H2O LLM Studio

Fine-tuning LLMs with No-Code GUI

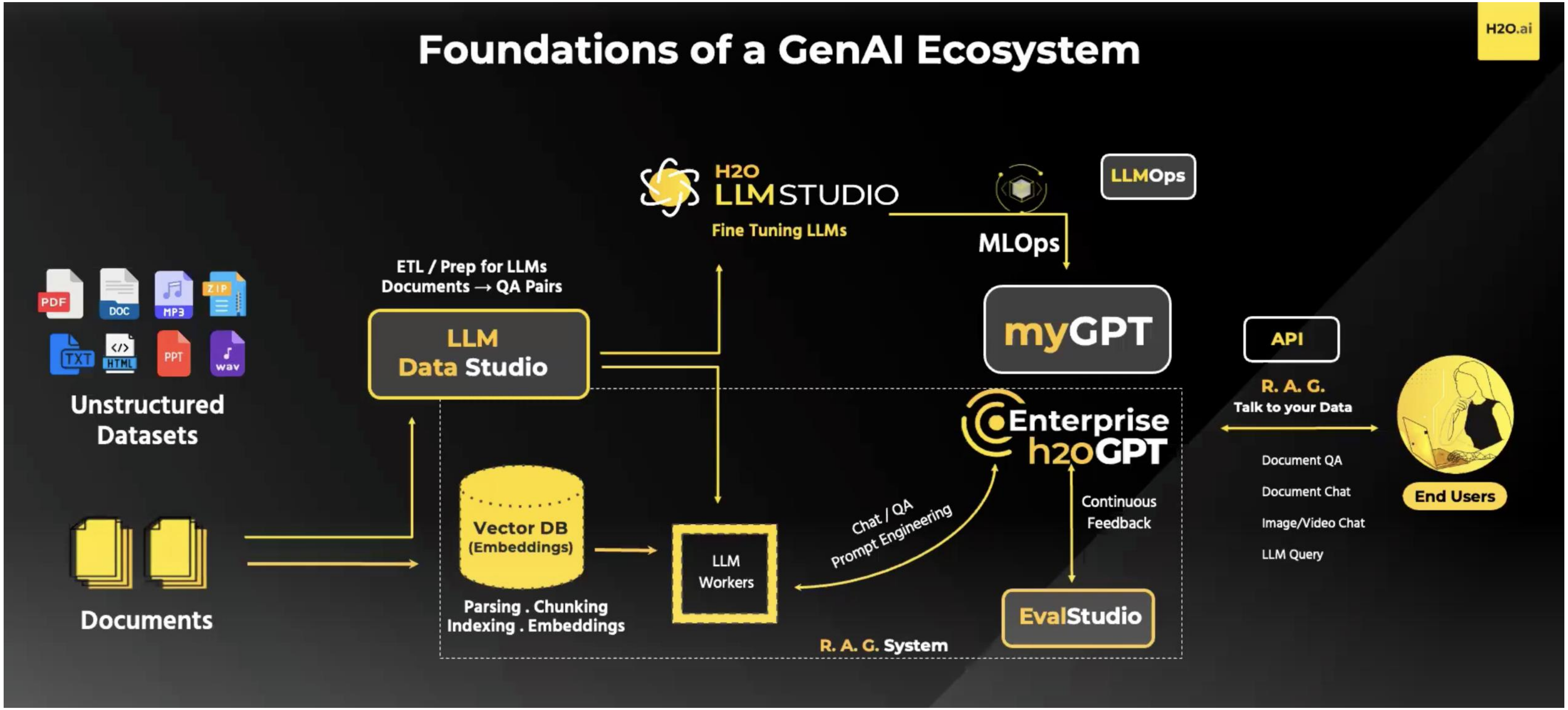
H2O Studio allows for LLM fine-tuning using a no-code Graphical User Interface (GUI). It can be installed locally or accessed via H2O AI Cloud (HAIC).

Krishna Damarla

October 6, 2025

Reference Architecture

The below diagram illustrates the core components and data flow within the H2O LLM Studio environment.

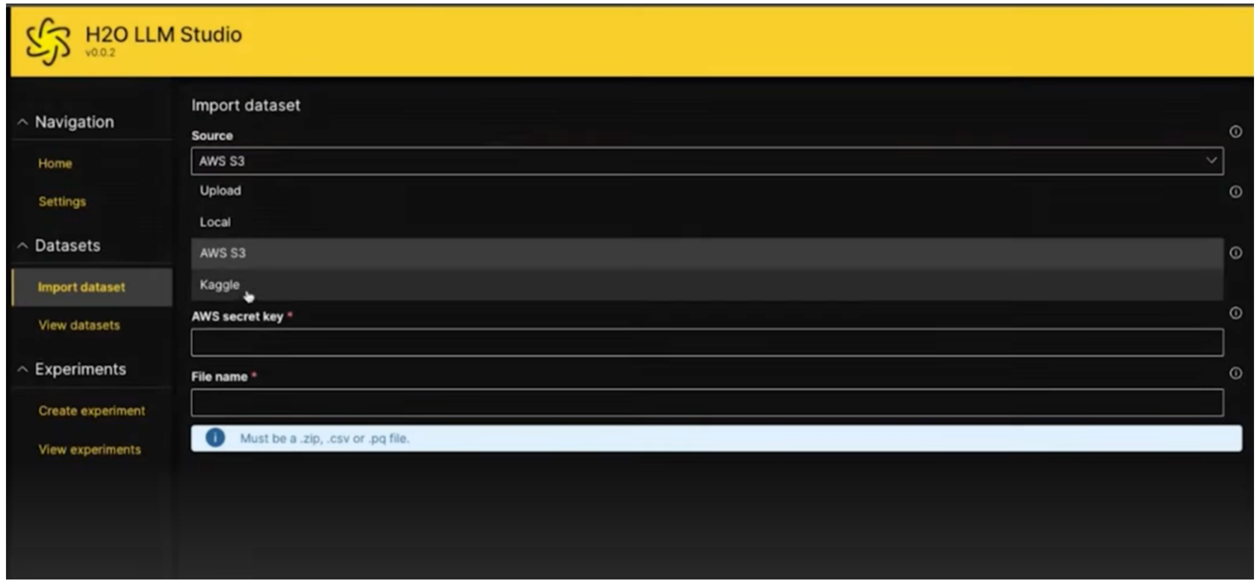


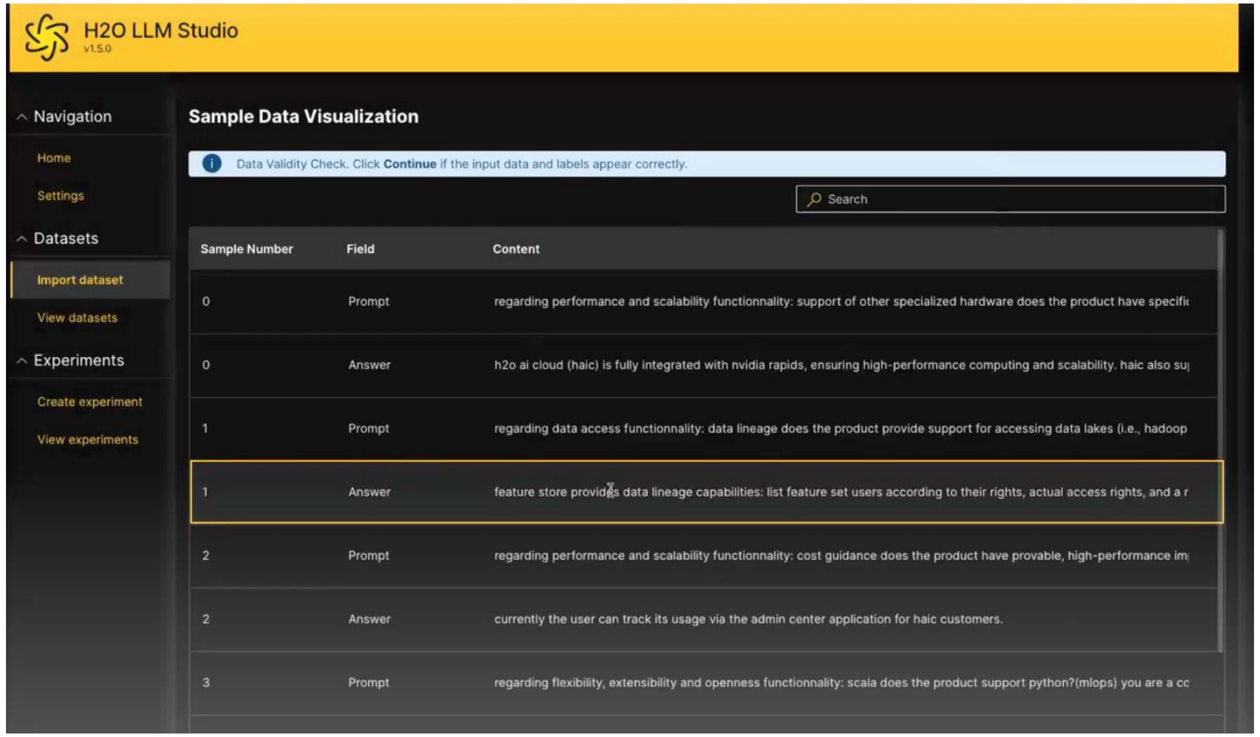
Key Steps in H2O LLM Studio

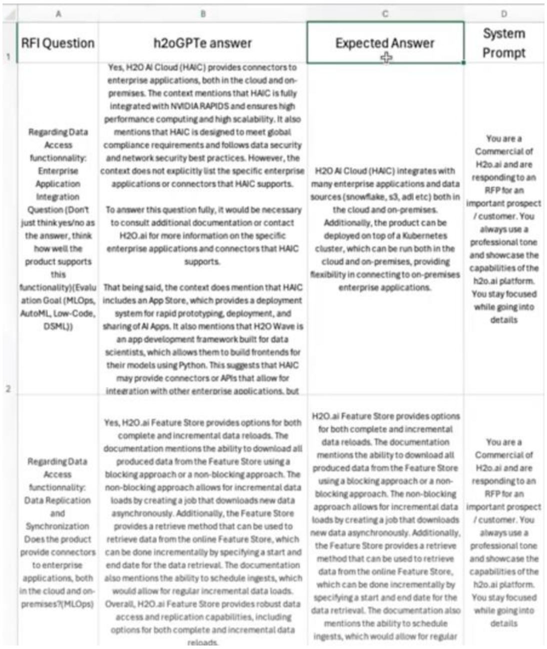
1. Import Dataset
2. View the loaded dataset
3. Fine-tune the model
4. Monitor model health
5. Test the Model
6. Push checkpoint to HuggingFace
7. Deploy via H2O MLOps

