

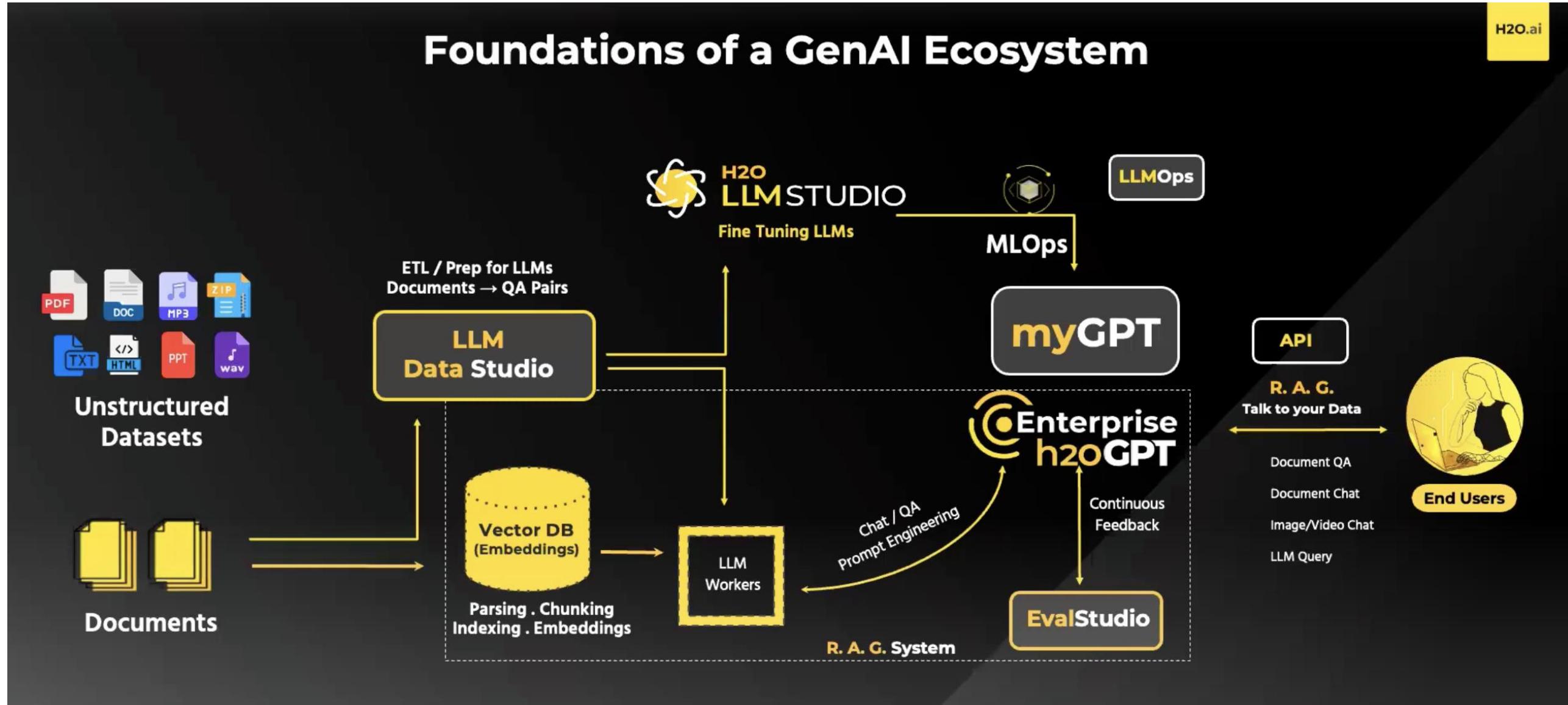
# **H2O LLM Studio**

## **Fine-tuning LLMs with No-Code GUI**

H2O Studio allows for LLM fine-tuning using a no-code graphical user interface. It can be installed locally or accessed via H2O AI Cloud (HAIC).

# 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

H2O LLM Studio v0.0.2

Import dataset

Source: AWS S3

Datasets: AWS S3

File name \*: Must be a .zip, .csv or .pq file.

Configure dataset

Dataset name: 40218556-285d-4305-ac79-1e2a0b175839

Train Dataframe: 40218556-285d-4305-ac79-1e2a0b175839.csv

Validation Dataframe: None

Import dataset

Prompt Column: RFI Question

Answer Column: RFI Question

Parent Id Column: None

H2O LLM Studio v1.5.0

Sample Data Visualization

Data Validity Check. Click Continue if the input data and labels appear correctly.

Sample Number	Field	Content
0	Prompt	regarding performance and scalability functionnality: support of other specialized hardware does the product have specific
0	Answer	h2o ai cloud (haic) is fully integrated with nvidia rapids, ensuring high-performance computing and scalability. haic also su
1	Prompt	regarding data access functionnality: data lineage does the product provide support for accessing data lakes (i.e., hadoop
1	Answer	feature store provides data lineage capabilities: list feature set users according to their rights, actual access rights, and a r
2	Prompt	regarding performance and scalability functionnality: cost guidance does the product have provable, high-performance im
2	Answer	currently the user can track its usage via the admin center application for haic customers.
3	Prompt	regarding flexibility, extensibility and openness functionnality: scala does the product support python?(mllops) you are a cc

A	B	C	D
1	RFI Question	h2oGPTe answer	Expected Answer
	Regarding Data Access functionality: Enterprise Application Integration Question (Don't just think yes/no as the answer, think how well the product supports this functionality!)(Evaluation Goal: MLOps, AutoML, Low-Code, DSHL)	Yes, H2O AI Cloud (HAIC) provides connectors to enterprise applications, both in the cloud and on-premises. The context mentions that HAIC is fully integrated with NVIDIA RAPIDS and ensures high performance and scalability. Additionally, it also mentions that HAIC is designed to meet global compliance requirements and follows data security and network security best practices. However, the context does not explicitly list the specific enterprise applications or connectors that HAIC supports. To answer this question fully, it would be necessary to consult additional documentation or contact H2O.ai for more information on the specific enterprise applications and connectors that HAIC supports.	You are a Commercial of H2oai and are responding to an RFP for an important prospect / customer. You always use a professional tone and showcase the capabilities of the h2oai platform. You stay focused while going into details
2	Regarding Data Access functionality: Data Replication and Synchronization Does the product provide connectors to enterprise applications, both in the cloud and on-premises?(MLOps)	Yes, H2O.ai Feature Store provides options for both complete and incremental data reloads. The documentation mentions the ability to download all produced data from the Feature Store using a blocking approach or a non-blocking approach. The non-blocking approach allows for incremental data loads by creating a job that downloads new data asynchronously. Additionally, the Feature Store provides a retrieve method that can be used to retrieve data from the online Feature Store, which can be done incrementally by specifying a start and end date. The documentation also mentions the ability to schedule ingest, which would allow for regular incremental data loads. Overall, H2O.ai Feature Store provides robust data access and replication capabilities, including options for both complete and incremental data reloads.	You are a Commercial of H2oai and are responding to an RFP for an important prospect / customer. You always use a professional tone and showcase the capabilities of the h2oai platform. You stay focused while going into details

