



H2O LLM DataStudio

DOCUMENTATION

For questions, please contact **support@h2o.ai**

Contents

What is H2O LLM DataStudio?	6
Who is H2O LLM DataStudio for?	6
Importance of cleaned data in NLP Downstream tasks	6
Video guide	7
Access H2O LLM DataStudio	8
Step 1: Access HAIC	8
Step 2: Search H2O LLM DataStudio	8
Step 3: Run H2O LLM DataStudio	8
Step 4: H2O LLM DataStudio instance	8
Pause or terminate instance	9
Data curation flow	11
Overview	11
Step 1: Create a new curate project	11
Step 2: Upload documents	11
Step 3: Configure and run pipeline	11
Step 4: Use the output for data preparation	11
Video guide	12
Data preparation flow	13
Overview	13
Step 1: Ingest data	13
Step 2: Build the workflow	13
Step 3: Configure the parameters	13
Step 4: Review and execute	13
Step 5: Analyze the output	13
Step 6: Compare datasets	13
Video guide	13
Tutorials	14
Learning path	14
Question and Answer	14
Text Summarization	14
Instruct Tuning	14
Human - Bot Conversations	14
Continued PreTraining	14
Data ingestion	15
Overview	15
Instructions	15
Workflow builder	18
Overview	18
Instructions	18
Video guide	18
Configuration	19
Overview	19
Instructions	19
General configurations	19
Source datasets / Augmentation	19
Filter by column	19
Text cleaning	19
Profanity check	19
Detoxify	19
Length check	20
Text quality check	20

Sensitive info check	20
Bias check	20
Add your own code	20
Pad sequence	20
Truncate sequence	21
Configurations for question-answering	21
Question relevance check	21
Configurations for text summarization	21
Filter compression	21
Add special tokens	21
Configurations for human-bot conversation	21
Flatten conversation	21
Review and execute	22
Overview	22
Instructions	22
Output	23
Overview	23
Instructions	23
Insights	24
Overview	24
Instructions	24
Tutorial 1A: Dataset preparation for Question Answering	26
Overview	26
Prerequisites	26
Step 1: Explore the project	26
Step 2: Ingest data	26
Step 3: Build the workflow	26
Step 4: Configure the parameters	27
Step 5: Review the configured parameters and execute the workflow	27
Step 6: Review and analyze the output	27
Step 7: Compare input and output datasets	27
Summary	28
Video guide	28
Tutorial 2A: Dataset preparation for Text Summarization	29
Overview	29
Prerequisites	29
Step 1: Explore the project	29
Step 2: Ingest data	29
Step 3: Build the workflow	29
Step 4: Configure the parameters	30
Step 5: Review the configured parameters and execute the workflow	30
Step 6: Review and analyze the output	31
Step 7: Compare input and output datasets	32
Summary	32
Tutorial 3A: Dataset preparation for Instruct Tuning	33
Overview	33
Prerequisites	33
Step 1: Explore the project	33
Step 2: Ingest data	33
Step 3: Build the workflow	34
Step 4: Configure the parameters	35
Step 5: Review the configured parameters and execute the workflow	35
Step 6: Review and analyze the output	36
Step 7: Compare input and output datasets	37

Summary	37
Tutorial 4A: Dataset preparation for Human - Bot Conversations	38
Overview	38
Prerequisites	38
Step 1: Explore the project	38
Step 2: Ingest data	38
Step 3: Build the workflow	39
Step 4: Configure the parameters	40
Step 5: Review the configured parameters and execute the workflow	40
Step 6: Review and analyze the output	41
Step 7: Compare input and output datasets	42
Summary	42
Tutorial 5A: Dataset preparation for Continued PreTraining	43
Overview	43
Prerequisites	43
Step 1: Explore the project	43
Step 2: Ingest data	43
Step 3: Build the workflow	44
Step 4: Configure the parameters	45
Step 5: Review the configured parameters and execute the workflow	45
Step 6: Review and analyze the output	46
Step 7: Compare input and output datasets	47
Summary	47
Tutorial 1B: Convert documents into Q&A pairs for data preparation	48
Overview	48
Prerequisites	48
Step 1: Create a new project	48
Step 2: Configure settings	48
Step 3: Run pipeline	49
Step 4: View the project	49
Step 5: Publish the dataset as preparation project	49
Summary	50
Tutorial 1C: Create question-type and robust-type evaluation datasets	51
Overview	51
Prerequisites	51
Step 1: Create a new evaluation dataset	51
Step 2: Upload question answering (QA) dataset	51
Step 3: Configure settings	51
Step 4: Run pipeline	52
Step 5: View the project	52
Step 6: Generate a robust evaluation dataset	52
Summary	53
Supported problem types	54
Overview	54
Question and Answer	54
Text Summarization	54
Instruct Tuning	55
Human - Bot Conversations	55
Continued PreTraining	55
Video guide	56
Supported functionalities	57
Video guide	57
Create a project	58

Overview	58
Instructions	58
View projects	59
Overview	59
View all projects	59
Delete a project	60
Overview	60
Instructions	60
Key functionalities	61
Variety of data types	61
LLM-Based question-answer pair generation	61
LLM-Based context summarization pair generation	61
Fast QA Mode	61
View and customize the output	61
Use the new structured dataset to finetune LLM in H2O LLM Studio.	61
Create a new project for data curation	62
Overview	62
Instructions	62
View a Curate project	66
Overview	66
View a specific Curate project	66
Delete a Curate project	68
Overview	68
Instructions	68
Create your own evaluation datasets	69
Overview	69
Instructions	69
View an evaluation dataset	71
Overview	71
View a specific Custom Eval project	71
Delete an evaluation dataset(s)	72
Overview	72
Instructions	72
Augment	73
Instructions	73
Augment Datasets (RLHF, Improve Content, Extra Rows)	73
Preloaded datasets	73
Settings	76
Overview	76
Integrate h2ogpte	76
Key terms	77
LLM (Large Language Model)	77
Data curation	77
Tokenization	77
Truncation	77
RLHF (Reinforcement Learning with Human Feedback)	77
Augmentation	77
HAMC	77
Relevance score	78