Step 1: Import Dataset

From AWS S3 Bucket / Kaggle / Local Dataset







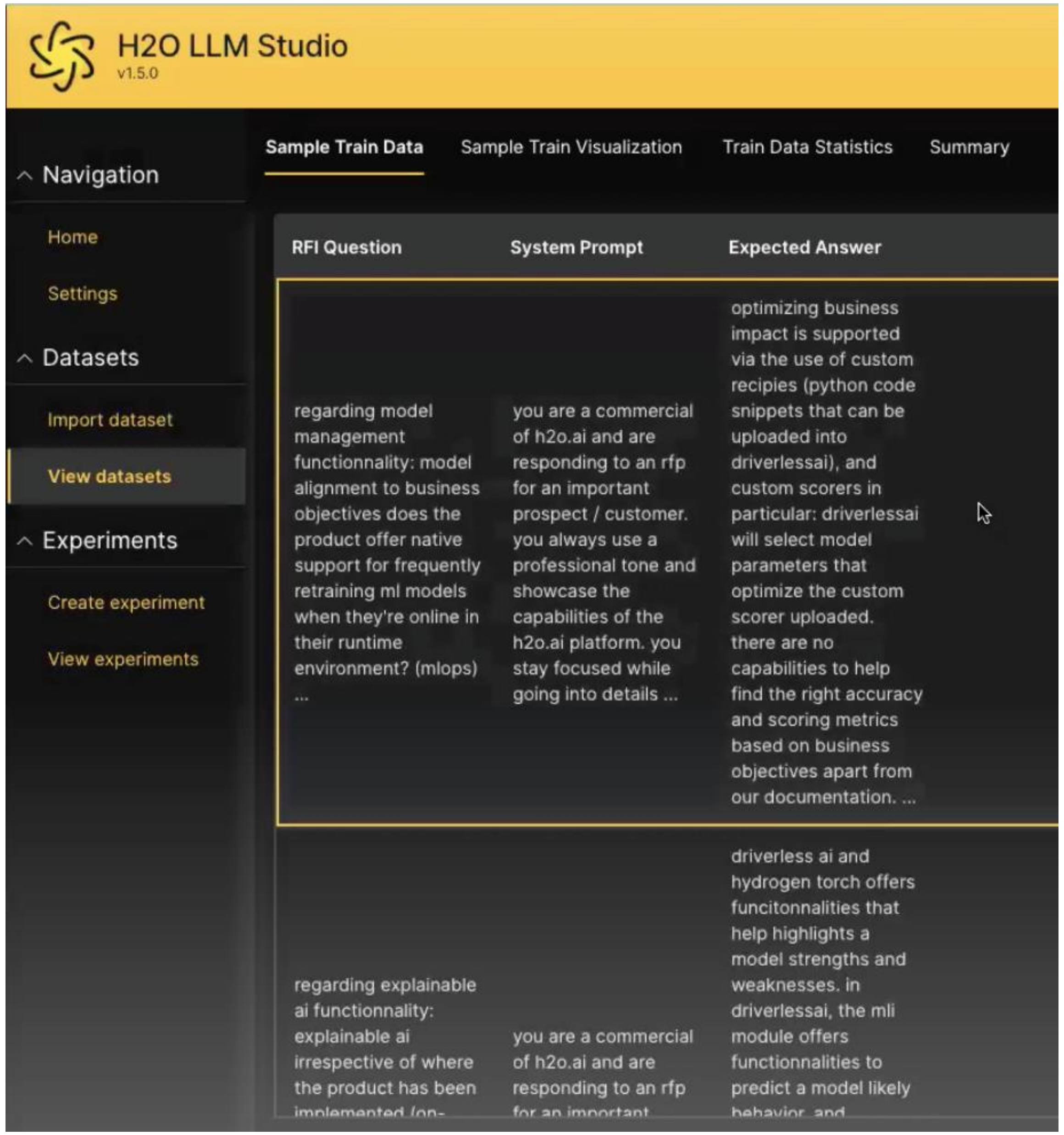
Step 2: View the Loaded Dataset

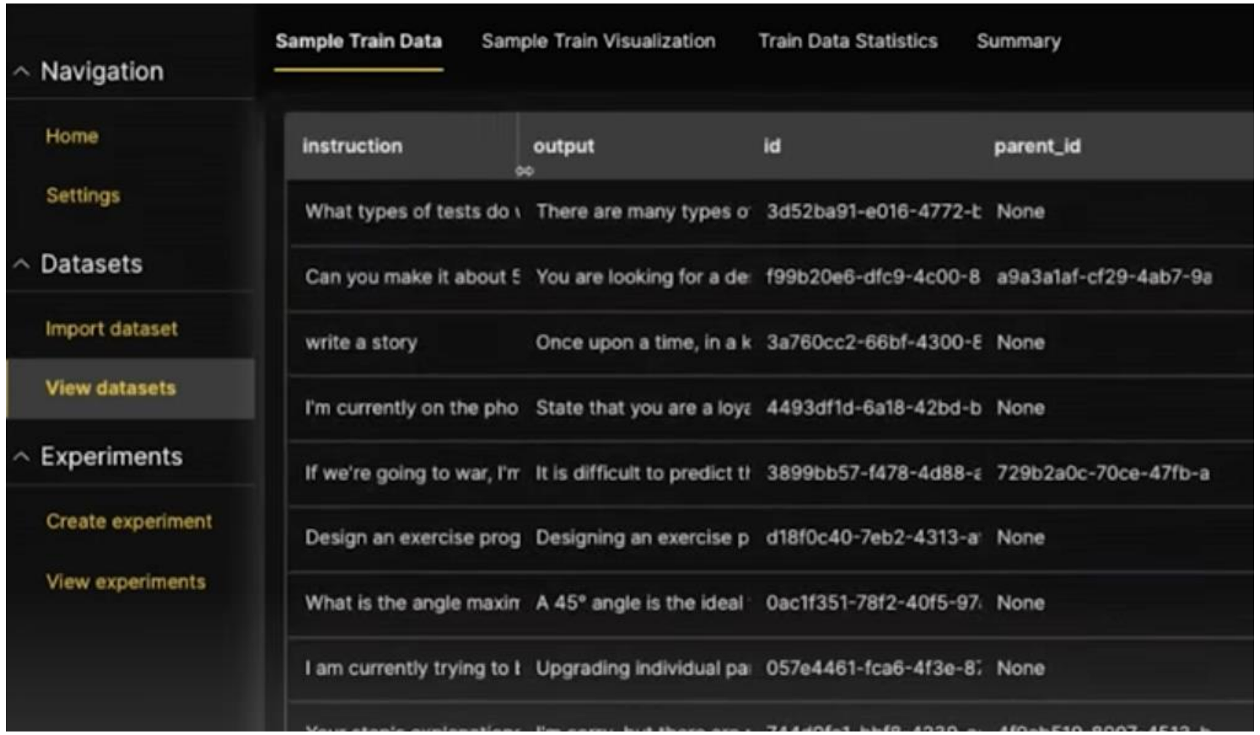
Example 1: RFI Dataset

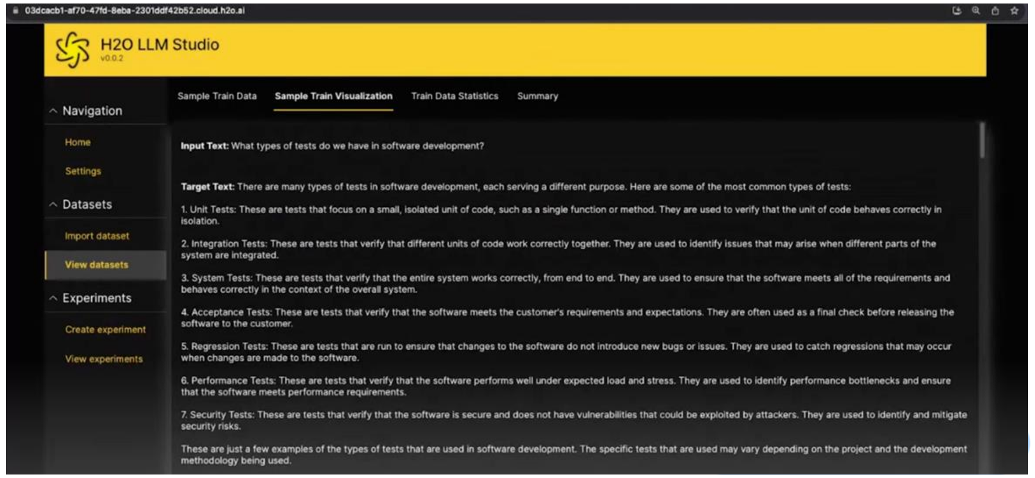
Example 2: Open Assistant Dataset

After importing, the studio provides tools to view the loaded dataset, examine its structure, columns, and initial rows.

This step is crucial for understanding the data quality and content before proceeding with fine-tuning.







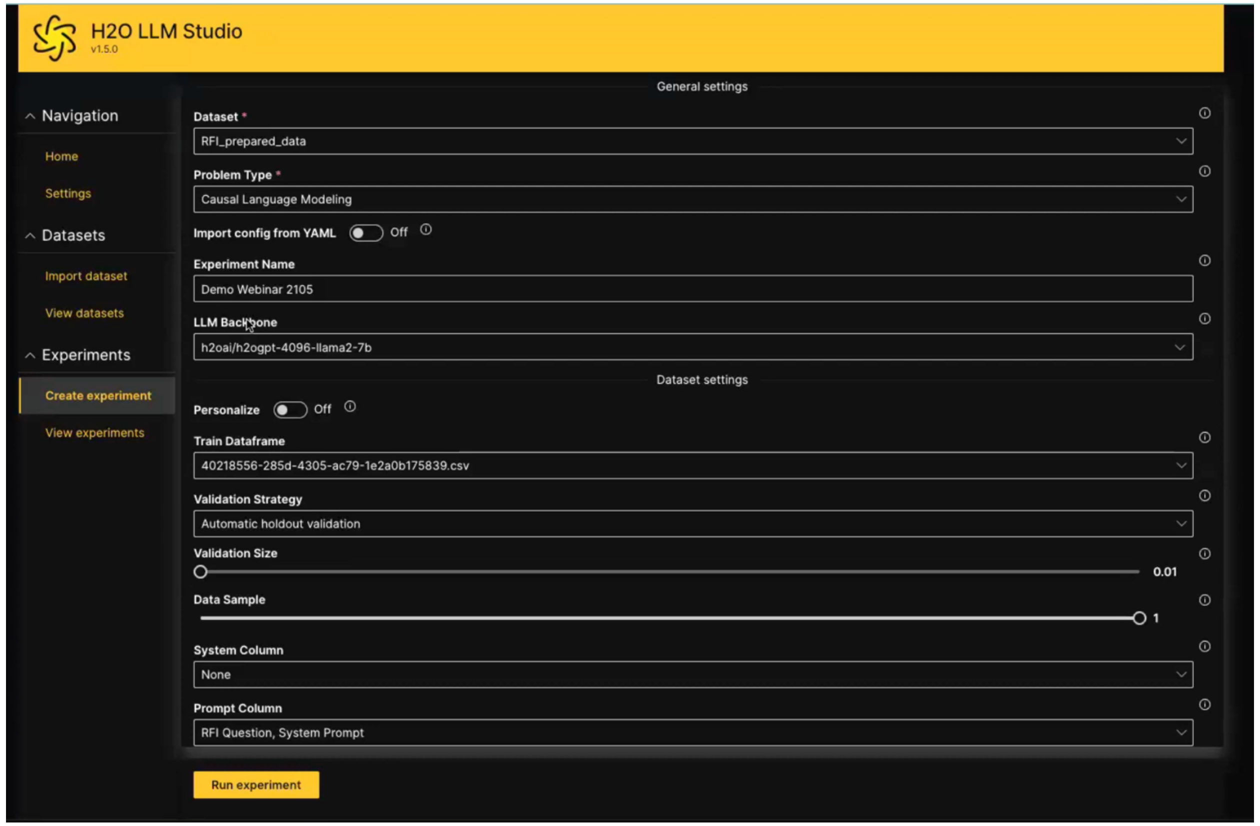
Step 3: Fine-tune the Model

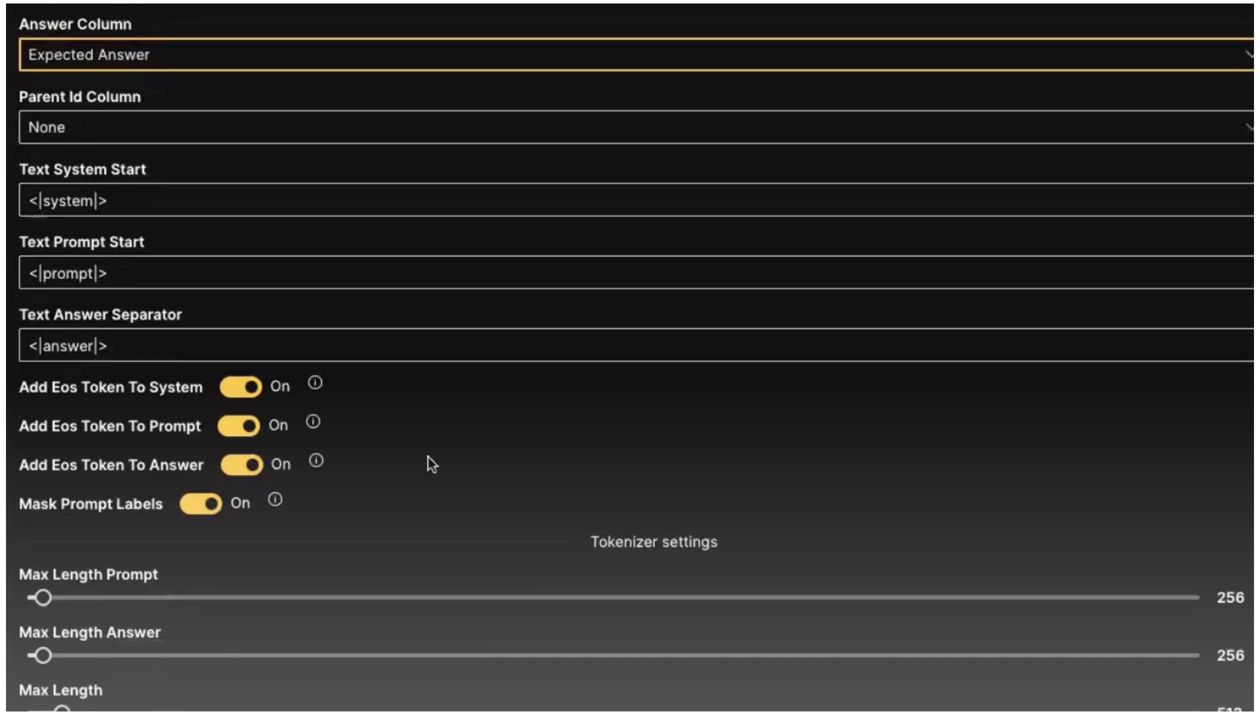
This step involves configuring various parameters and settings to fine-tune your chosen Large Language Model.