# Step 2: View the Loaded Dataset

## Example 1: RFI 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.

H2O LLM Studio v1.5.0

Navigation

Sample Train Data   Sample Train Visualization   Train Data Statistics   Summary

RFI Question   System Prompt   Expected Answer

RFI Question	System Prompt	Expected Answer
regarding model management	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 h2o.ai platform. you stay focused while going into details ...	optimizing business impact is supported via the use of custom recipies (python code snippets that can be uploaded into driverlessai), and custom scorers in particular: driverlessai will select model parameters that optimize the custom scorer uploaded. there are no capabilities to help find the right accuracy and scoring metrics based on business objectives apart from our documentation. ...
regarding explainable ai functionnalities:	you are a commercial of h2o.ai and are responding to an rfp for an important	driverless ai and hydrogen torch offers functionnalities that help highlights a model strengths and weaknesses. in driverlessai, the mli module offers functionnalities to predict a model likely behavior. and

## Example 2: Open Assistant Dataset

Navigation

Sample Train Data   Sample Train Visualization   Train Data Statistics   Summary

instruction	output	id	parent_id
What types of tests do we have in software development?	There are many types of tests such as unit tests, integration tests, system tests, acceptance tests, regression tests, performance tests, and security tests.	3d52ba91-e016-4772-bf00-000000000000	None
Can you make it about 5	You are looking for a developer who can write a story	f99b20e6-dfc9-4c00-8a9a3a1af-cf29-4ab7-9a00	None
I'm currently on the phone with my boss	Once upon a time, in a kingdom far, far away...	3a760cc2-66bf-4300-8a9a3a1af-cf29-4ab7-9a00	None
If we're going to war, I'll need to predict the enemy's movements	It is difficult to predict the enemy's movements	3899bb57-f478-4d88-e729b2a0c-70ce-47fb-a000	None
Design an exercise program	Designing an exercise program	d18f0c40-7eb2-4313-a000-000000000000	None
What is the angle maximized by a 45° angle?	A 45° angle is the ideal angle for a right-angled triangle	0ac1f351-78f2-40f5-97e0-000000000000	None
I am currently trying to learn how to play the piano	Upgrading individual parts of the piano	057e4461-fca6-4f3e-8000-000000000000	None
Moving from one location to another	Moving from one location to another	74440f41-5569-4230-9000-000000000000	None

Navigation

Input Text: What types of tests do we have in software development?

Target Text: There are many types of tests in software development, each serving a different purpose. Here are some of the most common types of tests:

1. Unit Tests: These are tests that focus on a small, isolated unit of code, such as a single function or method. They are used to verify that the unit of code behaves correctly in isolation.
2. Integration Tests: These are tests that verify that different units of code work correctly together. They are used to identify issues that may arise when different parts of the system are integrated.
3. System Tests: These are tests that verify that the entire system works correctly, from end to end. They are used to ensure that the software meets all of the requirements and behaves correctly in the context of the overall system.
4. Acceptance Tests: These are tests that verify that the software meets the customer's requirements and expectations. They are often used as a final check before releasing the software to the customer.
5. Regression Tests: These are tests that are run to ensure that changes to the software do not introduce new bugs or issues. They are used to catch regressions that may occur when changes are made to the software.
6. Performance Tests: These are tests that verify that the software performs well under expected load and stress. They are used to identify performance bottlenecks and ensure that the software meets performance requirements.
7. Security Tests: These are tests that verify that the software is secure and does not have vulnerabilities that could be exploited by attackers. They are used to identify and mitigate security risks.

These are just a few examples of the types of tests that are used in software development. The specific tests that are used may vary depending on the project and the development methodology being used.