Sampling ratio	78
Robust evaluation dataset	78
FAQs	79
General	79
What are the main workflows supported by H2O LLM DataStudio?	79
What are some of the main features of H2O LLM DataStudio?	79
What is the importance of good data for training Large Language Models?	79
Is the LLM DataStudio multi-user or single user?	79
Is there an API for LLM DataStudio?	79
If GPT is used, is LLM DataStudio still necessary, or is filtering handled automatically?	79
What does the relevance mean for the question-answer pairs?	79
What are the models used in LLM DataStudio?	79
Prepare	79
How is detoxifying performed in LLM DataStudio?	79
What are the file formats supported for exporting the processed data?	80
How does H2O LLM DataStudio handle data quality?	80
Can I manage multiple data tasks in H2O LLM DataStudio?	80
How does H2O LLM DataStudio ensure data privacy and safety?	80
Are there redundancy checks in LLM DataStudio? (i.e. two question-answer pairs are very similar)	80
Curate and Custom Eval	80
When converting documents to question-answer pairs, is there a way of seeing the progress of the data curation process in percentage?	80
Can we stop the process midway at some point and download the questions-answer pairs generated so far before it's fully complete?	80
Can you create question-answer pairs on multiple files in one go?	80
How does LLM DataStudio work with audio files?	80
For question and answer generation in LLM DataStudio, can you select the LLM to use to create question-answer pairs?	80
Can I merge datasets in LLM DataStudio?	81
How does chunking work in LLM DataStudio?	81
Augment	81
Does H2O LLM DataStudio support data augmentation?	81

What is H2O LLM DataStudio?

H2O LLM DataStudio is a no-code web application specifically designed to streamline and facilitate data curation, preparation, and augmentation tasks for Large Language Models (LLMs). H2O LLM DataStudio has 4 major components.

- **Curate:** You can convert unstructured data such as PDFs, DOCs, audio, and video files into structured question-answer pairs, chunk summaries, and file summaries.
- **Prepare:** You can manage your data tasks by creating, organizing, and tracking projects using the **Prepare** tab. H2O LLM DataStudio is its support for various workflows and task types. These tasks span question and answer models, text summarization, instruction tuning, human-bot conversation models, and continued pretraining of language models. Each workflow is accompanied by a customized set of functionalities, which assist in optimal preparation and structuring of datasets for the desired tasks.
- **Custom Eval:** You can create your own evaluation datasets with different evaluation types (question type, multi-choice, token presence) from documents (PDFs, DOCs, audio or video files) or datasets.
- **Augment:** You can combine external and RLHF datasets with your own data to make them rich and bias-free.

H2O LLM DataStudio underlines the [importance of clean data in training LLMs](#). The platform ensures the quality and suitability of the data being fed into the models by offering tools such as a profanity checker, text quality checker, and sensitive information detector. It allows the trained models to be more reliable, accurate, and effective in real-world applications.

Who is H2O LLM DataStudio for?

H2O LLM DataStudio is a complete solution that caters to the needs of developers, data scientists, and AI practitioners working with LLMs, providing a wide range of functionalities to effectively handle and manipulate large datasets. H2O LLM DataStudio is built to be user-friendly and easily accessible by offering a no-code web interface. This versatility allows both experienced developers and those with limited coding expertise to use the platform effectively.

Importance of cleaned data in NLP Downstream tasks

Cleaned data plays a vital role in fine-tuning and improving the performance, fairness, and ethical considerations of NLP models in downstream tasks. Here are key reasons why cleaned data are important in NLP downstream tasks:

- **Enhanced model performance:** Cleaned data eliminates noise, errors, and inconsistencies that could restrain model performance. By removing irrelevant or misleading information, the model can focus on learning patterns and relationships that are more relevant to the task at hand. It leads to improved accuracy, precision, and overall performance of the model in downstream tasks.
- **Reduced bias and unwanted influences:** Cleaning the data helps mitigate biases and unwanted influences that may have been present in the training data. Bias in the data can be reflected in the model's predictions and outputs. You can try to minimize the impact of biases by carefully curating and cleaning the data. This will lead to more unbiased and equitable results.
- **Consistency and coherence:** Cleaned data ensures consistency and coherence in the input to the model. Inconsistencies, such as conflicting information or contradictory statements, can confuse the model and negatively affect its responses. You provide the model with a more coherent and reliable input by cleaning and standardizing the data. It enables to generate more meaningful and accurate outputs.
- **Improved generalization:** Cleaning the data helps the model generalize better to new or unseen examples. The model can focus on learning robust and transferable patterns by removing irrelevant or noisy data. It improves the model's ability to handle diverse inputs in real-world scenarios and produce more reliable predictions.
- **Ethical considerations:** Cleaning the data allows for the removal of offensive, hateful, or inappropriate content. Models trained on such data can generate responses that promote harmful behavior or propagate misinformation. You can mitigate the risks of the model generating undesirable or harmful outputs by ensuring that the data is free from offensive or unethical content.
- **User experience and trust:** Cleaned data leads to more accurate and reliable outputs, enhancing the user experience and building trust in the model's performance. Users are more likely to trust and rely on models that consistently produce high-quality and trustworthy results. Cleaned data contributes to the development of more dependable and user-friendly NLP applications.

Cleaned data enables models to perform better, generalize effectively, and generate reliable and trustworthy outputs by removing noise, biases, and inconsistencies.

Video guide

Watch [this](#) video guide to learn more about the importance of data preparation and how clean data enhances the reliability and ethical integrity of NLP models.

Access H2O LLM DataStudio

You can access H2O LLM DataStudio through an instance that you can create on the H2O AI Cloud (HAIC). To access LLM Data Studio:

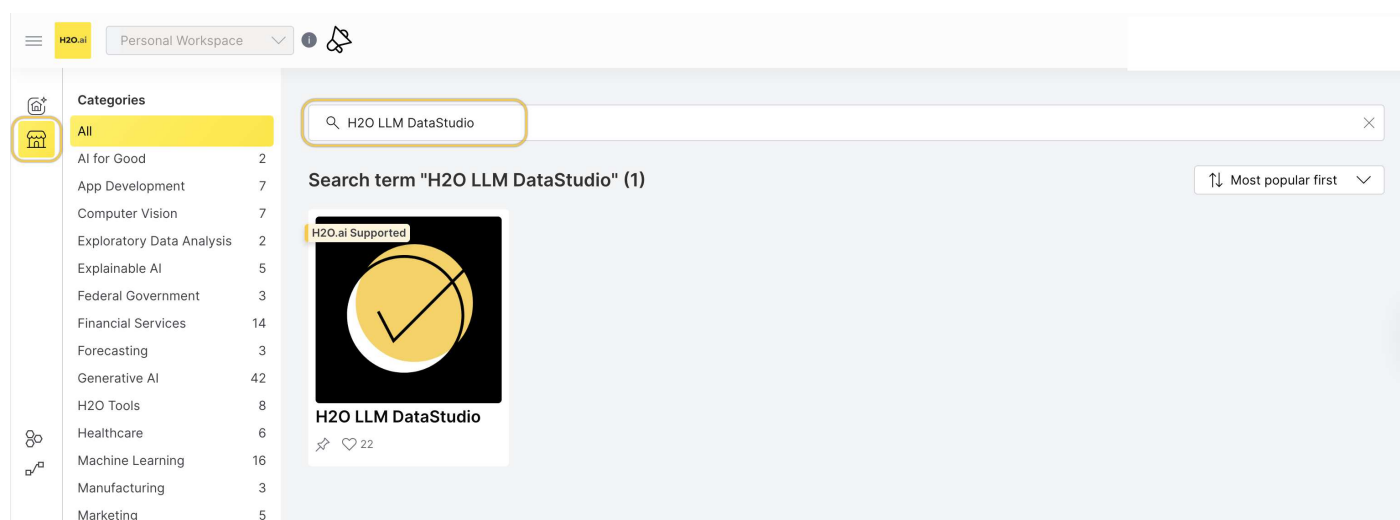
- [Step 1: Access HAIC](#)
- [Step 2: Search H2O LLM DataStudio](#)
- [Step 3: Run H2O LLM DataStudio](#)
- [Step 4: H2O LLM DataStudio instance](#)

Step 1: Access HAIC

Access your H2O AI Cloud (HAIC) account.

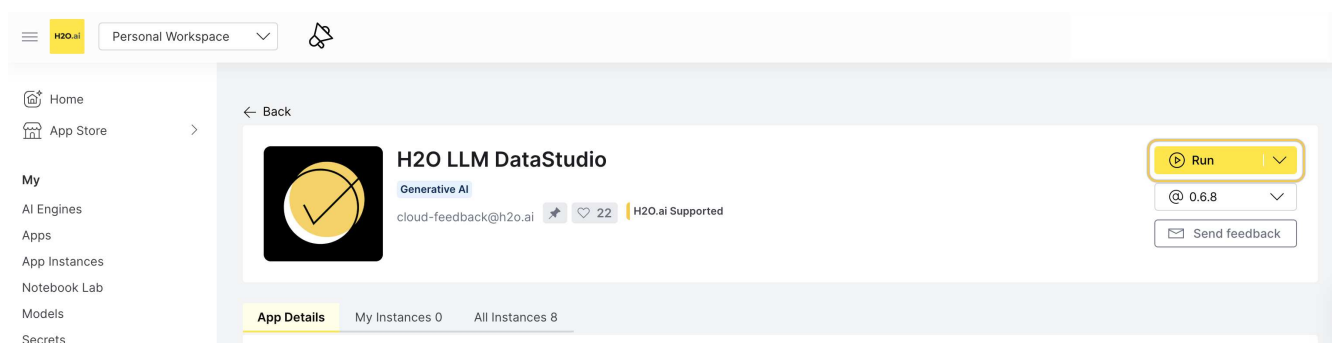
Step 2: Search H2O LLM DataStudio

1. In **HAIC** navigation menu, click **App Store**.
2. In the *app store* search bar, search H2O LLM DataStudio.
3. Now, select the H2O LLM DataStudio tile. Details about H2O LLM DataStudio appear.



Step 3: Run H2O LLM DataStudio

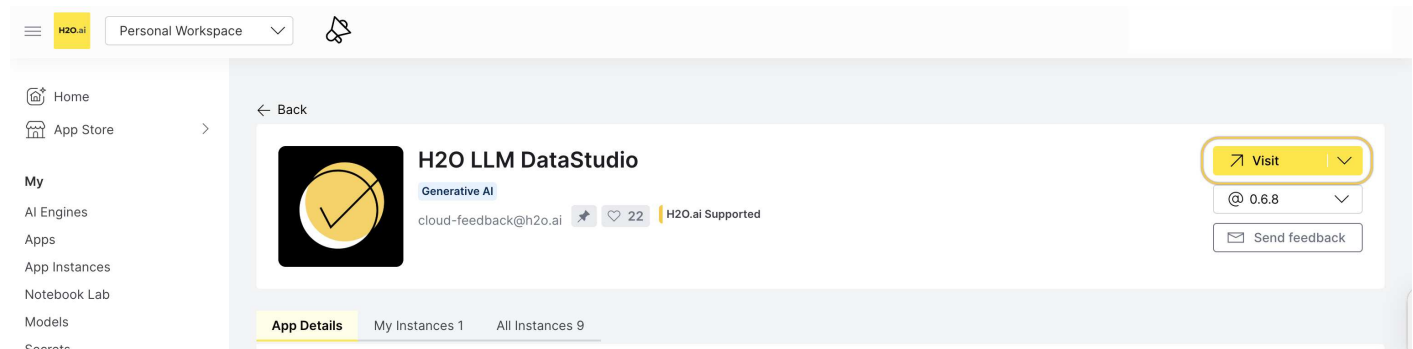
1. To start an H2O LLM DataStudio instance, click **Run**.



Step 4: H2O LLM DataStudio instance

Now, the H2O AI Cloud is starting an instance of H2O LLM DataStudio for you. While you have a starting/running instance, the **Run** button changes its name to **Visit**.

1. To open H2O LLM DataStudio in a new tab, click **Visit**.



Note:

- The latest version of H2O LLM DataStudio is preselected.
- The instance takes a couple of minutes to get started; you can use the application instance right after.