The no-code GUI simplifies this complex process, allowing users to specify training details without writing code.

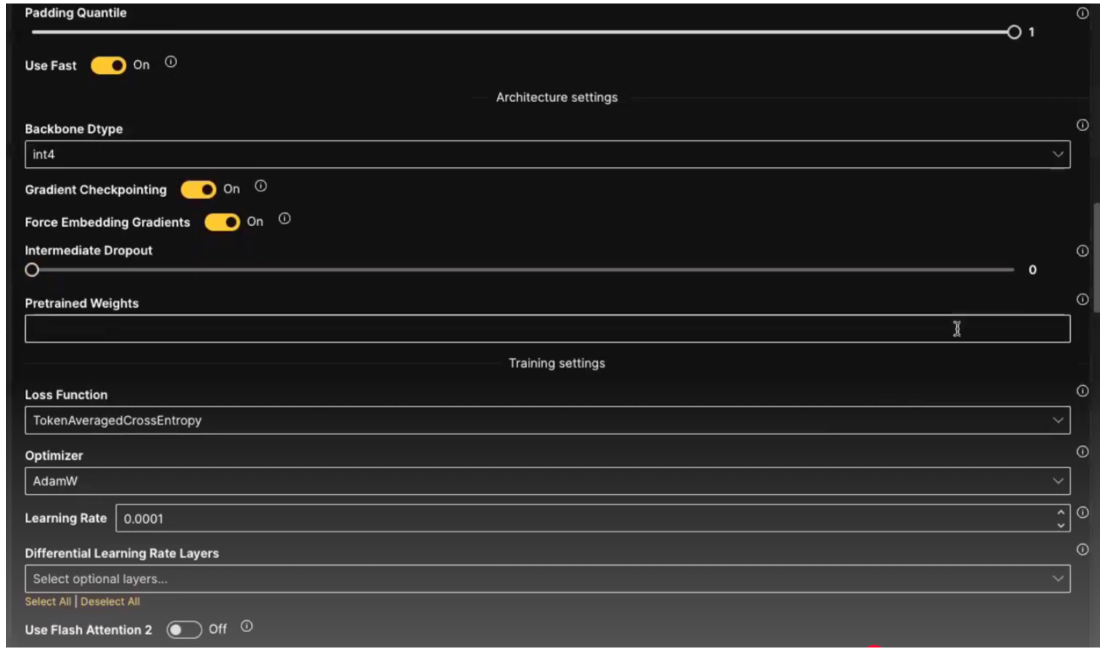
Users can easily adjust critical parameters such as learning rate, number of epochs, batch size, and select the specific base model for fine-tuning.

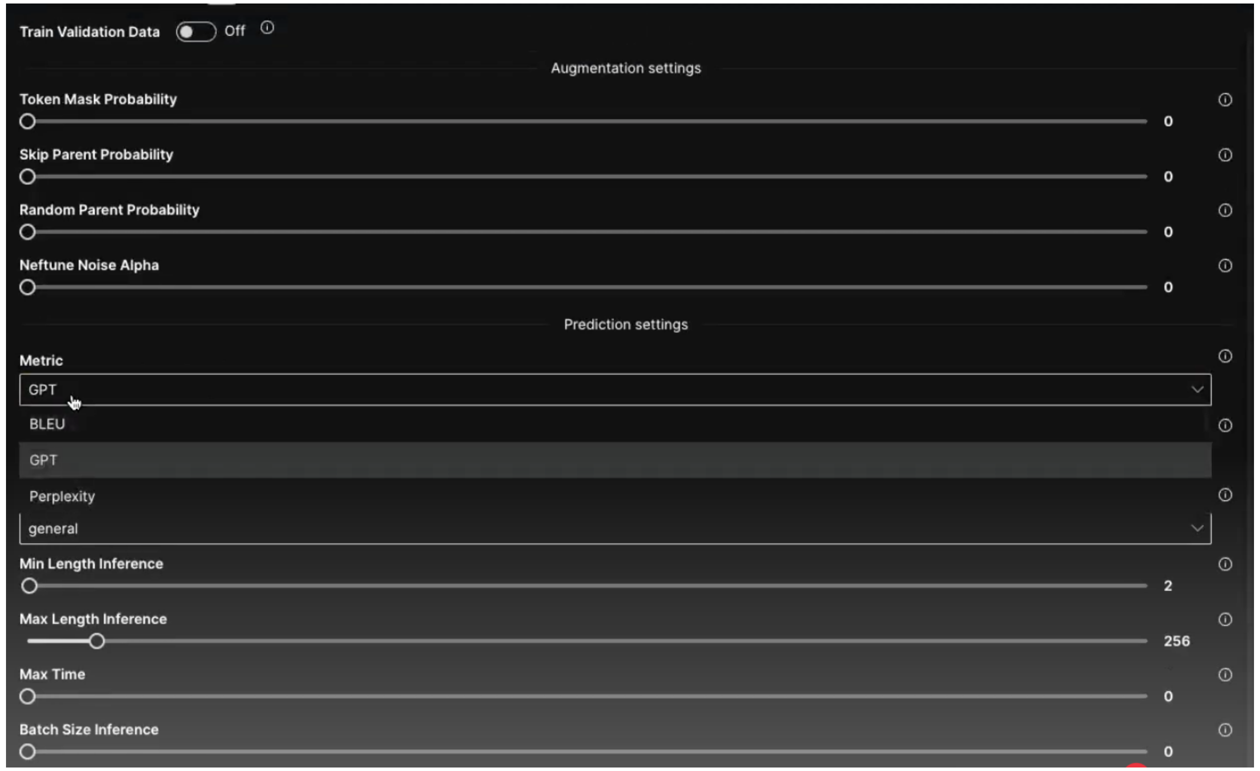
The interface also allows for defining evaluation metrics and saving preferences.