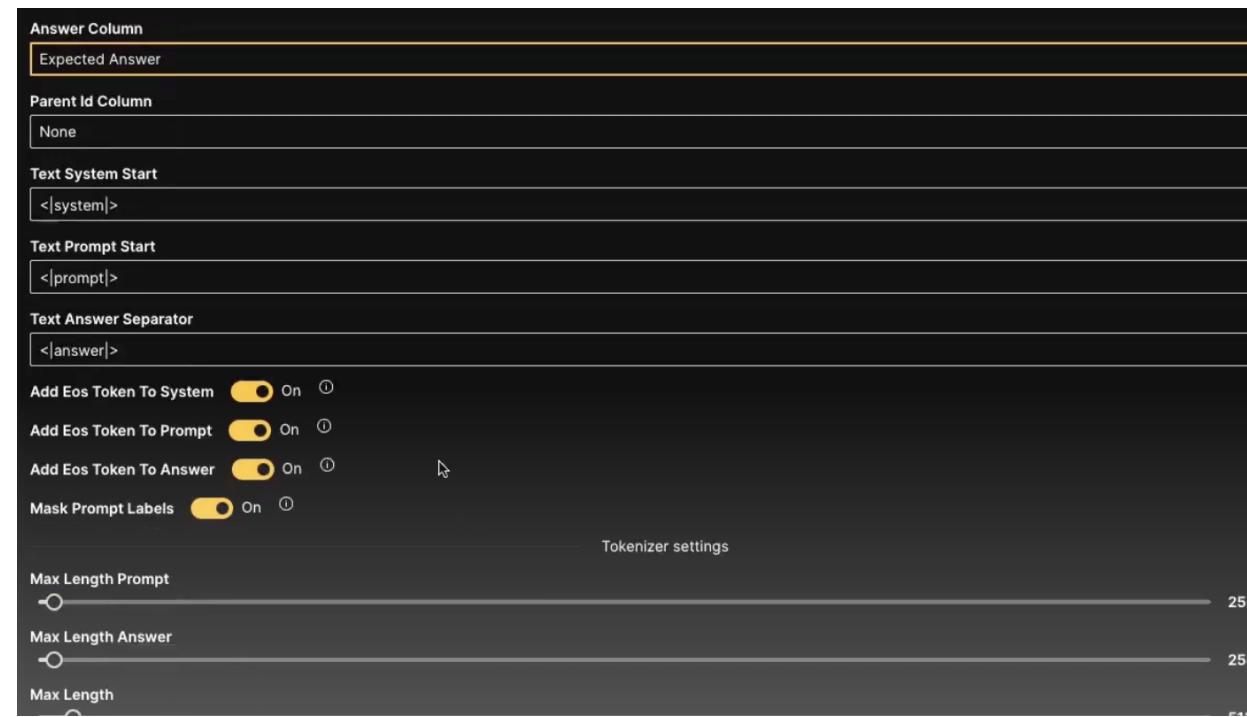
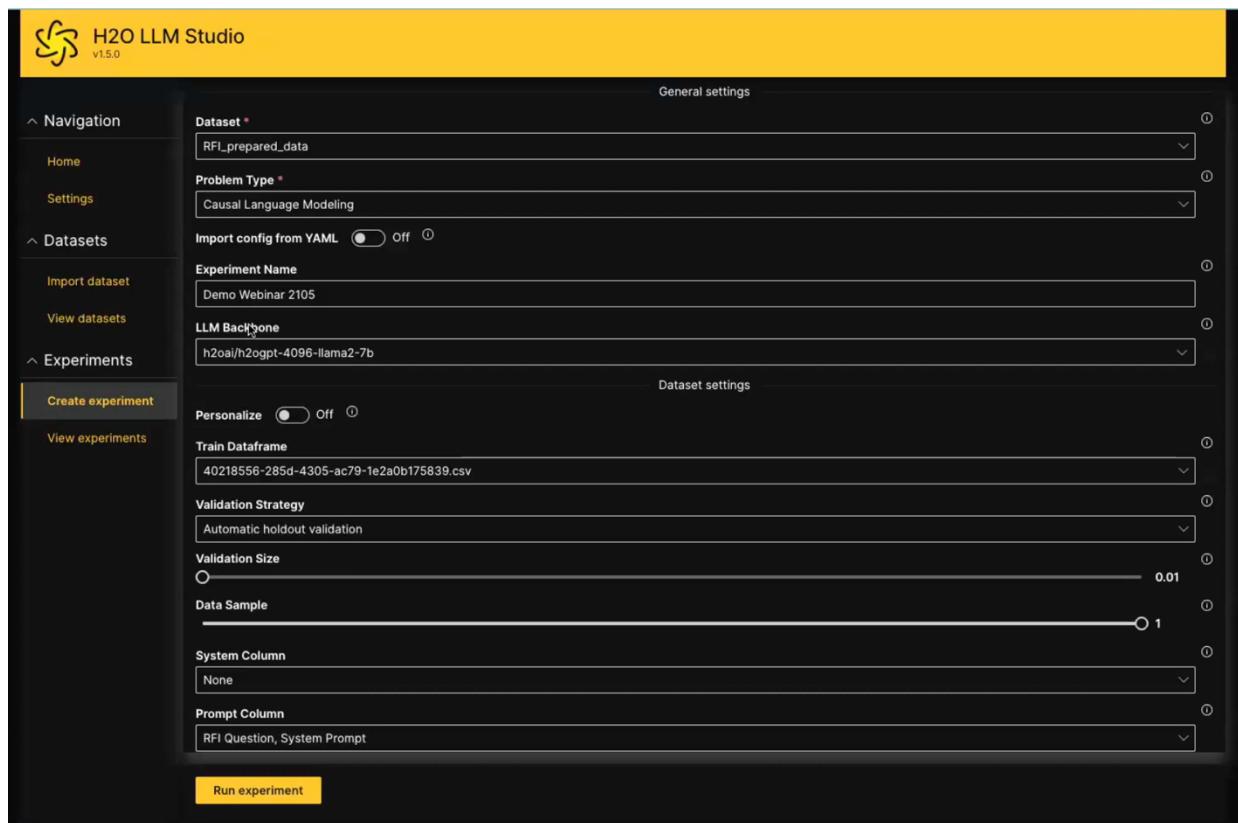
## Step 3: Fine-tune the Model

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

Padding Quantile

Architecture settings

Use Fast  On ⓘ

Backbone Dtype

Gradient Checkpointing  On ⓘ

Force Embedding Gradients  On ⓘ

Intermediate Dropout

Pretrained Weights

Training settings

Loss Function

Optimizer

Learning Rate

Differential Learning Rate Layers

Select optional layers...