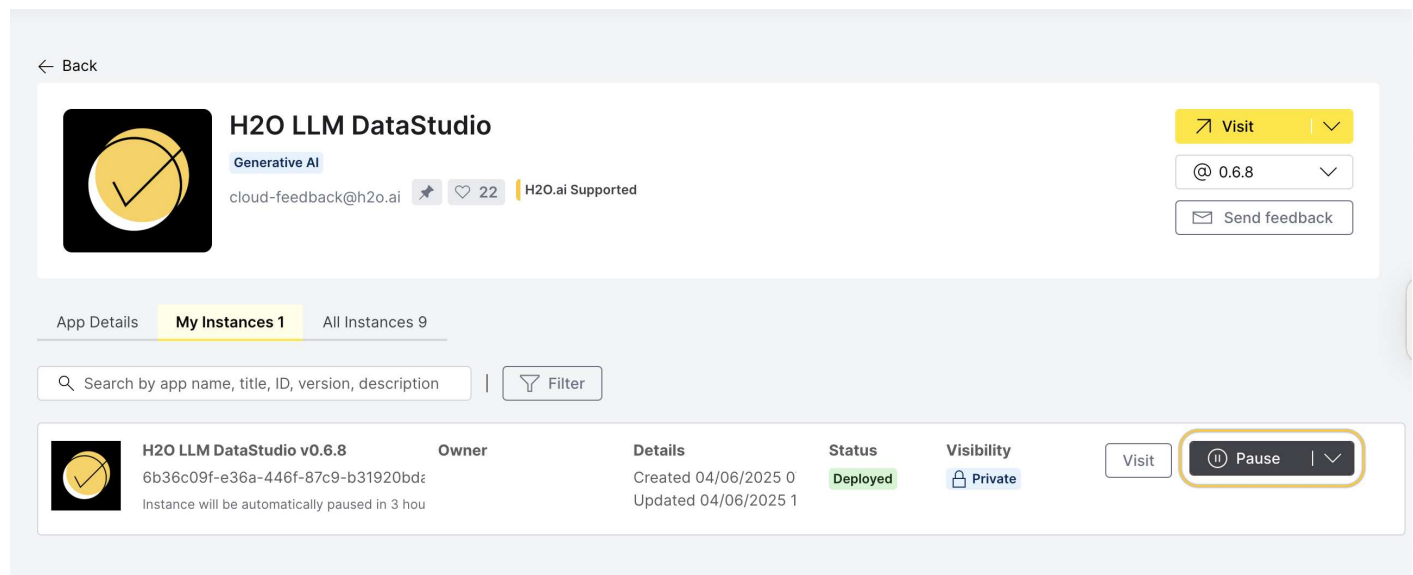
Pause or terminate instance

You can pause or terminate an instance of H2O LLM DataStudio.

- **Pause:** Pausing an instance saves computational time and resources.
- **Terminate:** Terminating an instance deletes the instance permanently.

Pause You can **Pause** an instance in the app details page:

- In the **app details** page, click the **My instances** tab.
- Locate the instance you want to pause.
- Click **Pause**.



You can also **Pause** an instance in the **My App Instances** page:

- In the **H2O AI Cloud** menu, click **App Instances**.
- Locate the instance you want to pause.
- Click **Pause**.

My App Instances
You have 5 instances

Search by app name, title, ID, version, description | Filter

App Icon	App Name	ID	Details	Status	Visibility	Actions
	H2O LLM DataStudio v0.6.8	6b36c09f-e36a-446f-87c9-b31920bda4e1	Created 04/06/2025 07:05:47 Updated 04/06/2025 11:19:47	Deployed	Private	Visit Pause

Terminate You can **Terminate** an instance in the app details page:

- In the **app details** page, click the **My instances** tab.
- Locate the instance you want to terminate.
- Click expand_more **Expand**.
- Click **Terminate**.

← Back

H2O LLM DataStudio
Generative AI
cloud-feedback@h2o.ai | 22 | H2O.ai Supported

Visit @ 0.6.8 Send feedback

App Details **My Instances 1** All Instances 9

Search by app name, title, ID, version, description | Filter

App Icon	App Name	Owner	ID	Details	Status	Visibility	Actions
	H2O LLM DataStudio v0.6.8	Owner	6b36c09f-e36a-446f-87c9-b31920bda4e1	Created 04/06/2025 0 Updated 04/06/2025 1	Deployed	Private	Visit Pause Terminate

You can also **Terminate** an instance in the **My App Instances** page:

- In the **H2O AI Cloud** menu, click **App Instances**.
- Locate the instance you want to terminate.
- Click expand_more **Expand**.
- Click **Terminate**.

My App Instances
You have 5 instances

Search by app name, title, ID, version, description | Filter

App Icon	App Name	ID	Details	Status	Visibility	Actions
	H2O LLM DataStudio v0.6.8	6b36c09f-e36a-446f-87c9-b31920bda4e1	Created 04/06/2025 07:05:47 Updated 04/06/2025 11:19:47	Deployed	Private	Visit Pause Terminate
	H2O Hydrogen Torch v1.5.0	2dd15882-b5c1-4b27-904c-aed256dbafe4	Created 29/05/2025 18:08:07 Updated 29/05/2025 22:08:51	Paused	All Users	Visit Edit Postpone auto-pause Terminate
	H2O Label Genie v1.1.0	d80d3497-7933-46ed-b6e6-f451bfde8acc	Created 03/04/2025 23:40:42	Paused	All Users	Visit Terminate

Data curation flow

Overview

The LLM DataStudio's **Curate** component is a no-code capability to build structured LLM datasets from unstructured data. You can import documents in PDFs, DOCs, Audio and Video file formats and convert those documents to question-answer pairs, summarization pairs and file summaries for downstream tasks.

H2O LLM DataStudio utilizes the Llama2 LLM model family to generate 10 high-quality question-answer pairs from each chunk of text within a document. By default, each chunk is 4000 characters long and has a small overlap with adjacent chunks.

Note: The dataset curation component supports documents in multiple languages including English, Spanish, French, German, Italian and Portuguese.

The flow of the data curation process of H2O LLM DataStudio, can be summarized in the following sequential steps:

- [Step 1: Create a new curate project](#)
- [Step 2: Upload documents](#)
- [Step 3: Perform the configuration and run pipeline](#)
- [Step 4: Use the output for data preparation](#)

Each of the steps given above has been summarised in the sections below.

Note: Before starting with Data curation, integrate **h2ogpte** by providing the required credentials. For more information, see [Settings](#).

Step 1: Create a new curate project

The first step in data curation process is creating a new **Curate** project. Click **New** on **New Project / Curate Data for LLMs** page. Provide the project name and the description. To learn how to create a new curate project, see [Create a new project for data curation](#).

Step 2: Upload documents

As the second step in data curation, select the task type of the experiment (question-answer/summarization/file summary) and upload the document or enter the webpage URL. H2O LLM DataStudio supports PDF, DOCX, TXT, MD file formats, MP3, M4A, WAV audio and video file formats. If you have multiple documents you can upload them in a ZIP file. For more information, see [Create a new project for data curation](#).

Step 3: Configure and run pipeline

After uploading the document, configure **Smart chunking (Fast)** and **Sampling ratio**.

If you turn **Smart chunking (Fast)** on, it will find the unique information within the dataset, filter some of the best chunks and generate question-answer pairs from those chunks. The smart chunking feature is recommended to use when you have large documents and want to generate question-answer pairs quickly.