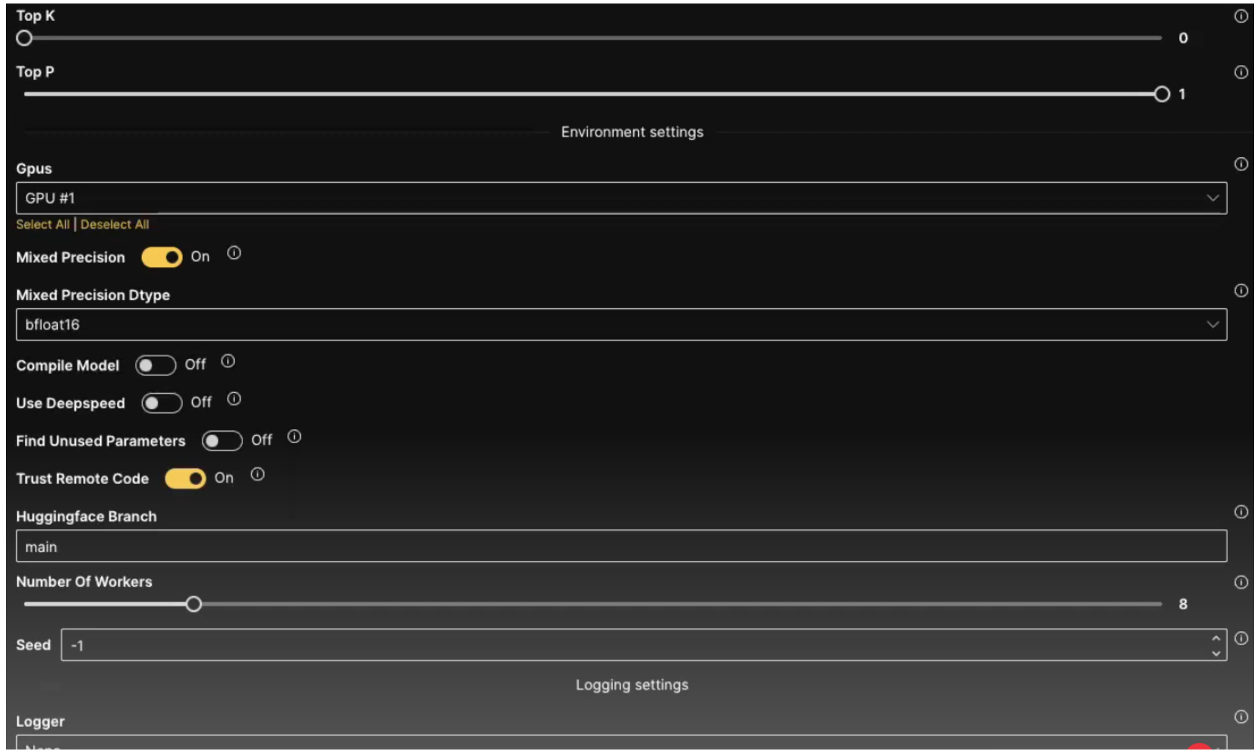




Step 3: Fine-tune the Model



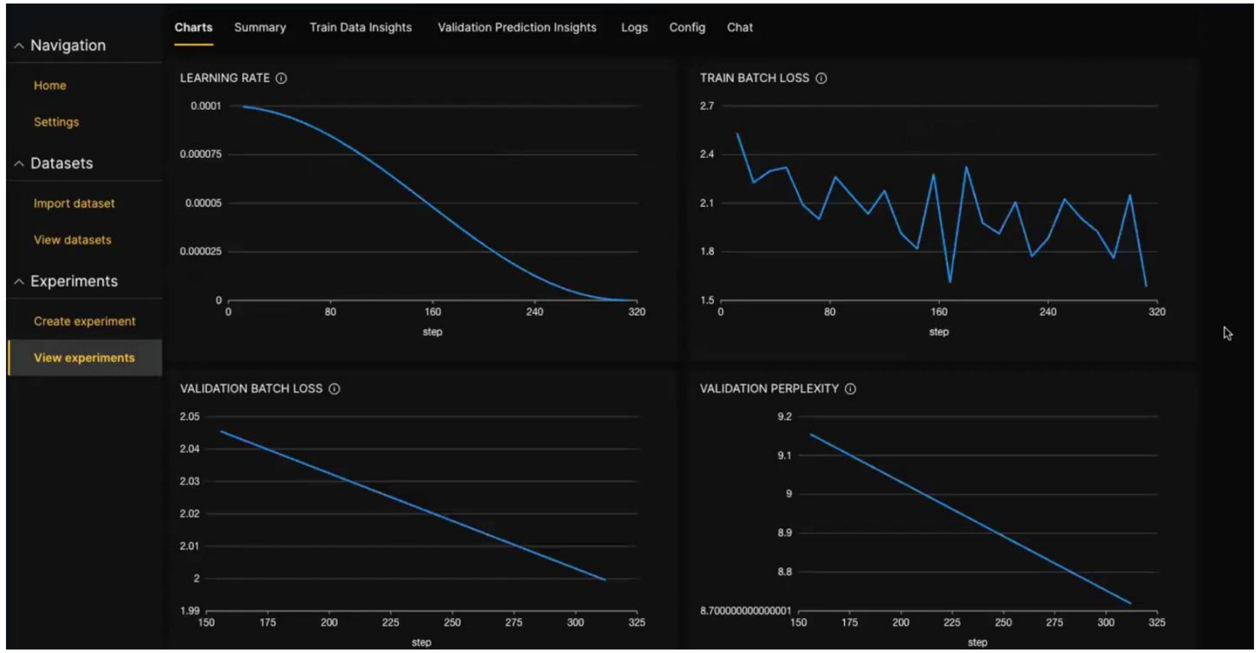


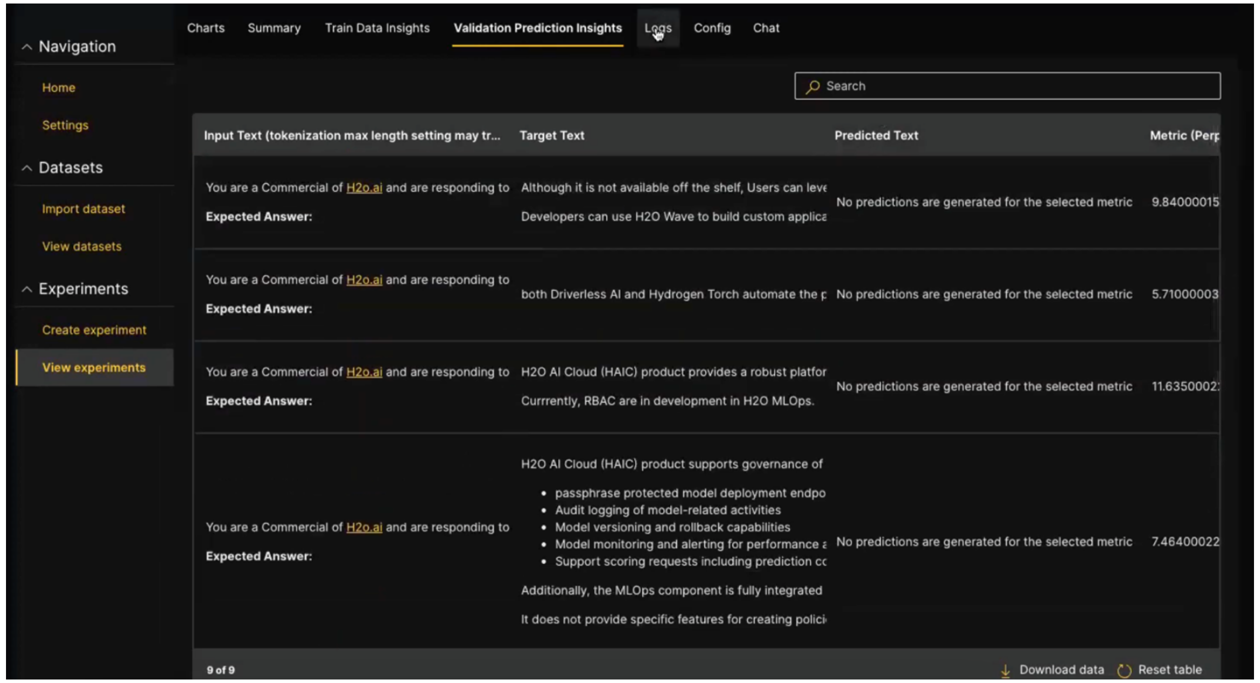


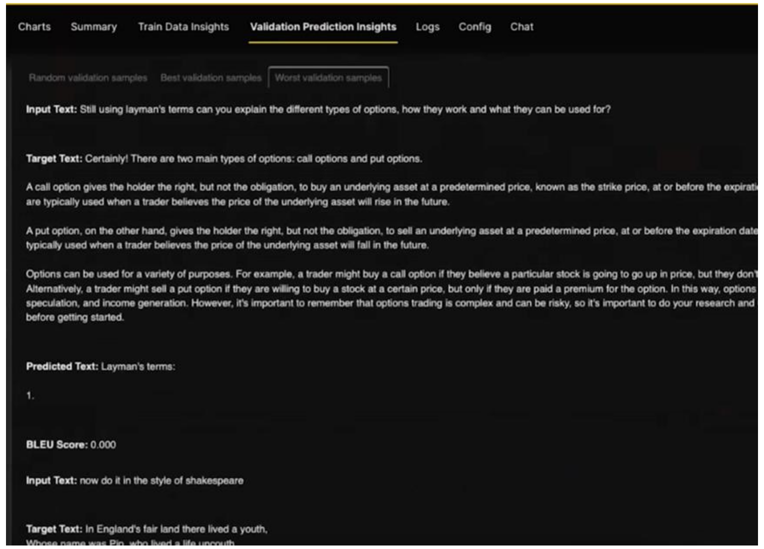
Step 4: Monitor Model Health

After fine-tuning, it's crucial to monitor the model's health and performance. This involves reviewing training progress and analyzing key performance metrics to ensure the model is learning effectively and performing as expected.

This step allows users to identify any issues early, such as overfitting or underfitting, and make necessary adjustments to the fine-tuning process.





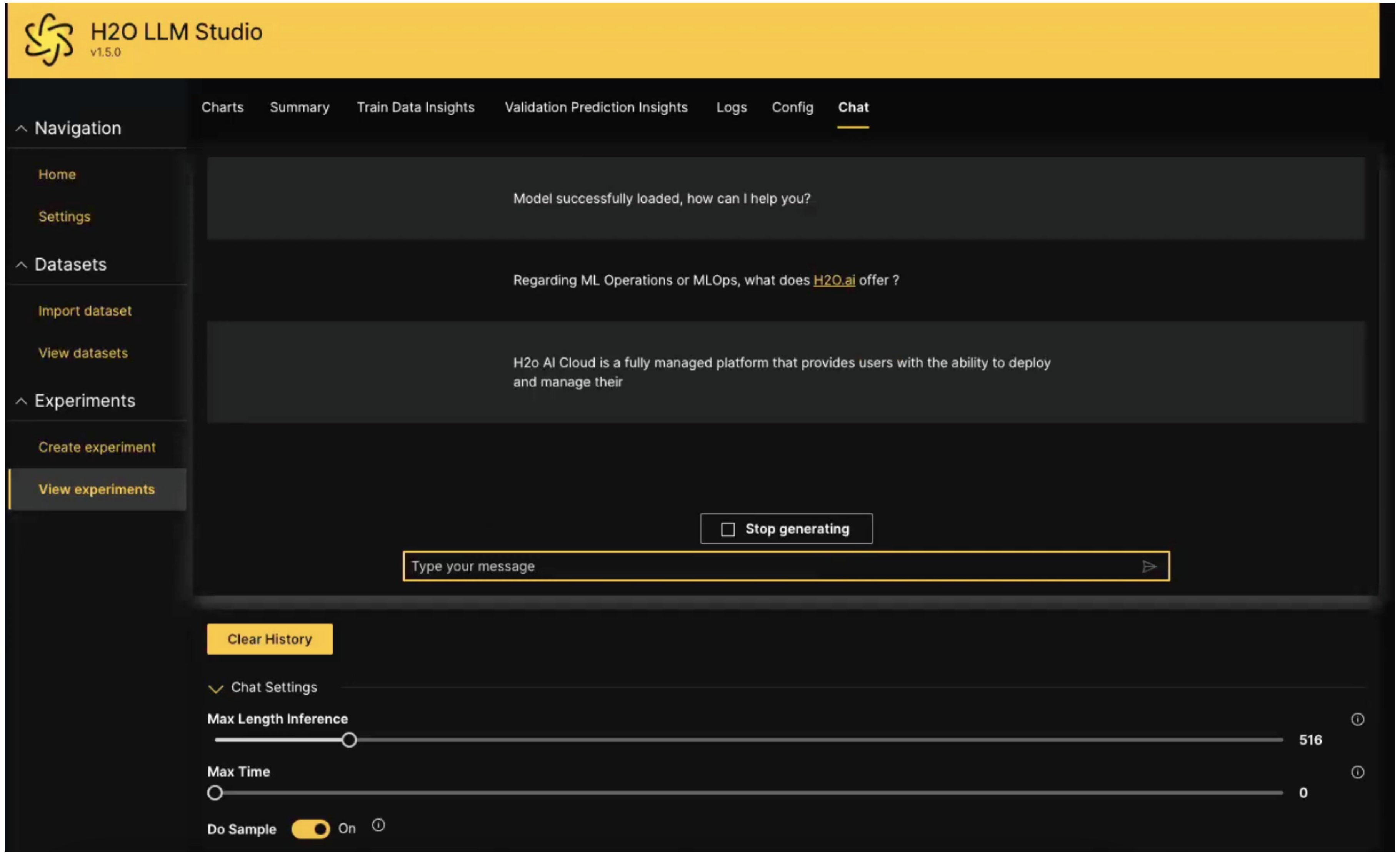


Step 5: Test the Model

Interactive Chat Window

Once the model is fine-tuned, you can interact with it directly through an interactive chat window.

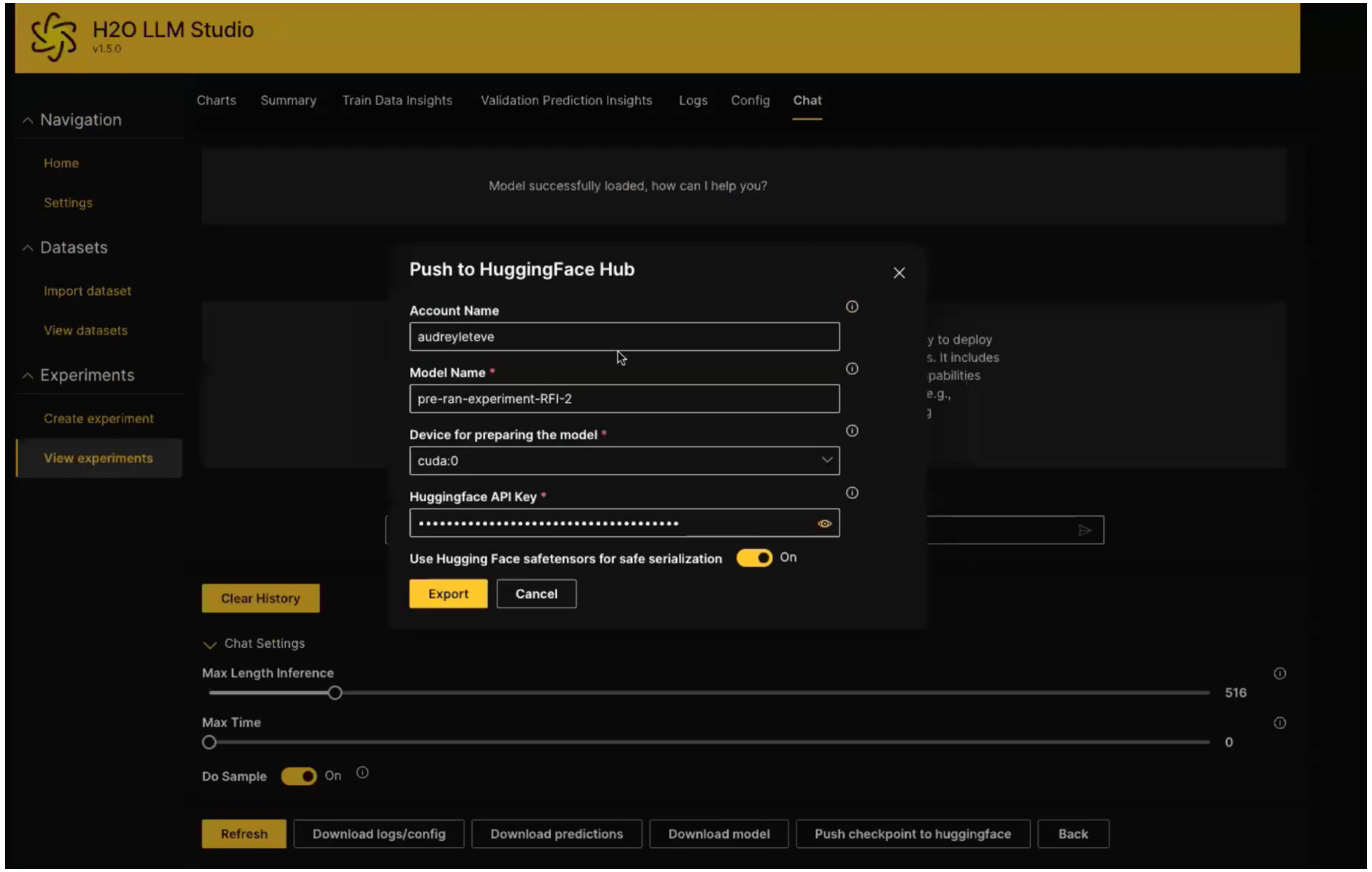
This allows for immediate testing of the model's responses to various prompts and scenarios, ensuring it behaves as expected in a conversational context.



This interactive testing environment provides a practical way to evaluate the model's performance, identify any areas for further improvement, and validate its readiness for deployment.

Step 6: Push to HuggingFace

Deploy the fine-tuned model checkpoint to HuggingFace Hub for sharing and distribution.