Use Flash Attention 2  Off ⓘ

Epochs

Schedule

Warmup Epochs

Weight Decay

Gradient Clip

Grad Accumulation

Lora  On ⓘ

Lora R

Lora Alpha

Lora Dropout

Lora Target Modules

Save Best Checkpoint  Off ⓘ

Evaluation Epochs

Evaluate Before Training  Off ⓘ

Train Validation Data  Off ⓘ

Augmentation settings

Train Validation Data  Off ⓘ

Token Mask Probability

Skip Parent Probability

Random Parent Probability

Neftune Noise Alpha

Prediction settings

Metric

BLEU

GPT

Perplexity

Min Length Inference

Max Length Inference

Max Time

Batch Size Inference

Environment settings

Top K

Top P

Opus

Select All  Deselect All

Mixed Precision  On ⓘ

Mixed Precision Dtype

Compile Model  Off ⓘ

Use DeepSpeed  Off ⓘ

Find Unused Parameters  Off ⓘ

Trust Remote Code  On ⓘ

Huggingface Branch

Number Of Workers

Logging settings

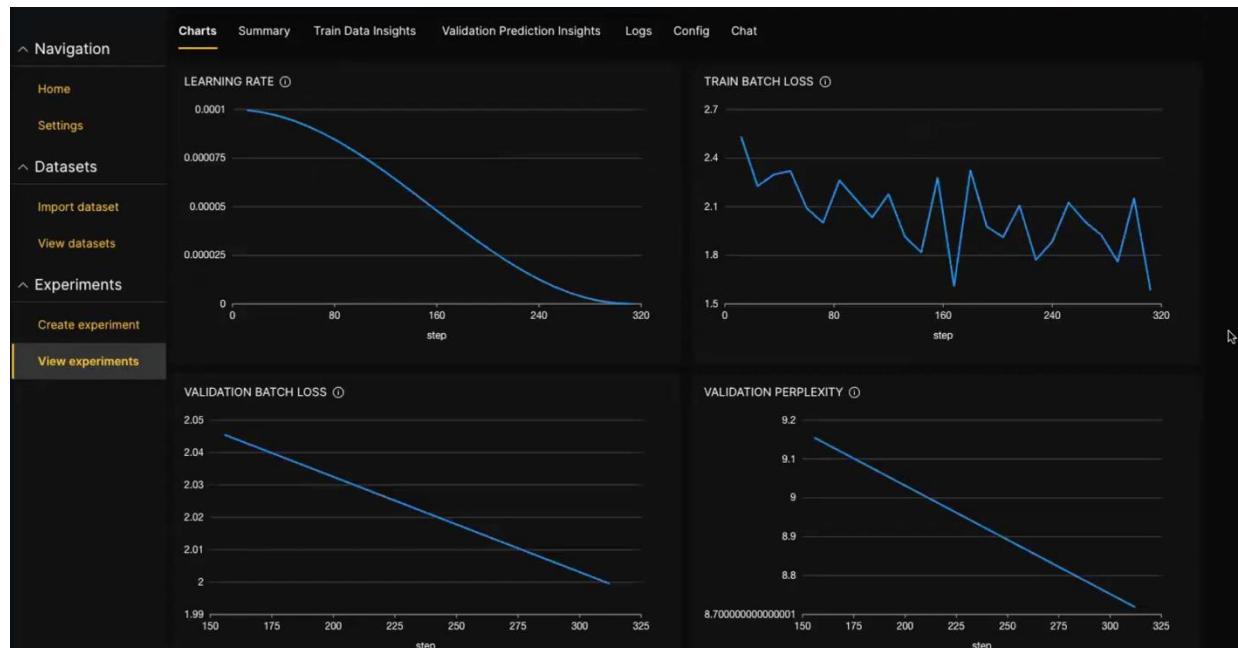
Seed

Logger

## 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.



Input Text (tokenization max length setting may tr...)	Target Text	Predicted Text	Metric (Perf)
You are a Commercial of <a href="#">H2o.ai</a> and are responding to	Although it is not available off the shelf, Users can leve	No predictions are generated for the selected metric	9.84000015
Expected Answer:	Developers can use H2O Wave to build custom applic		
You are a Commercial of <a href="#">H2o.ai</a> and are responding to	both Driverless AI and Hydrogen Torch automate the p	No predictions are generated for the selected metric	5.71000003
Expected Answer:	Currently, RBAC are in development in H2O MLOps.		
You are a Commercial of <a href="#">H2o.ai</a> and are responding to	H2O AI Cloud (HAIC) product provides a robust platfor	No predictions are generated for the selected metric	11.6350002
Expected Answer:	Currently, RBAC are in development in H2O MLOps.		
	H2O AI Cloud (HAIC) product supports governance of		
	<ul style="list-style-type: none"><li>• passphrase protected model deployment endpo</li><li>• Audit logging of model-related activities</li><li>• Model versioning and rollback capabilities</li><li>• Model monitoring and alerting for performance</li><li>• Support scoring requests including prediction cc</li></ul>		
You are a Commercial of <a href="#">H2o.ai</a> and are responding to	Additionally, the MLOps component is fully integrated		
Expected Answer:	It does not provide specific features for creating polici		

This screenshot shows a detailed view of a validation sample. It includes:

- Input Text:** Still using layman's terms can you explain the different types of options, how they work and what they can be used for?
- Target Text:** Certainly! There are two main types of options: call options and put options.
- Predicted Text:** Layman's terms:

  - 1.

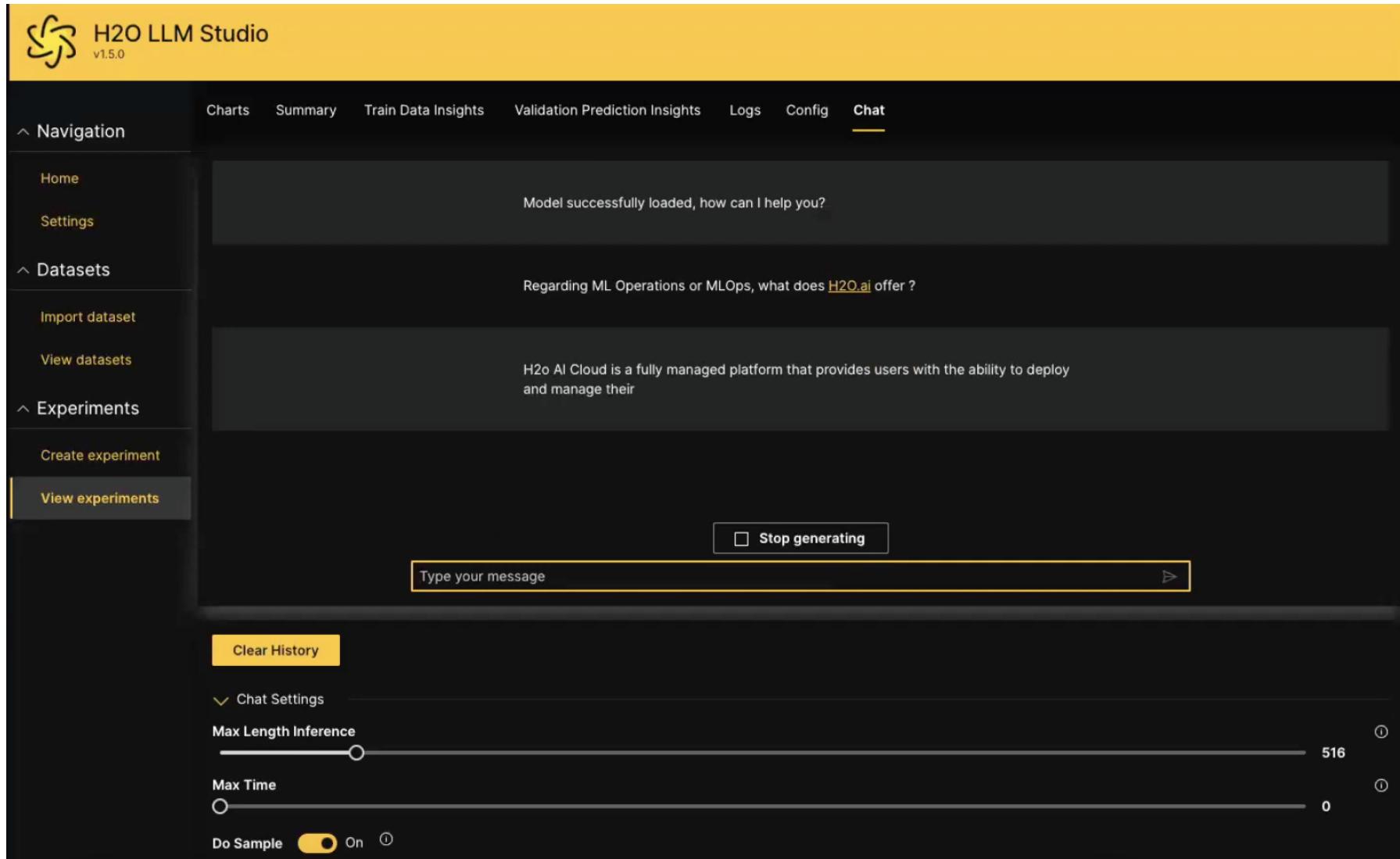
- BLEU Score:** 0.000
- Input Text:** now do it in the style of shakespeare
- Target Text:** In England's fair land there lived a youth, Whose name was ...