You can manually specify the sampling ratio, or H2O LLM DataStudio will automatically set the sampling ratio depending on the length of the document. The sampling ratio is used to sample the documents based on the specified percentage. It selects the best chunks out of all available chunks for faster question-answer pair generation. If the sampling ratio is set to 0, LLM Data Studio will automatically choose the best ratio. However, it is recommended to set the sampling ratio to a value greater than or equal to 0.5.

After configuring the uploaded document, click **Run pipeline**. For more information, see the [instructions for creating a new project for data curation](#).

Step 4: Use the output for data preparation

Once the output is generated from Curate, you can input the dataset to the [data preparation flow](#). Inside the new Curation project, click file_open (Publish as Preparation Project) to publish the new Curation project as a Preparation project in the data preparation flow. For more information, see [View a specific Curation project](#).

Video guide

Watch [this](#) video guide for a walkthrough of the H2O LLM DataStudio interface and to learn more about its data curation process.

Data preparation flow

Overview

Data preparation involves a series of steps to transform and clean your dataset. The flow of data preparation using H2O LLM DataStudio can be summarized in the following sequential steps:

- [Step 1: Ingest data](#)
- [Step 2: Build the workflow](#)
- [Step 3: Configure the parameters](#)
- [Step 4: Review and execute](#)
- [Step 5: Analyze the output](#)
- [Step 6: Compare datasets](#)

Each of the steps given above has been summarised in the sections below.

Step 1: Ingest data

As the first step in the data preparation flow, upload your datasets to H2O LLM DataStudio in the specified format for the project type. Select the appropriate dataset files and map column names in your dataset to the required formats. It is essential to ensure that your dataset is compatible with the application's supported formats.

To learn about data ingestion, see [Data ingestion](#).

Step 2: Build the workflow

As the second step in the data preparation flow, drag the required steps to form the desired sequence. The workflow builder lets you to configure the order of data preparation steps. It lets you to define how the dataset will be prepared and transformed according to your requirements.

To learn how to build the workflow, see [Workflow builder](#).

Step 3: Configure the parameters

As the third step in the data preparation flow, customize the behavior of the function by setting parameters. You can use default parameters or configure them based on your specific requirements and the characteristics of your dataset.

To learn how to configure the parameters, see [Configuration](#).

Step 4: Review and execute

As the fourth step in the data preparation flow, carefully review the configured parameters to ensure accuracy. Once you are satisfied, initiate the execution of the workflow by clicking the **Run pipeline** button at the bottom of the page. The application will process the dataset according to the defined steps and parameters.

To learn how to review and execute the workflow, see [Review and execute](#).

Step 5: Analyze the output

As the fifth step in the data preparation flow, take time to review and analyze the output to ensure that it meets your expectations. You can export the obtained output dataset in JSON or CSV file format.

To learn about the generated resulting dataset, see [Output](#).

Step 6: Compare datasets

Using the **Insights** tab, you can compare the input and output datasets and the new columns generated. To learn about comparing datasets, see [Insights](#).

Video guide

Watch [this](#) step-by-step video guide to learn more about data preparation using LLM DataStudio.

Tutorials

Learn about H2O LLM DataStudio a no-code application and toolkit designed to streamline data curation, preparation, and augmentation tasks for large language models (LLMs).

Learning path

graph LR;

```
User[User] --> A1[Tutorial 1: Dataset preparation for question-answering models];
User[User] --> B1[Tutorial 2: Dataset preparation for text summarization models];
User[User] --> C1[Tutorial 3: Dataset preparation for instruct tuning models];
User[User] --> D1[Tutorial 4: Dataset preparation for human - bot conversations models];
User[User] --> E1[Tutorial 5: Dataset preparation for continued pretraining models];
```

```
%% Apply custom color to nodes
```

```
style User fill:#FEC925;
```

```
style A1 fill:#FEC925;
```

```
style B1 fill:#FEC925;
```

```
style C1 fill:#FEC925;
```

```
style D1 fill:#FEC925;
```

```
style E1 fill:#FEC925;
```

```
%% Add links to each node
```

```
click A1 "/h2o-llm-data-studio/tutorials/tutorial-1a"
```

```
click B1 "/h2o-llm-data-studio/tutorials/tutorial-2a"
```

```
click C1 "/h2o-llm-data-studio/tutorials/tutorial-3a"
```

```
click D1 "/h2o-llm-data-studio/tutorials/tutorial-4a"
```

```
click E1 "/h2o-llm-data-studio/tutorials/tutorial-5a"
```

The H2O LLM DataStudio tutorials are available for all the supported workflows. The workflows include:

Question and Answer

- [Tutorial: Preparation of a dataset for the problem type of Question Answering](#) > This tutorial describes the process of preparing a dataset that consists of contextual information, questions, and corresponding answers.

Text Summarization

- [Tutorial: Preparation of a dataset for the problem type of Text Summarization](#) > This tutorial describes the process of preparing a dataset that consists of articles and their associated summaries.

Instruct Tuning

- [Tutorial: Preparation of a dataset for the problem type of Instruct Tuning](#) > This tutorial describes the process of preparing a dataset that consists of prompts and their respective responses.

Human - Bot Conversations

- [Tutorial: Preparation of a dataset for the problem type of Human - Bot Conversations](#) > This tutorial describes the process of preparing a dataset comprising multiple dialogues between human users and chatbots.

Continued PreTraining

- [Tutorial: Preparation of a dataset for the problem type of Continued PreTraining](#) > This tutorial describes the process of preparing datasets with extensive texts for further pretraining of language models.

Data ingestion

Overview

Data preparation involves a series of steps to transform and clean your dataset. The first step is data ingestion.

Instructions

To ingest data, consider the following instructions:

1. On the H2O LLM DataStudio left navigation menu, click **Projects**.
2. Click the name of the project that you created before.
3. On the H2O LLM DataStudio left navigation menu inside the project, Click **Ingestion**.
4. The Ingestion tab will appear. Visit the **Add new dataset** section.
5. In the **Dataset name** text box, enter the name of the dataset.
6. In the **Description** text box, enter a description for the dataset.
7. Click **Browse** in the **Upload file** section to select and upload the dataset file in the supported file format. **Note:**
 - You can upload a single file, multiple files, or multiple files compressed in a **zip** file.
 - The supported file formats are **csv**, **json**, **jsonl** and **parquet**.
 - The size of the file should be less than 1GB.
 - There are required columns for each workflow type. For more information, see [Supported problem types](#).
 - To find sample datasets, see [Example datasets](#) in the H2O LLM DataStudio GitHub repository.
8. Click **Ingest** to ingest the dataset into the data preparation flow.