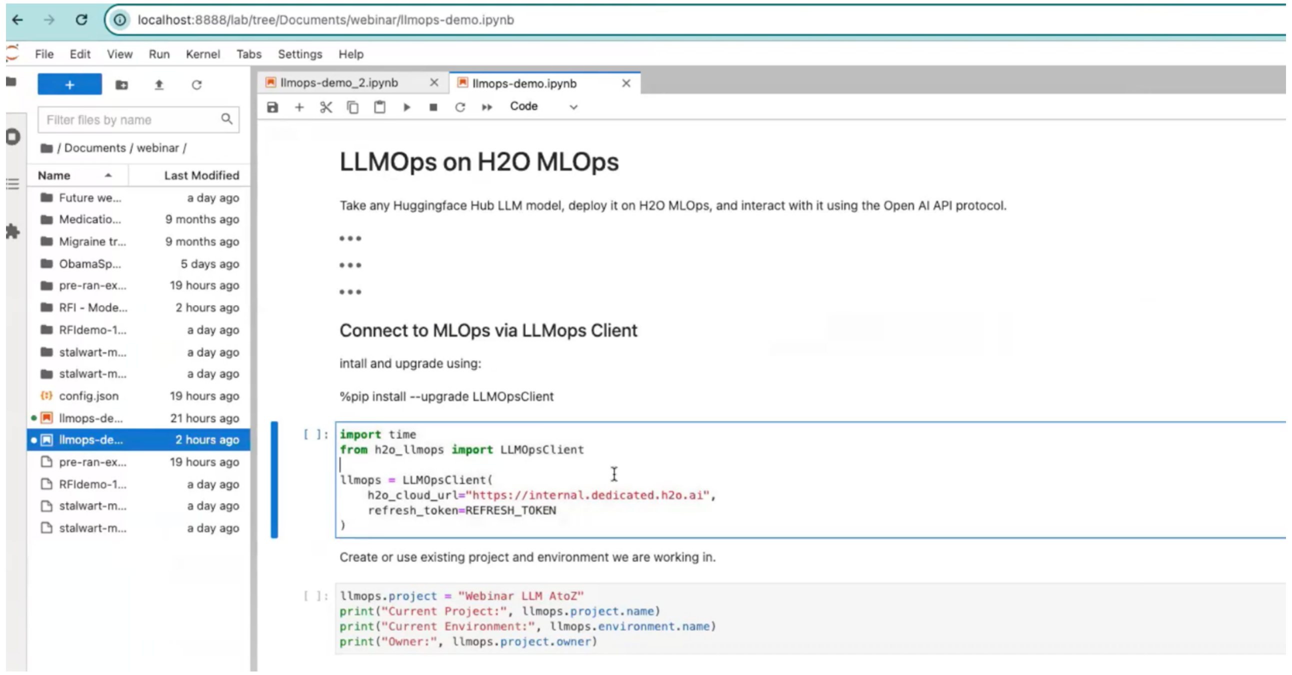


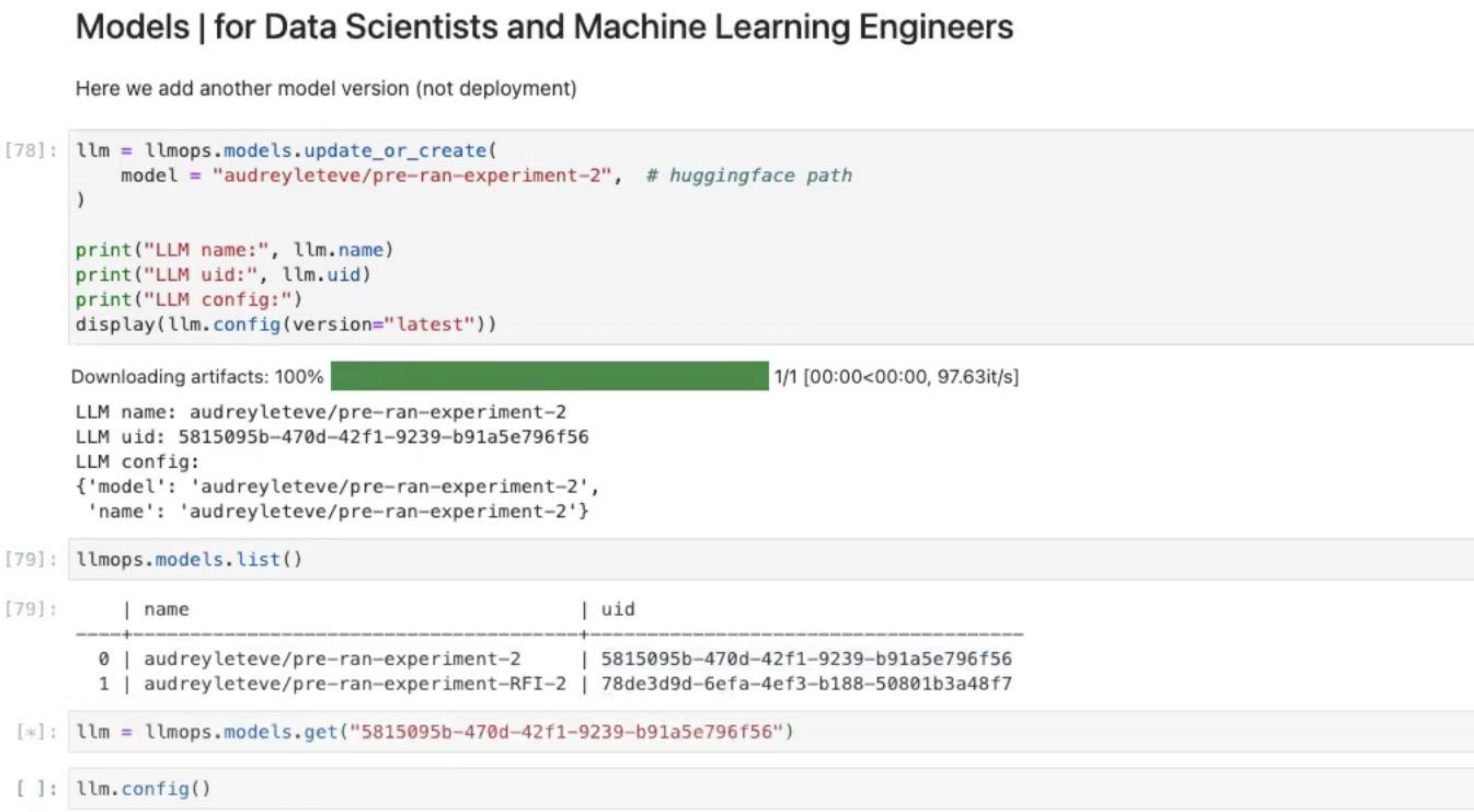
Login to your HuggingFace account & check the deployed model

Step 7: H2O MLOps Integration

After successful fine-tuning and validation, the model can be seamlessly integrated into H2O MLOps for production deployment.

Below code snippet describes how to interact with OpenAPI via H2O MLOps to manage, monitor, and scale your LLM in a production environment.
environment.





Step 7: H2O MLOps Integration

H2O MLOps provides comprehensive tools for model lifecycle management, allowing for robust version control, A/B testing, and continuous performance monitoring of your deployed LLMs.

Deployments | for Machine Learning Engineers

Deployments served by vLLM on H2O MLOps.

https://github.com/vllm-project/vllm

Existing deployments can be replaced without changing the endpoint URL, which is automatically generated based on the model name.

```
[ ]: llmd = llmops.deployments.replace_or_create(
         model=llm,
         gpu_type = "nvidia.com/gpu",
         gpu_count=1,
         replicas=1,
         passphrase=passphrase
     )
```

Let's use the deployment of the same model

```
[82]: llmops.deployments.list()

[82]:     | name                                  | mode          | uid
      ----+---------------------------------------+---------------+--------------------------------------
        0 | audreyleteve/pre-ran-experiment-RFI-2 | Single Model  | de0180e7-b989-4da3-8da7-fda1af491c3c

[*]: llmd = llmops.deployments.list()[0]

[ ]: print("Deployment name:", llmd.name)
     print("Deployment uid:", llmd.uid)

     print("Deployment status:", llmd.status())
```

OpenAI API Protocol | for LLM Consumers

The deployment exposes a basic OpenAI API endpoint for easy drop in to existing patterns. https://platform.openai.com/docs/api-reference

```
[85]: from openai import OpenAI

      client = OpenAI(
          api_key=passphrase,
          base_url=llmd.url_for_scoring,
      )
```

View available model.

```
[86]: model = client.models.list().data[0].id
```

Stream completion - based on RFP related questions :

```
[ ]: question = "Regarding model deployment mode, what options does the H2o platform offer to the user?"
     # question = "Regarding ML Operations or MLOps, what does H2O.ai offer ?"
     # question = "regarding hyperparameters optimisation, what does the H2O.ai platform offer to the user?"
     # question = "What are H2O MLOps main functionalities ?"
     # question = "Does H2o MLOps platform integrate with third party mlops platform?"

[ ]: system_prompt  = "You are a Commercial of H2o.ai and are responding to an RFP for an important prospect / customer. You always use a professional tone and showcase the capabilities of the h

[ ]: completion = client.completions.create(
         model=model,
         prompt= system_prompt + question,
         temperature=0.09,
         stream=True,
         max_tokens = 280,
         frequency_penalty = 0.25,
         presence_penalty = 0.15,
     )

     for c in completion:
         print(c.choices[0].text, sep="", end="")
```

```
H2o.ai offers a wide range of deployment options for its platform, including:
Cloud-based: H2o.ai provides a cloud-based deployment option that allows users to access the platform through a web browser or API. This option is ideal for organizations that do not have t
he resources or expertise to manage their own infrastructure.
On-premises: H2o.ai also offers an on-premises deployment option that allows organizations to install the platform on their own servers and manage it themselves. This option is ideal for or
ganizations that have strict security requirements or need to maintain control over their data.
Hybrid: H2o.ai also supports a hybrid deployment model that allows organizations to use both cloud and on-premises resources to deploy their models. This option provides the best of both wo
rlds, allowing organizations to take advantage of the scalability and flexibility of the cloud while maintaining control over their data and infrastructure.
In addition to these deployment options, H2o.ai also offers a range of services and support options to help users get the most out of their platform, including training, consulting, and man
aged services.
```

H2O LLM Data Studio

Data Preparation for LLMs Fine-tuning with No-Code GUI

Data Preparation Best Practices

The CRISP-DM (Cross-Industry Standard Process for Data Mining) model highlights the critical importance of proper data curation. Following its structured approach, along with 9 fundamental laws of data mining, ensures robust and reliable insights from your data.

Phase 1: Business Understanding

Translating business needs into data mining objectives. This phase underscores **Law 1: Business goals drive everything**, ensuring our efforts are strategically aligned from the outset. It also highlights **Law 2: We need business knowledge at every step**, as domain expertise is crucial for setting relevant goals.

Phase 2: Data Understanding & Exploration

Collecting, describing, and exploring financial transaction data. Through thorough exploration, we acknowledge **Law 5: Patterns Law 5: Patterns always exist in the data**, even if hidden. This initial step also leverages **Law 6: Data mining helps you see better, you see better**, providing early insights into data characteristics through exploratory data analysis.

Phase 3: Data Preparation & Feature Engineering

Cleaning, transforming, and constructing datasets for modeling. This often resource-intensive phase directly illustrates **Law 3: Data** illustrates **Law 3: Data preparation takes the most time**, emphasizing its critical role in successful model development. development.

Phase 4: Modeling & Algorithm Selection

Choosing and applying appropriate machine learning models. Here, we embrace **Law 4: Find the best model through through experimentation**, iteratively testing various approaches. Furthermore, this stage often involves generating new insights, generating new insights, aligning with **Law 7: Prediction adds new information**.

07

Phase 5: Evaluation & Performance Metrics

Assessing model effectiveness and identifying actionable insights. This phase directly ties into **Law 8: Value comes from business from business impact not just accuracy**, ensuring our metrics reflect real-world benefits. It also reinforces the iterative nature, iterative nature, aligning with **Law 9: As data and business goals keep evolving so does the model**, as evaluation informs necessary informs necessary adjustments.

08

Phase 6: Deployment & Monitoring

Integrating the solution and ensuring continuous performance. This final phase completes the iterative cycle, embodying **Law 9: As data and business goals keep evolving so does the model**, as continuous monitoring ensures the model adapts to new data and changing business needs.

Phase 3: Data Preparation & Feature Engineering

Data Consolidation

The first step in data preparation is to integrate diverse datasets such as transaction logs, customer profiles, and device information into a single, unified repository. This ensures a comprehensive and coherent view, essential for robust analysis, truly highlighting why **(Law 3: Data preparation takes the most time)** as this foundational work is critical for all subsequent steps

Data Cleaning

Addressing the critical "Garbage In, Garbage Out" principle, this phase meticulously identifies meticulously identifies and rectifies errors, imputes missing values, detects and manages manages outliers, and resolves inconsistencies. This rigorous cleaning is paramount for paramount for ensuring integrity and reliability, clearly demonstrating why **(Law 3: Data 3: Data preparation takes the most time)** to get right, but is absolutely essential for reliable for reliable outcomes.