## 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.

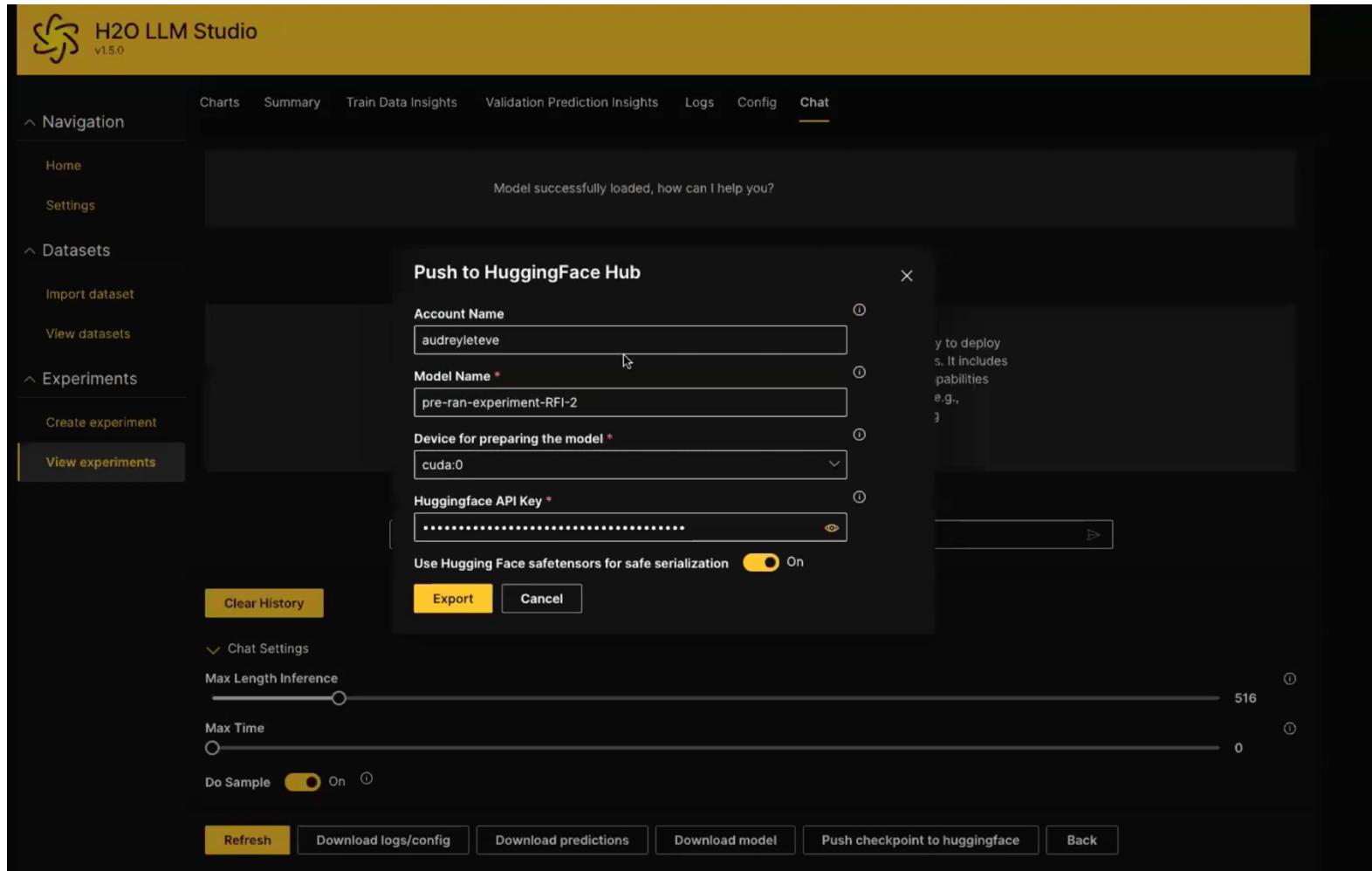
This allows for immediate testing of the model's responses to various prompts and scenarios, ensuring it behaves as expected in a conversational context.



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



Search models, datasets, users...