[← Back to Proje...](#)

- 1. Ingestion
- 2. Workflow
- 3. Configurati...
- 4. Review
- 5. Output
- 6. Insights

New project / Ingestion

> 1.1 Project Details

✓ 1.2 Add New Datasets

Add Datasets

Dataset Name

tweet qua

Dataset Description

tweet qua

Upload Datasets

Can upload single file or multiple files or a zip file with multiple files (csv, json, jsonl,parquet max size: 1GB)

tweet_qa.csv

Browse

+ Ingest

× Discard

Mandatory Columns: **question, answer, context**

Expected Input File (Workflow: Question Answering)

i

'csv', 'json', 'jsonl','parquet' file type, less than 1GB in size

- Once you add the dataset to the project, you can check the available datasets from the right side of the page. Select the dataset and click **Dataset preview** to preview the dataset.
- Under **Configure columns**, select the relevant columns for question, answer, and context from the given options.
- Click **Save**.

← Back to Proje...

1. Ingestion

2. Workflow

3. Configurati...

4. Review

5. Output

6. Insights

New project / Ingestion

> 1.1 Project Details

> 1.2 Add New Datasets

✓ Dataset(s) added successfully! Please Configure the Columns

> 1.3 Configure Datasets

View dataset (Total: 1)

tweet qua0

Preview

Delete

Configure Columns

Question Column * ⓘ
question

Answer Column * ⓘ
answer

Context Column * ⓘ
context

Save

Workflow builder

Overview

The workflow builder allows you to configure the order of data preparation steps. This lets you to define how the dataset will be prepared and transformed according to your specific requirements.

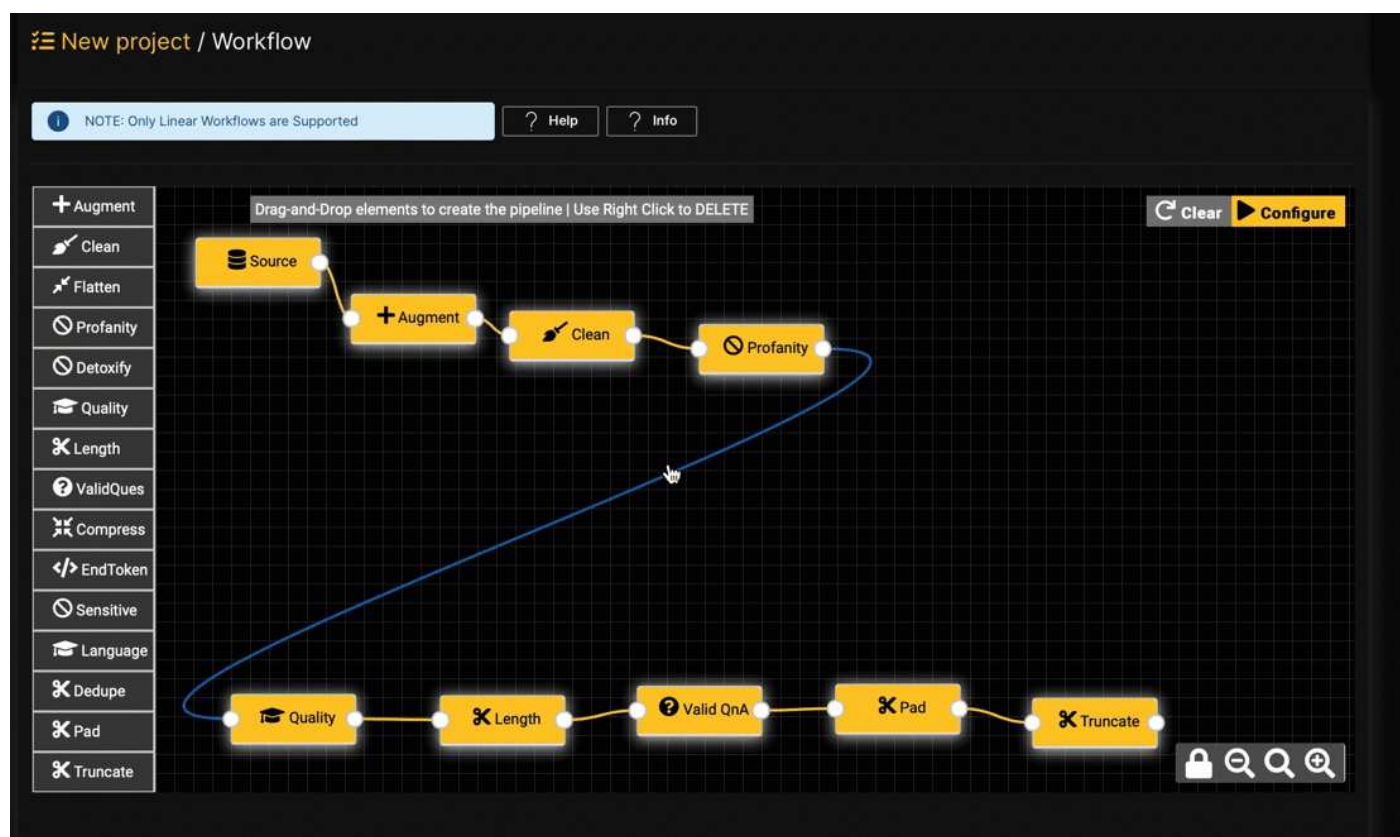
Instructions

To build the workflow, consider the following instructions:

1. On the H2O LLM DataStudio left navigation menu, click **Projects**.
2. Click the name of the project that you created before.
3. On the H2O LLM DataStudio left navigation menu inside the project, Click **Workflow**.
4. Drag the required steps (nodes) from the left panel to form the desired sequence to create the pipeline of tasks. For more information about the data preparation steps, see [Supported functionalities](#).

Note:

- To delete a node or connection, select the node or connection and right-click to delete.
 - To zoom in and zoom out, use the buttons in the bottom right corner.
 - To clear the workflow and reset the screen, click on the **Clear** button.
5. Click **Configure** to run the workflow and move to the next screen.



Video guide

Watch [this](#) video guide to learn more about projects and the **workflow builder** tool for data preparation in LLM DataStudio.

Configuration

Overview

For each data preparation step, you can customize the behavior of the function by configuring parameters. You can use default parameters or configure them based on your specific requirements and the characteristics of your dataset.

The configuration settings can be different from one problem type to another according to the required data preparation steps.