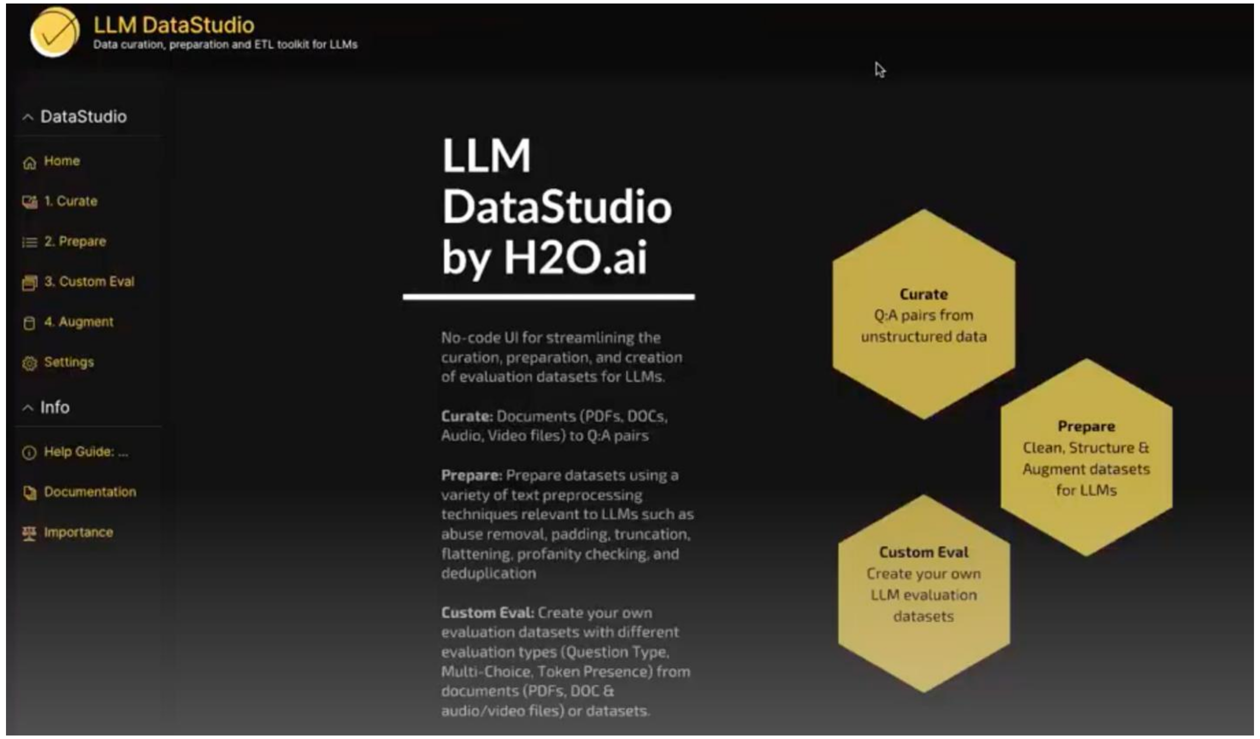
Data Transformation & Feature Engineering

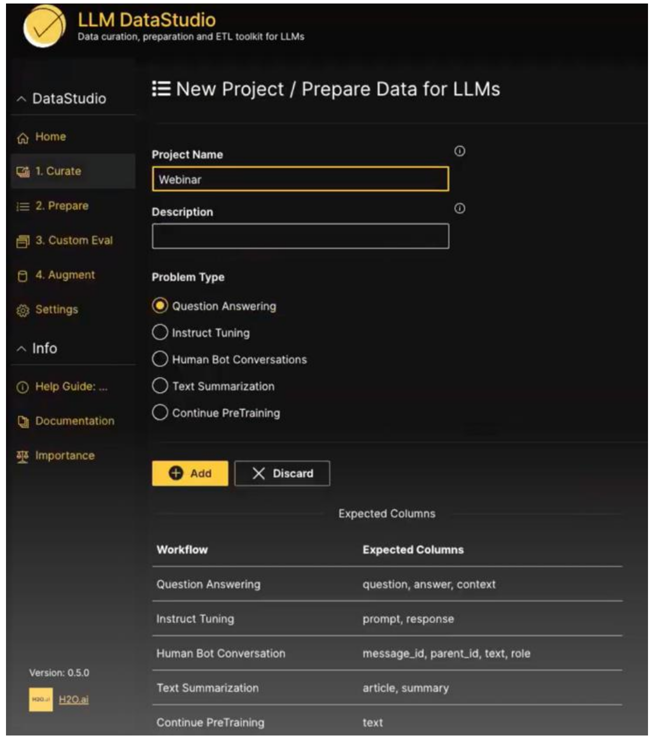
Raw data is rarely in an optimal format for modeling. This stage involves converting data converting data types, scaling numerical features, encoding categorical variables, and and crucially, engineering new features (e.g., 'frequency of transactions' or 'average spend'). 'average spend'). This labor-intensive process, aligning with **(Law 3: Data preparation takes preparation takes the most time)**, requires deep domain knowledge to create variables that variables that significantly enhance a model's predictive power

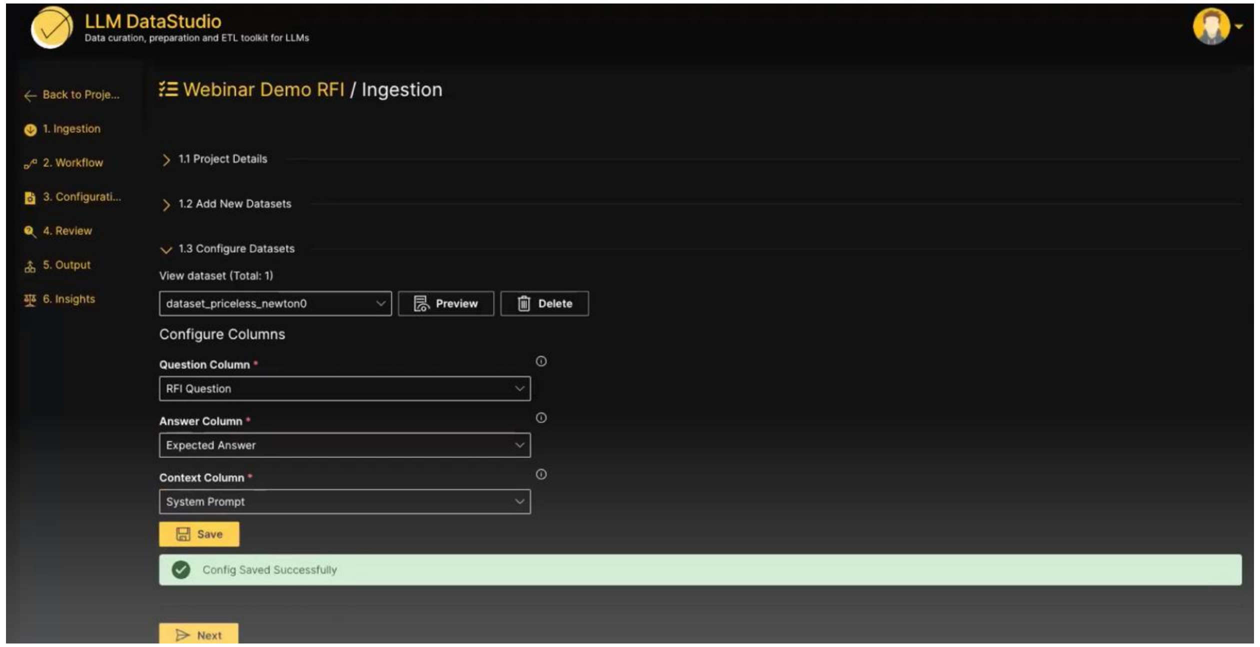
Data Reduction

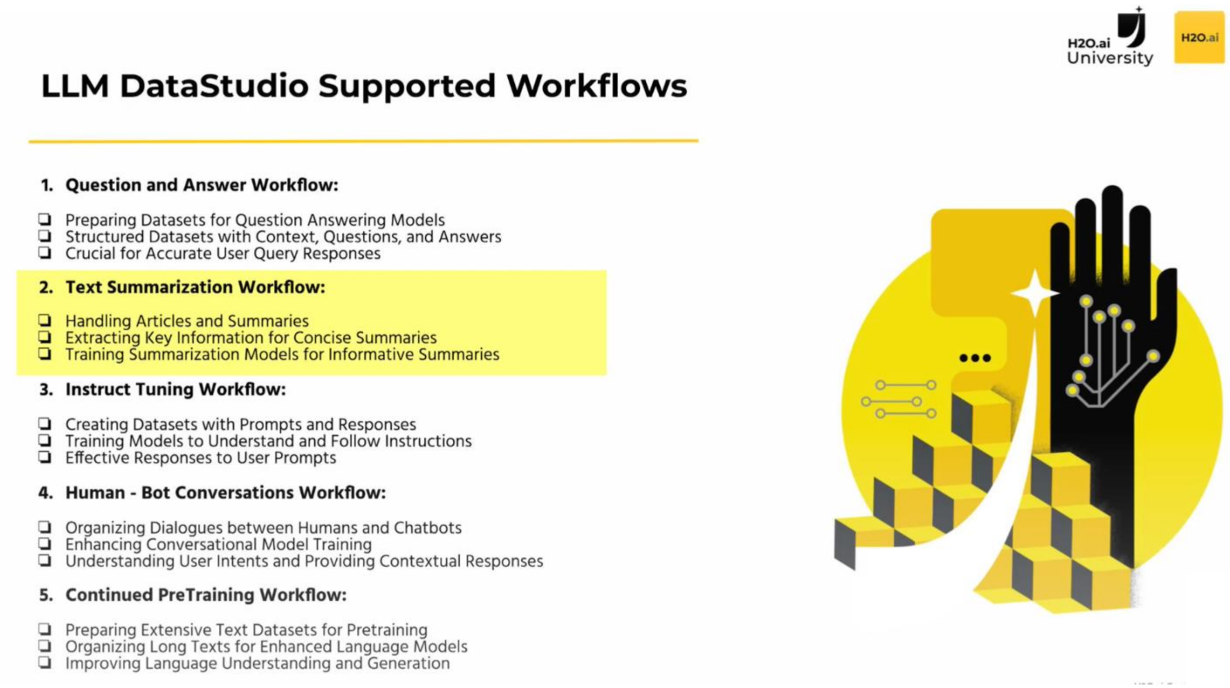
While more data can be beneficial, redundancy and irrelevant/repeated correlated features correlated features can hinder model performance. Data reduction employs techniques like techniques like Principal Component Analysis (PCA) to distill the dataset to its most most informative components. This meticulous streamlining, a key part of the phase where phase where **(Law 3: Data preparation takes the most time)**, focuses the model on key fraud on key fraud indicators, optimizing computation and mitigating overfitting.