Models

Datasets

Spaces

Posts

Docs

Solutions Pricing

## 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.

The screenshot shows a Jupyter Notebook interface with two tabs: 'llmops-demo\_2.ipynb' and 'llmops-demo.ipynb'. The left pane displays a file tree for a 'webinar' directory containing various Jupyter notebooks and configuration files. The right pane shows code snippets:

- LLMops on H2O MLOps:** A section describing how to deploy a Huggingface Hub LLM model to H2O MLOps using the OpenAI API protocol.
- Connect to MLOps via LLMOps Client:** Instructions for installing and upgrading the LLMOpsClient using pip.
- Code Snippets:**
  - [1]:

```
import time
from h2o_llmops import LLMOpsClient
llmops = LLMOpsClient(
    h2o_cloud_url="https://internal.dedicated.h2o.ai",
    refresh_token=REFRESH_TOKEN
)
```
  - [1]:

```
llmops.project = "Webinar LLM AtoZ"
print("Current Project:", llmops.project.name)
print("Current Environment:", llmops.environment.name)
print("Owner:", llmops.project.owner)
```

### Models | for Data Scientists and Machine Learning Engineers

Here we add another model version (not deployment)

```
[78]: llm = llmops.models.update_or_create(
    model = "audreyleteve/pre-ran-experiment-2", # huggingface path
)

print("LLM name:", llm.name)
print("LLM uid:", llm.uid)
print("LLM config:")
display(llm.config(version="latest"))

Downloading artifacts: 100% [1/1 [00:00<00:00, 97.63it/s]
LLM name: audreyleteve/pre-ran-experiment-2
LLM uid: 5815095b-470d-42f1-9239-b91a5e796f56
LLM config:
{'model': 'audreyleteve/pre-ran-experiment-2',
 'name': 'audreyleteve/pre-ran-experiment-2'}
```

```
[79]: llmops.models.list()
```

	name	uid
0	audreyleteve/pre-ran-experiment-2	5815095b-470d-42f1-9239-b91a5e796f56
1	audreyleteve/pre-ran-experiment-RFI-2	78de3d9d-6efa-4ef3-b188-50801b3a48f7

```
[*]: llm = llmops.models.get("5815095b-470d-42f1-9239-b91a5e796f56")
```

```
[ ]: llm.config()
```

# 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
[82]:   +-----+
[82]:   0 | audreyteteve/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 H2o.ai platform. Your answers should be concise and informative, highlighting the unique features and benefits of the platform. You should also be able to answer any questions related to the deployment mode, integration with other platforms, and overall functionality of the H2o.ai platform.""
[ ]: 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 their own 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 organizations 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 worlds, 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 managed 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 3: Data Preparation & Feature Engineering

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

07

## Phase 5: Evaluation & Performance Metrics

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

## Phase 2: Data Understanding & Exploration

Collecting, describing, and exploring financial transaction data. Through thorough exploration, we acknowledge **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 4: Modeling & Algorithm Selection

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

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



## Data Transformation & Feature Engineering

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



## Data Reduction

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

# Data Preparation in H2O LLM Data Studio

The screenshot shows the LLM DataStudio homepage. On the left is a navigation sidebar with links for Home, Curate, Prepare, Custom Eval, Augment, Settings, Info, Help Guide, Documentation, and Importance. The main content area features the title "LLM DataStudio by H2O.ai" and a dark background with yellow hexagonal callouts. The first callout is "Curate Q:A pairs from unstructured data". The second is "Prepare Clean, Structure & Augment datasets for LLMs". The third is "Custom Eval Create your own LLM evaluation datasets". Below these are descriptions for each feature: Curate (Documents PDFs, DOCs, Audio, Video files), Prepare (Text preprocessing techniques), and Custom Eval (Create your own evaluation datasets with different types). A note at the bottom states: "No-code UI for streamlining the curation, preparation, and creation of evaluation datasets for LLMs."

This screenshot shows the "New Project / Prepare Data for LLMs" interface. It includes fields for "Project Name" (set to "Webinar"), "Description", and "Problem Type" (set to "Question Answering"). Below this is an "Info" section with "Help Guide" and "Documentation" links. The "Importance" section has "Add" and "Discard" buttons. At the bottom, there's a table titled "Workflow" with columns "Workflow" and "Expected Columns". The table lists five workflows: Question Answering (question, answer, context), Instruct Tuning (prompt, response), Human Bot Conversation (message\_id, parent\_id, text, role), Text Summarization (article, summary), and Continue PreTraining (text). The version of the software is listed as 0.5.0.

This screenshot shows the "Webinar Demo RFI / Ingestion" screen. The navigation sidebar includes links for Back to Project, 1. Ingestion, 2. Workflow, 3. Configuration, 4. Review, 5. Output, and 6. Insights. The main area shows a tree view of project steps: 1.1 Project Details, 1.2 Add New Datasets, 1.3 Configure Datasets, and 1.4 Configure Columns. Under "Configure Columns", there are dropdowns for "Question Column" (RFI Question), "Answer Column" (Expected Answer), and "Context Column" (System Prompt). Buttons for "Preview" and "Delete" are also present. A success message "Config Saved Successfully" is shown at the bottom.

## LLM DataStudio Supported Workflows

### 1. Question and Answer Workflow:

- Preparing Datasets for Question Answering Models
- Structured Datasets with Context, Questions, and Answers
- Crucial for Accurate User Query Responses

### 2. Text Summarization Workflow:

- Handling Articles and Summaries
- Extracting Key Information for Concise Summaries
- Training Summarization Models for Informative Summaries

### 3. Instruct Tuning Workflow:

- Creating Datasets with Prompts and Responses
- Training Models to Understand and Follow Instructions
- Effective Responses to User Prompts

### 4. Human - Bot Conversations Workflow:

- Organizing Dialogues between Humans and Chatbots
- Enhancing Conversational Model Training
- Understanding User Intents and Providing Contextual Responses