Instructions

1. On the H2O LLM DataStudio left navigation menu, click **Projects**.
2. Click the name of the project that you created before.
3. On the H2O LLM DataStudio left navigation menu inside the project, Click **Configuration**.
4. Once all the parameters are configured, click **Review** to move to the next step.

General configurations

The following configurations can be used for all the workflow/task types.

Source datasets / Augmentation

Select the dataset(s) you need to use. If you select multiple datasets, the datasets will undergo [data augmentation](#). Data augmentation involves mixing the current training dataset with additional datasets. It helps increase the diversity and size of the training dataset and enhances the model's performance. If none of the datasets are chosen, the first dataset will be used by default.

1. Click **Select** to select an augmentation dataset from the [Augment](#) to augment with your input dataset.
2. Click **Add** to add the dataset.

Filter by column

You can use this configuration to filter rows from the dataset based on another categorical column. It enables you to filter out curation pairs marked as irrelevant.

Text cleaning

This configuration is responsible for text cleaning and preprocessing tasks. Configure the data cleaning steps for the text by selecting or deselecting the tags. The **H2O LLM DataStudio** provides options to,

- remove newline characters,
- remove whitespaces,
- lowercase capital letters,
- remove URLs,
- remove HTML characters, and
- remove ASCII characters.

Profanity check

Adjust the slider to control the level of sensitivity in detecting offensive language within the text. This configuration helps filter out content that may be offensive or inappropriate for certain applications. For example, if the threshold is set to 0.9, any text in which profanity detection exceeds 0.9 will be filtered out.

Detoxify

Detoxify parameter checks for toxicity in the texts and filters based on the threshold. It includes four sub-configurations.

- **Acceptable toxicity threshold:** Adjust the slider to control the level of toxicity within the text. For example, if the value is set to 0.9, any text having value above 0.9 will be dropped.
- **Acceptable identity attack threshold:** Adjust the slider to control the level of 'identity attack' within the text. Example, if the value is set to 0.9, any text having value above 0.9 will be dropped.

- **Acceptable insult threshold:** Adjust the slider to control the level of ‘insult’ within the text. Example, if the value is set to 0.9, any text having value above 0.9 will be dropped.
- **Acceptable threat threshold:** Adjust the slider to control the level of ‘threat’ within the text. Example, if the value is set to 0.9, any text having value above 0.9 will be dropped.

Note: If there is no GPU, the detoxify function will take a long time to run.

Length check

Adjust the sliders to set the minimum and maximum text lengths for each column of the dataset to ensure the text falls within the desired length criteria. This configuration helps to ensure that the input data meets specific length requirements to truncate or pad the text to a desired length for model compatibility.

Text quality check

Adjust the slider to set the minimum and maximum text grade to include texts within the desired grade range to ensure the quality of the texts. This configuration assesses the quality or appropriateness of the data. It evaluates various criteria, such as grammar, relevance, or coherence, to identify potential issues or areas for improvement in the dataset.

Sensitive info check

Add the sensitive information you wish to drop from the text. The selected sensitive or confidential information will be removed from the text to ensure privacy and data protection.

The available options for the sensitive info check are,

- Email address
- Phone number
- Crypto wallet number
- The International Bank Account Number (IBAN)
- IP address
- Named entity removal

Data Anonymization Turn the toggle **On** to anonymize sensitive information. When enabled, all the sensitive data will be transformed into a format that cannot be traced back to the original data.

If the toggle is turned **Off**, the system will completely remove the sensitive information instead of anonymizing it.

Bias check

Use the slider to set the desired bias threshold. The threshold determines the level of bias that is acceptable within the text.

Example: If you set the threshold to 0.9, any text with a bias score detected to be above 0.9 will be automatically dropped from the dataset.

Add your own code

Upload your own python cleaning function inside a .py file. The code needs to be wrapped inside the following function to work:

```
def custom_function(df, text_columns):
```

You can refer the sample Python files in the [H2O LLM DataStudio GitHub repository](#) and create your own Python code. Additionally, you can download those examples and upload them to the application according to your problem type.

Pad sequence

Define the maximum padding length for sequences. This configuration is used for sequence padding. It adds padding tokens to sequences to make them equal in length. It is often necessary for efficient batch processing in neural networks.

Truncate sequence

This configuration is responsible for truncating or cutting the input text to a specific length. It removes excess text beyond the desired length, ensuring consistency and compatibility with model requirements. It includes three sub-configuration.

- Truncate max length: Define the maximum truncating length so that the sequences longer than the specified length will be truncated.
- Truncate ratio: Set the truncation ratio to summarize the sequence and extract the most informative parts using TextRank so that the most informative parts will not be truncated from the sequence.
- Model based: Toggle the button to enable or disable the model-based truncation.

Configurations for question-answering

The following configuration can only be used for the **Question and answer** task type.

Question relevance check

Toggle the button to determine whether the question for question-answer pairs is actually a question. If it finds that there is no question in the pair, the app filters out that particular question-answer pair from further processing.

Configurations for text summarization

The following configurations can only be used for the **Text summarization** task type.

Filter compression

Adjust the slider to set the threshold for compression ratio between the article and its summary. This configuration calculates the compression ratio between the article and its summary and removes the article-summary pairs with a compression ratio above a certain threshold. It helps in creating high-quality article-summary pairs for text summarization models.

Add special tokens

Toggle the button to determine whether or not to add the start and end token to the texts to indicate the beginning and end of each text sequence. It is commonly used in sequence-to-sequence models or language generation tasks.

Configurations for human-bot conversation

The following configuration can only be used for the **Human-Bot conversation** task type.

Flatten conversation

Toggle the button to determine whether to flatten the human-bot conversation dataset. When enabled, the dataset will be flattened into a single sequence, disregarding the individual turns.

Review and execute


Overview

Before executing the data preparation workflow, you can review the configured parameters to ensure accuracy.

Instructions

To review the configurations, consider the following instructions:

1. On the H2O LLM DataStudio left navigation menu, click **Projects**.
2. Click the name of the project that you created before.
3. On the H2O LLM DataStudio left navigation menu inside the project, Click **Review**.
4. Once you are satisfied with configured parameters, click **Run pipeline** to initiate the execution of the workflow. The application will process the dataset according to the defined steps and parameters.

 File Uploaded: MigraineTreatment.zip

Smart Chunking


Recommended to use when document files are hundreds/thousands of pages.


Perform Smart Chunking (Fast) ☐ Off

Sampling Ratio

0

Note: If the ratio=0, best ratio is chosen automatically.