Data Preparation in H20 LLM Data Studio

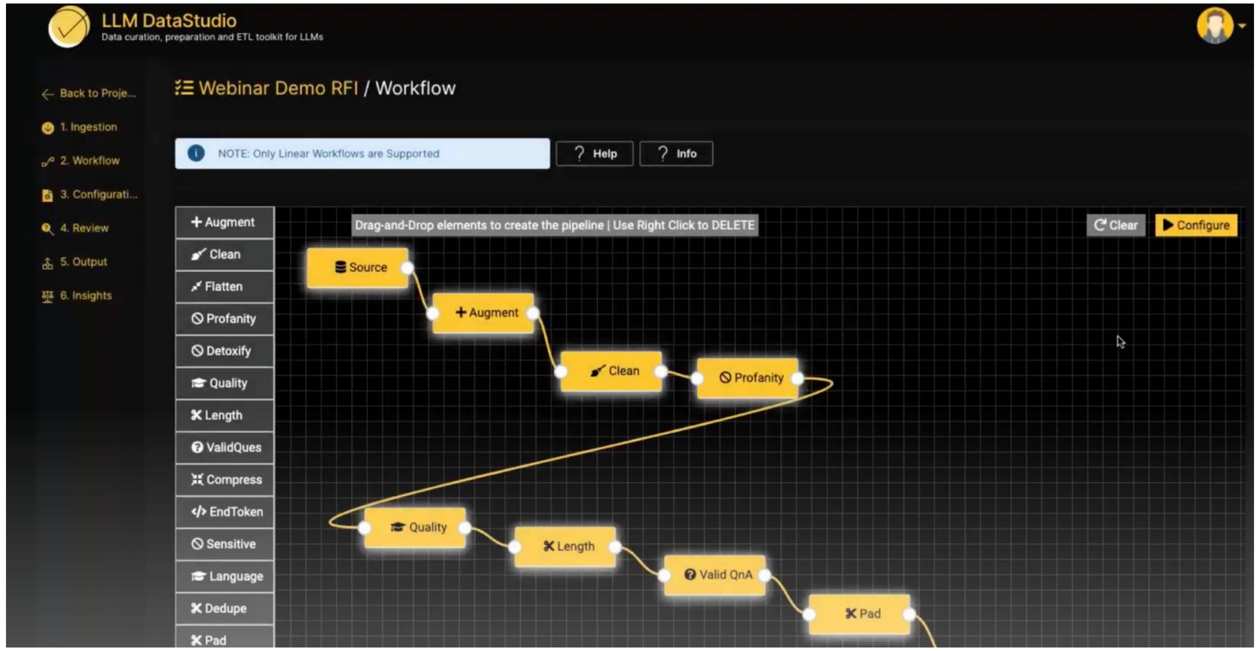


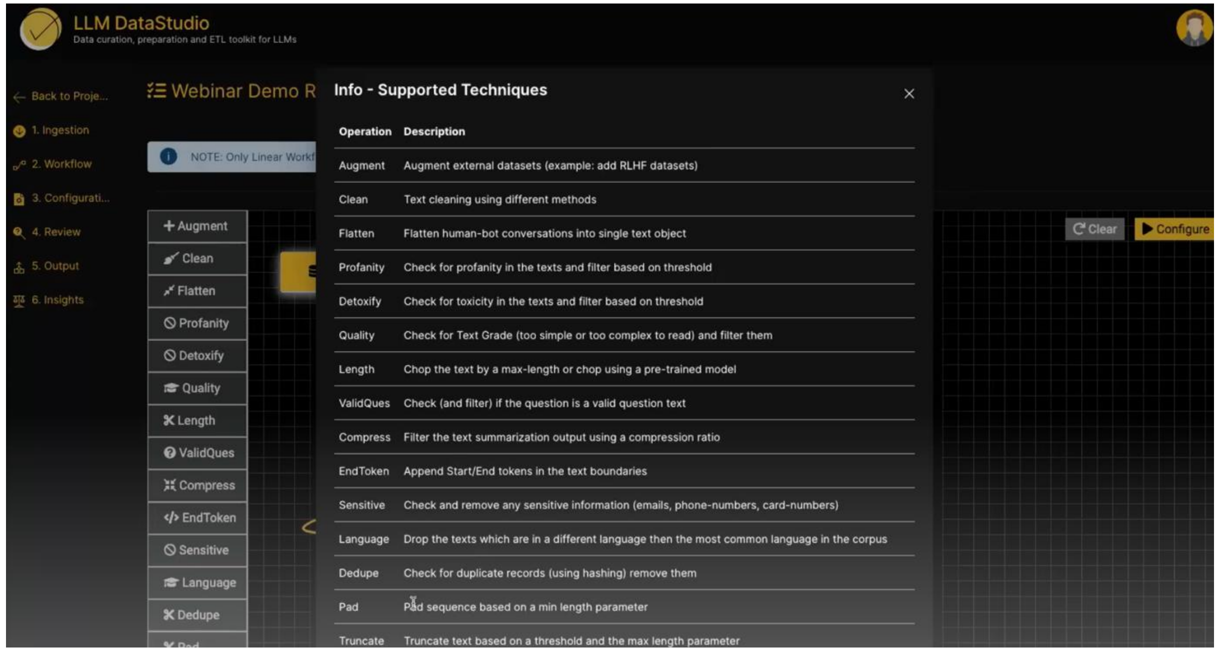


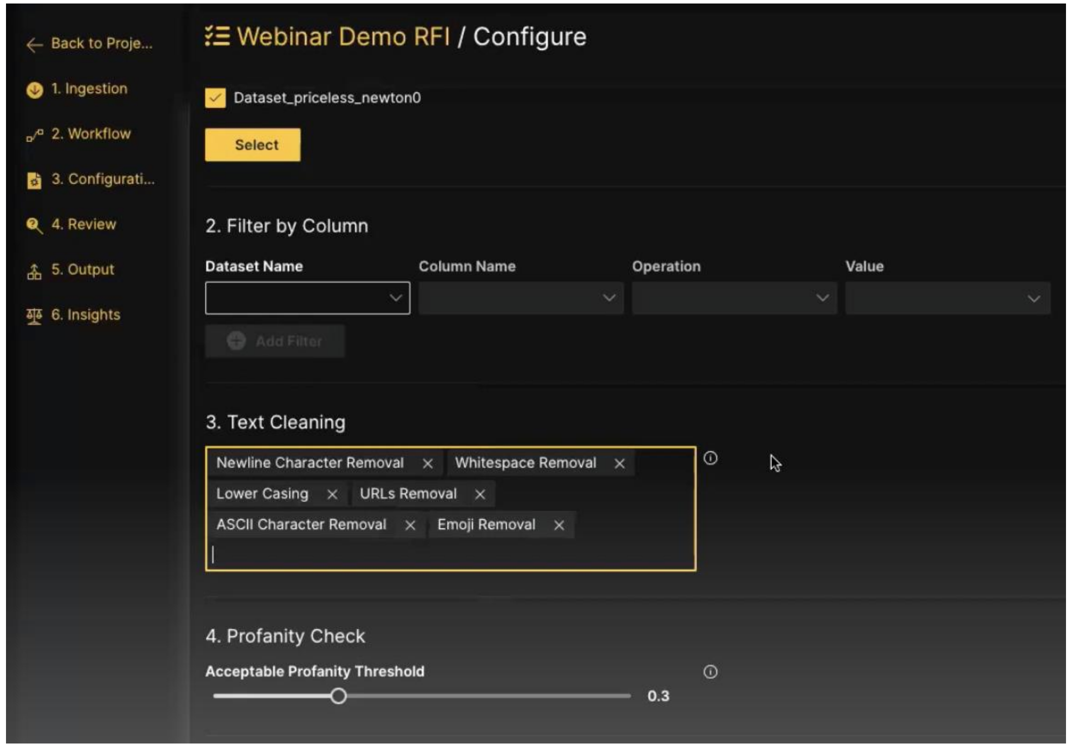


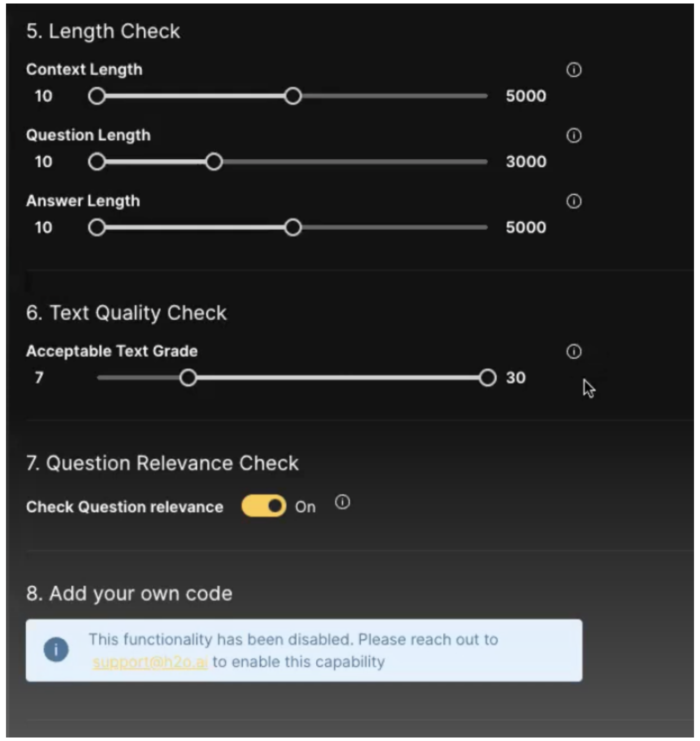


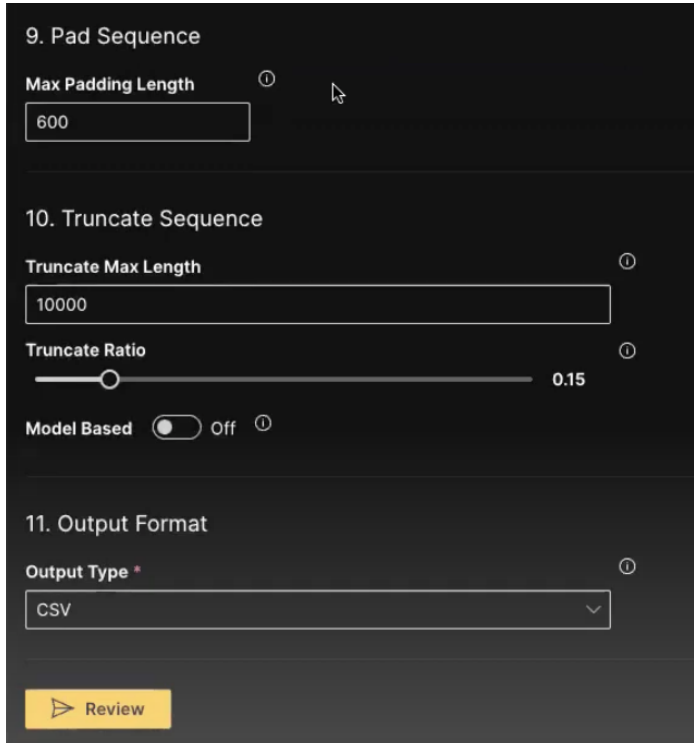
Data Preparation in H20 LLM Data Studio



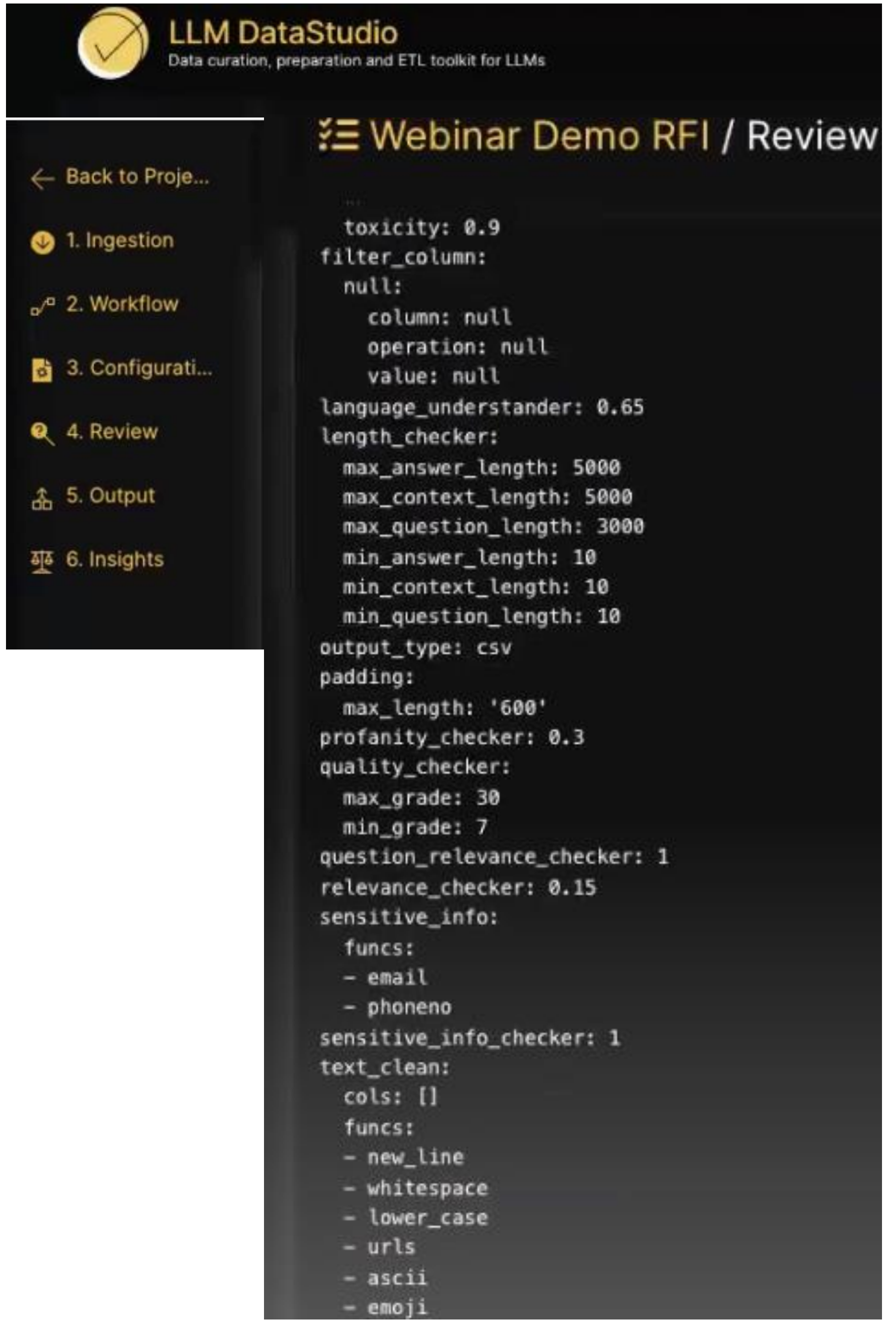


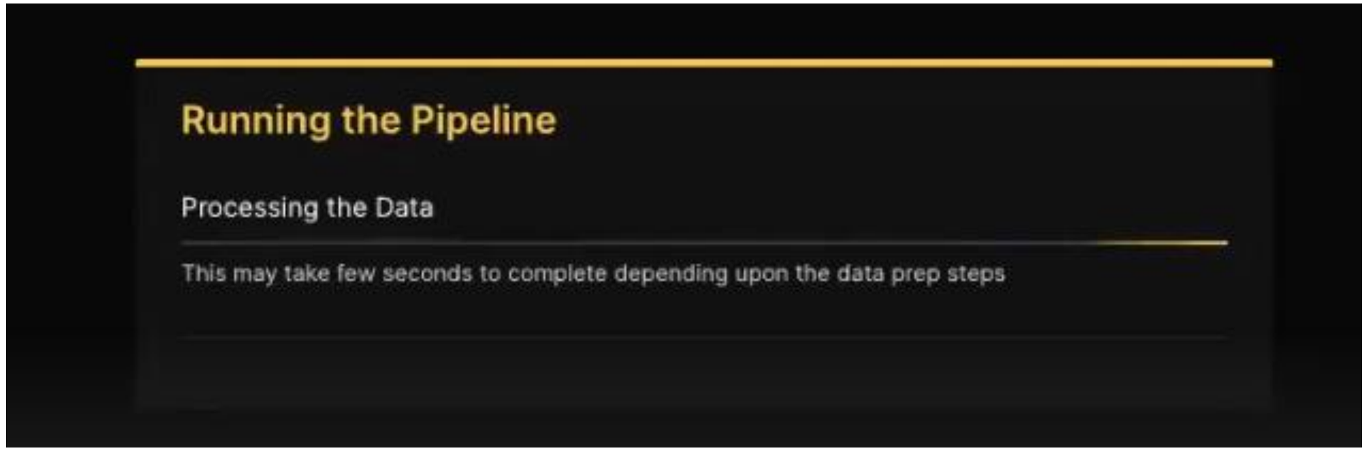




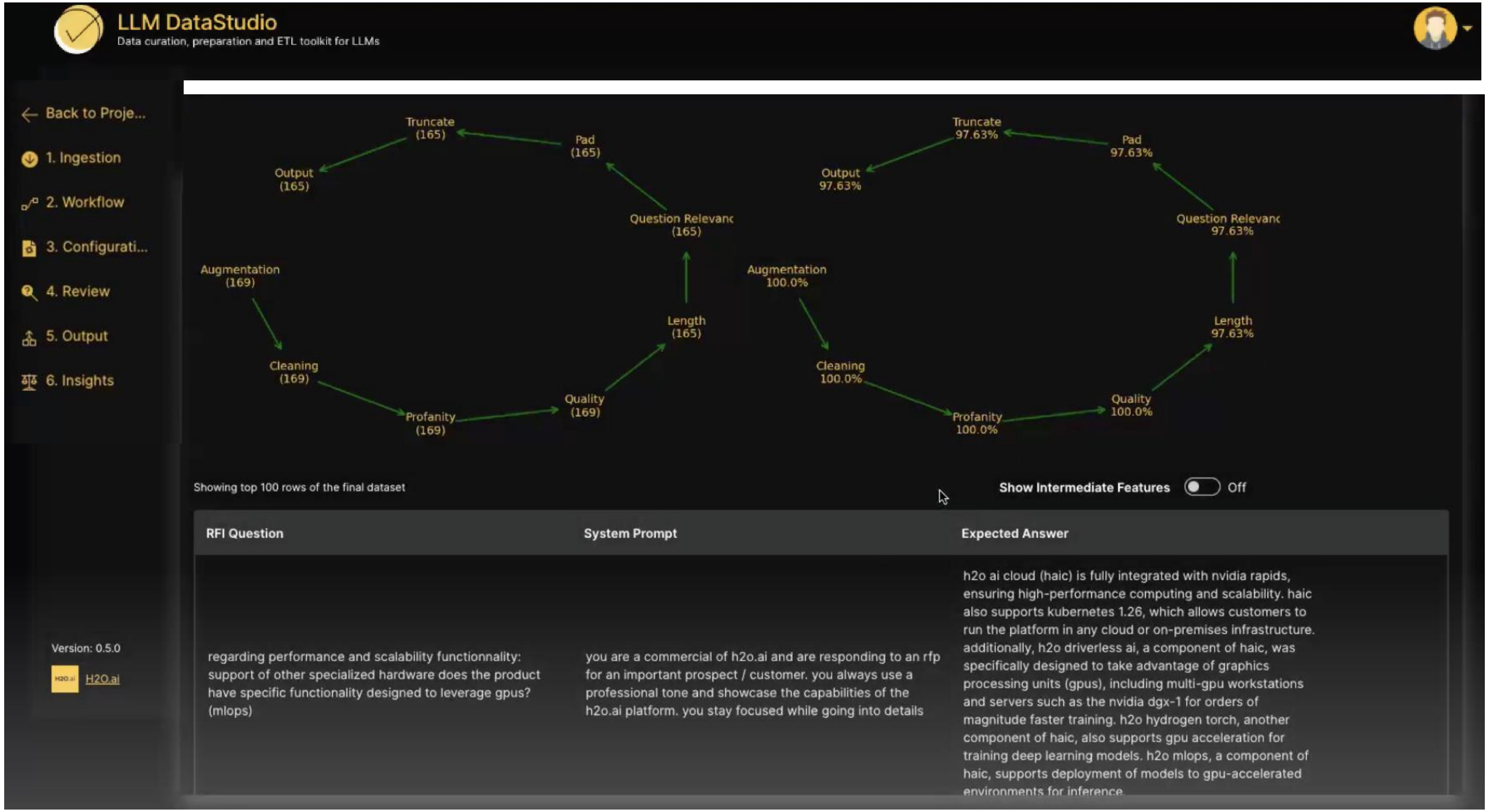


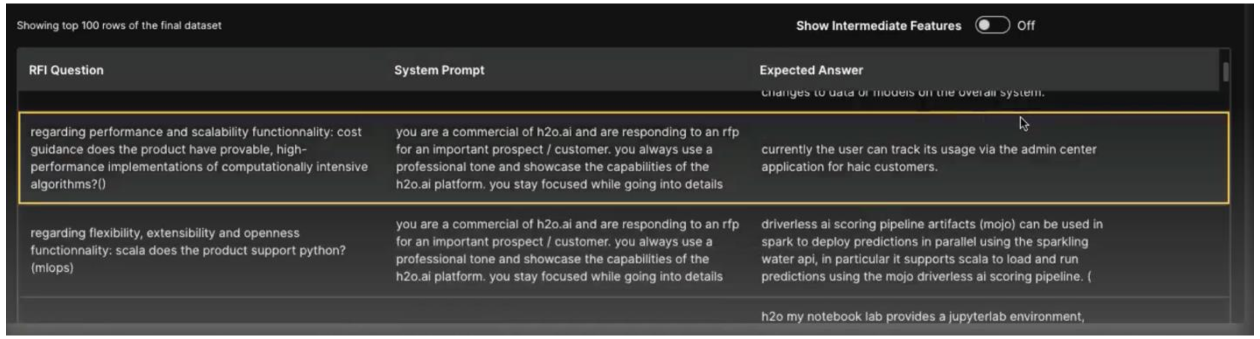
Data Preparation in H20 LLM Data Studio





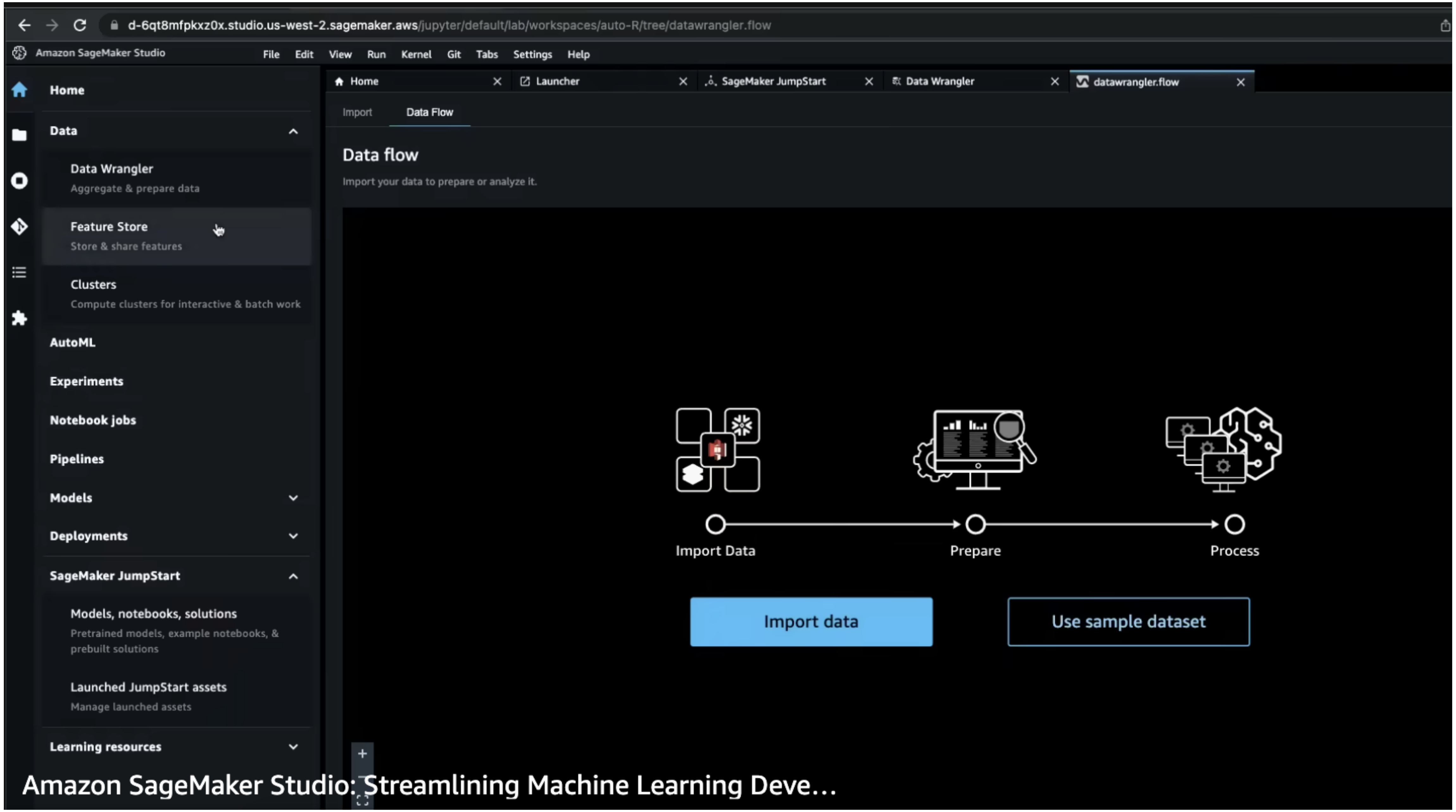
Insights





From the prepared dataset, which is ready to pass to model for fine-tuning, we can quickly scroll through the top 100 rows of the Structured RFI dataset in Insights of LLM Data Studio & review key data and spot trends. This table shows around 170 rows after reducing redundant or bad rows.

Data Preparation in AWS SageMaker Studio: A Quick Overview



References

- Amazon Web Services. (2024). Amazon SageMaker Studio: Streamlining machine learning development from data preparation to model deployment. AWS TV https://aws.amazon.com/awstv/watch/faa2c877499/