### 5. Continued PreTraining Workflow:

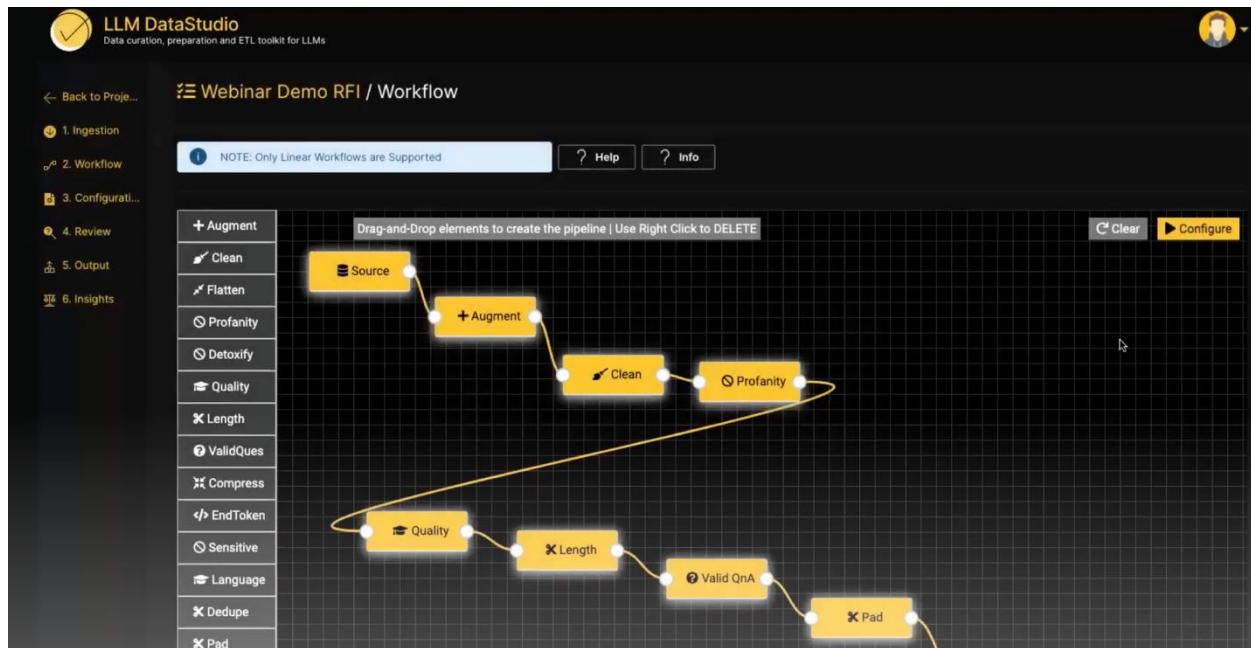
- Preparing Extensive Text Datasets for Pretraining
- Organizing Long Texts for Enhanced Language Models
- Improving Language Understanding and Generation



H2O.ai  
University

H2O.ai

# Data Preparation in H2O LLM Data Studio



The screenshot shows the 'Webinar Demo RFI / Info - Supported Techniques' screen. The sidebar has steps 1-6. The main area is a table of supported techniques:

Operation	Description
Augment	Augment external datasets (example: add RLHF datasets)
Clean	Text cleaning using different methods
Flatten	Flatten human-bot conversations into single text object
Profanity	Check for profanity in the texts and filter based on threshold
Detoxify	Check for toxicity in the texts and filter based on threshold
Quality	Check for Text Grade (too simple or too complex to read) and filter them
Length	Chop the text by a max-length or chop using a pre-trained model
ValidQues	Check (and filter) if the question is a valid question text
Compress	Filter the text summarization output using a compression ratio
EndToken	Append Start/End tokens in the text boundaries
Sensitive	Check and remove any sensitive information (emails, phone-numbers, card-numbers)
Language	Drop the texts which are in a different language than the most common language in the corpus
Dedupe	Check for duplicate records (using hashing) remove them
Pad	Pad sequence based on a min length parameter
Truncate	Truncate text based on a threshold and the max length parameter

The screenshot shows the 'Webinar Demo RFI / Configure' screen. The sidebar has steps 1-6. Sections include:

- 1. Ingestion: Dataset\_priceless\_newton0 selected.
- 2. Workflow: Select button.
- 3. Configuration: Filter by Column, Text Cleaning (Newline, Whitespace, Lower Casing, URLs, ASCII, Emoji Removal), Profanity Check (Acceptable Profanity Threshold slider at 0.3).
- 4. Review.
- 5. Output.
- 6. Insights.

The screenshot shows configuration details for steps 5 through 11:

- 5. Length Check: Context Length (10 to 5000), Question Length (10 to 3000), Answer Length (10 to 5000).
- 6. Text Quality Check: Acceptable Text Grade (7 to 30).
- 7. Question Relevance Check: Check Question relevance (On).
- 8. Add your own code: A note says 'This functionality has been disabled. Please reach out to support@h2o.ai to enable this capability'.
- 9. Pad Sequence: Max Padding Length (600).
- 10. Truncate Sequence: Truncate Max Length (10000), Truncate Ratio (0.15), Model Based toggle (Off).
- 11. Output Format: Output Type (CSV).

The screenshot shows the final 'Review' step:

Review

# Data Preparation in H2O LLM Data Studio

The screenshot shows the LLM DataStudio interface with the following details:

