfully. Moving forward, our vision extends to creating a versatile platform by integrating various data servers. DataVerse embodies the rewarding experience of our internship, contributing to the advancement of AI development and empowering engineers worldwide.