- H2O.ai. (2024). LLMs from A to Z: Data prep, fine-tuning & deployment with H2O tools. YouTube. https://www.youtube.com/watch?v=eOUvAY3trJc&t=343s
- H2O.ai. (2023). Hands-on workshop: Fine-tune with H2O LLM Studio. YouTube. https://www.youtube.com/watch?v=_jkedkMpJzo

https://www.youtube.com/watch?v=_jkedkMpJzo

- H2O.ai. (2023). LLM DataStudio: Streamline language model data prep. YouTube. https://www.youtube.com/watch?v=8NQmrDg9KjY

https://www.youtube.com/watch?v=8NQmrDg9KjY

- IBM Developer. (2025). Preparing data for LLM applications using Data Prep Kit. YouTube. https://www.youtube.com/watch?v=gxgHPFOqsNw&t=999s
- Murel, J. (n.d.). Feature engineering. *IBM*. https://www.ibm.com/think/topics/feature-engineering
- https://github.com/h2oai/h2o-llmstudio?tab=readme-ov-file#run-h2o-llm-studio-gui
- https://docs.h2o.ai/h2o-llm-data-studio/
- https://pdf-documentation.s3.amazonaws.com/h2o-llm-datastudio/H2O_LLM_DataStudio__Docs_v0.6.13.pdf
- https://www.youtube.com/channel/UCk6ONJlPzjw3DohAeMSgsng

THANK YOU