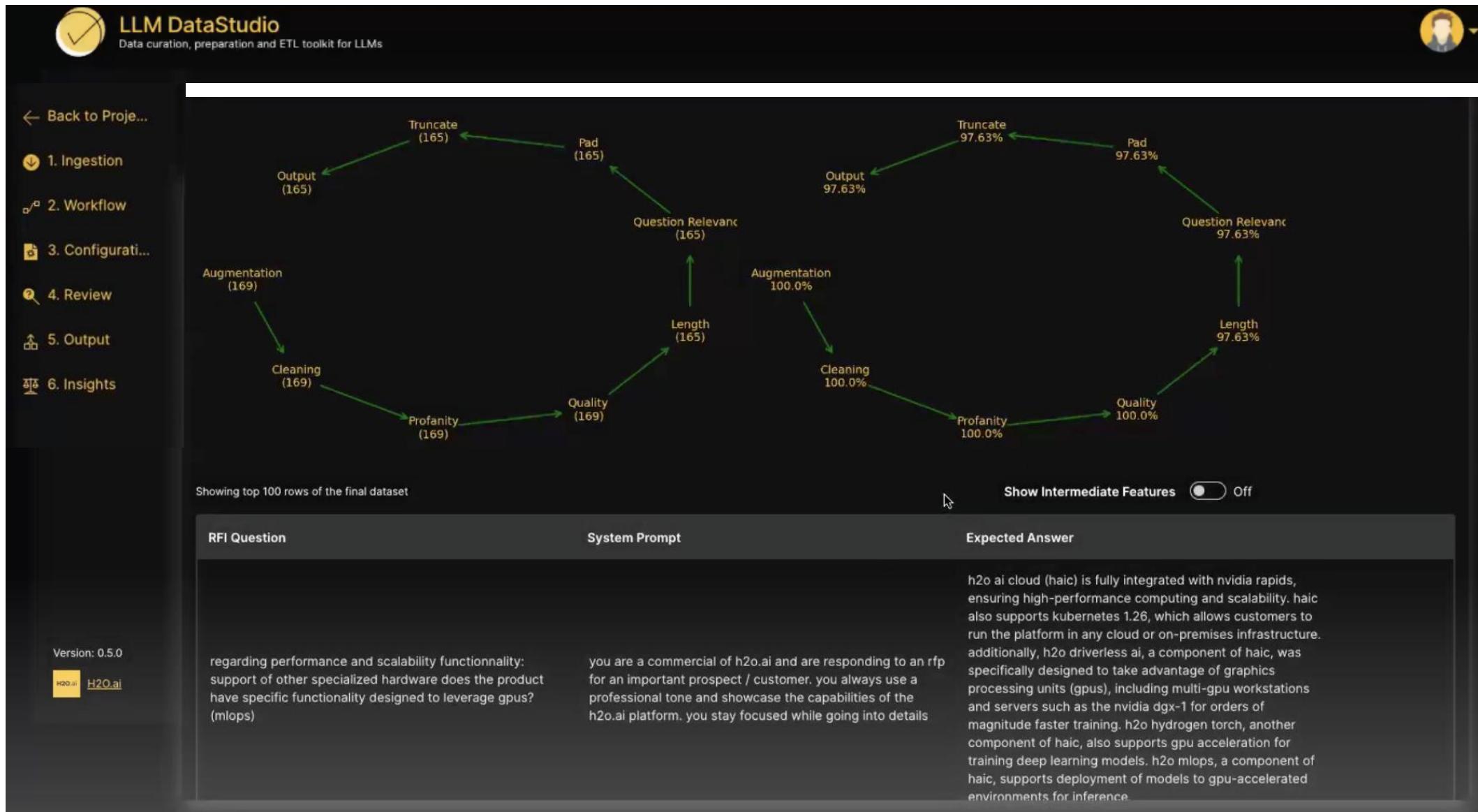
- Project Name:** Webinar Demo RFI / Review
- Navigation:** Back to Project, 1. Ingestion, 2. Workflow, 3. Configuration, 4. Review, 5. Output, 6. Insights.
- Pipeline Configuration:**

```
toxicity: 0.9
filter_column:
  null:
    column: null
    operation: null
    value: null
language_understander: 0.65
length_checker:
  max_answer_length: 5000
  max_context_length: 5000
  max_question_length: 3000
  min_answer_length: 10
  min_context_length: 10
  min_question_length: 10
output_type: csv
padding:
  max_length: '600'
profanity_checker: 0.3
quality_checker:
  max_grade: 30
  min_grade: 7
question_relevance_checker: 1
relevance_checker: 0.15
sensitive_info:
  funcs:
    - email
    - phoneno
sensitive_info_checker: 1
text_clean:
  cols: []
  funcs:
    - new_line
    - whitespace
    - lower_case
    - urls
    - ascii
    - emoji
```

The screenshot shows the 'Running the Pipeline' step with the following details:

- Section:** Processing the Data
- Text:** This may take few seconds to complete depending upon the data prep steps.

# Insights



This screenshot shows a similar view of the LLM DataStudio interface, focusing on the table of RFI questions, system prompts, and expected answers. A specific row is highlighted with a yellow box, indicating it has been selected or is being reviewed. The columns are RFI Question, System Prompt, and Expected Answer. The RFI Question row contains a question about cost guidance for computationally intensive algorithms. The System Prompt row contains a template response. The Expected Answer row contains the generated text. A "Show Intermediate Features" toggle switch is set to "Off".

RFI Question	System Prompt	Expected Answer
regarding performance and scalability functionality: cost guidance does the product have provable, high-performance implementations of computationally intensive algorithms?()	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 h2o.ai platform. you stay focused while going into details	currently the user can track its usage via the admin center application for haic customers.
regarding flexibility, extensibility and openness functionality: scala does the product support python? (mlops)	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 h2o.ai platform. you stay focused while going into details	driverless ai scoring pipeline artifacts (mojo) can be used in spark to deploy predictions in parallel using the sparkling water api, in particular it supports scala to load and run predictions using the mojo driverless ai scoring pipeline. (

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

The screenshot shows the AWS SageMaker Studio interface with the following details:

- Header:** The URL is `d-6qt8mpkxz0x.studio.us-west-2.sagemaker.aws/jupyter/default/lab/workspaces/auto-R/tree/datawrangler.flow`. The tabs at the top include Home, Launcher, SageMaker JumpStart, Data Wrangler, and the current tab, datawrangler.flow.
- Left Sidebar:** The sidebar under the Data category includes Data Wrangler, Feature Store, Clusters, AutoML, Experiments, Notebook jobs, Pipelines, Models, Deployments, and SageMaker JumpStart. Under SageMaker JumpStart, there are sections for Models, notebooks, solutions (with a note about pretrained models) and Launched JumpStart assets (with a note about managing launched assets). There is also a Learning resources section.
- Central Content:** The main area is titled "Data flow" with the sub-section "Import". It contains the text "Import your data to prepare or analyze it." Below this is a diagram illustrating the data preparation process:
  - The process starts with "Import Data", represented by a cluster icon.
  - An arrow points to the "Prepare" step, represented by a monitor icon with a magnifying glass.
  - Another arrow points to the "Process" step, represented by a cluster icon with gears.
- Buttons:** At the bottom, there are two buttons: "Import data" (highlighted in blue) and "Use sample dataset".
- Footer:** The footer displays the text "Amazon SageMaker Studio: Streamlining Machine Learning Deve...".

# 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>
- 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://pdf-documentation.s3.amazonaws.com/h2o-llm-datastudio/H2O_LLM_DataStudio_Docs_v0.6.13.pdf)
- <https://www.youtube.com/channel/UCK6ONJIPzjw3DohAeMSgsng>

THANK YOU