 Run Pipeline

 Discard

Output

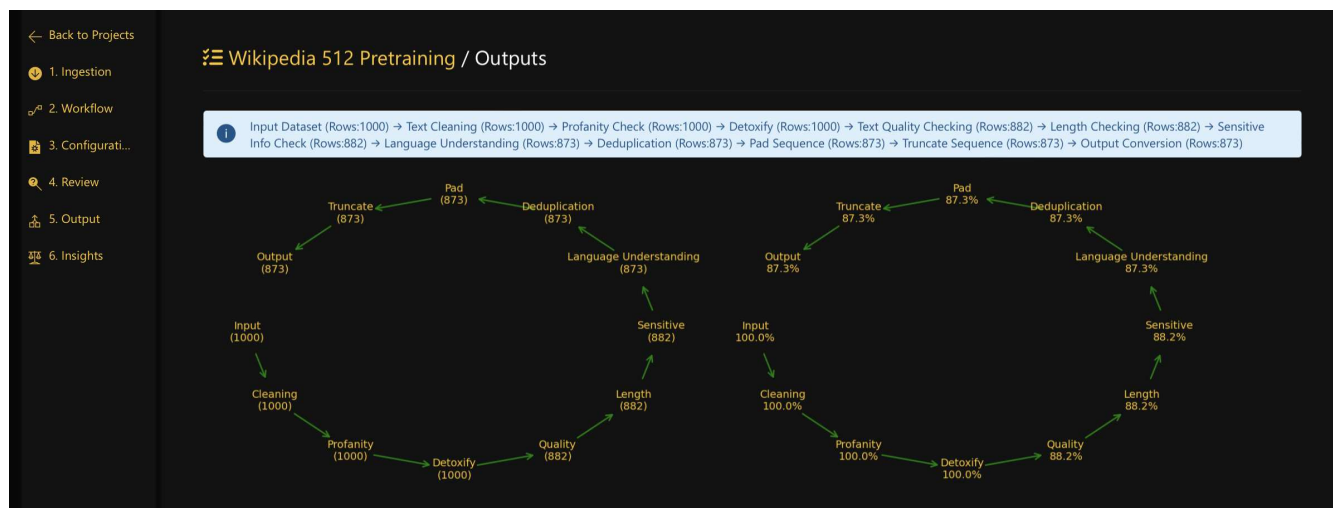
Overview

After the data preparation process is completed, a resulting dataset will be generated. Take the time to review and analyze the output to ensure it meets your expectations.

Instructions

To review and analyze the output, consider the following instructions:

1. On the H2O LLM DataStudio left navigation menu, click **Projects**.
2. Click the name of the project that you created before.
3. On the H2O LLM DataStudio left navigation menu inside the project, Click **Output**.



Note:

- The first section of the **Output** page graphically represents the number of rows and the percentage of rows against input rows in each data preparation step.
 - The second section of the **Output** page consists of a preview of the final dataset with the top 100 rows.
 - Enabling **Show intermediate features** will display values related to different analysis runs, such as the flesch score for text quality.
4. Click **Download CSV** to download the output dataset in the CSV file format.
Click **Download configurations** to download the project workflow settings.
Click **Export to H2O Drive** to export the output dataset to H2O Drive.
 5. Click **View insights** to view and compare the input and output datasets, as well as the newly generated columns.

Insights

Overview

Using the **Insights** tab, you can compare the input and output datasets and the new columns generated to see text cleaning differences, selected columns, and dataset row differences.

Instructions

To compare the datasets, consider the following instructions:

1. On the H2O LLM DataStudio left navigation menu, click **Projects**.
2. Click the name of the project that you created before.
3. On the H2O LLM DataStudio left navigation menu inside the project, Click **Insights**.
4. On the left navigation bar, click **Insights**.

The screenshot displays the 'Insights' tab in the H2O LLM DataStudio interface. The left sidebar contains a navigation menu with options: 'Back to Project...', '1. Ingestion', '2. Workflow', '3. Configuration...', '4. Review', '5. Output', and '6. Insights' (which is highlighted). The main content area is titled 'OpenAssistant Conversations / Insights' and is divided into two panels: 'Input Dataset' and 'Output Dataset'.

Input Dataset: The dataset shape is (2022, 7). It shows a 'Raw Dataset Snapshot (Top 10 rows)' with columns 'message_id' and 'parent_id'. The first row has 'message_id' as 'e0cf4fa2-e944-44fb-b502-11f155062092' and 'parent_id' as 'nan'. The second row has 'message_id' as '7d05acb7-9360-458c-8a1d-c0b6492b8f8a' and 'parent_id' as 'e0cf4fa2-e944-44fb-b502-11f155'.

Output Dataset: The dataset shape is (95, 1). It shows an 'Output Dataset Snapshot (Top 10 rows)' with a single column 'text'. The text content is a conversation between a human and a bot about learning Mandarin and writing Chinese characters. A yellow button labeled 'View Dropped Rows' is located below the text.

Click **View dropped rows** to view the dropped rows from the input dataset.

Dropped Rows

Showing top 100 rows

question	context	answer	Dropped_Reason
	privacy policy we appreciate the opportunity to interact with you on the internet and are committed to protecting and safeguarding your privacy. the purpose of this privacy policy is to inform you about the types of information we might collect about you when you visit jessica style, how we may use that information, whether we disclose it to anyone, and the choices you have regarding our use of, and your ability to correct, that information. what information we collect and how we use it any information we collect on jessica style generally falls into the following two categories: personally identifiable information or aggregate information. \. personally identifiable information this refers to information that lets us know who you are or things specifically about you. a. visitors you can browse jessica style without revealing any personally identifiable information. b. registration if you want to register and		

Download CSV

Click the **Download CSV** button below the **Dropped rows** table to download the dropped rows in CSV file format.

Tutorial 1A: Dataset preparation for Question Answering

Overview

This tutorial describes the process of preparing a dataset that consists of contextual information, questions, and corresponding answers. This task is essential for training question-answering models that can accurately respond to queries based on the provided context. The dataset preparation process focuses on building a well-structured dataset for training question-answering systems.

Prerequisites

- [Access H2O LLM DataStudio via H2O AI Cloud \(HAIC\)](#).
- Review [Data preparation flow](#).
- GPU is recommended to use the [Detoxify](#) technique in data cleaning.

Step 1: Explore the project

For this tutorial, we are going to use the prebuilt **Website Privacy** project, which consists of an online privacy policy question and answer dataset. Let's explore the project:

1. On the H2O LLM DataStudio navigation menu, click **Prepare**.
2. Explore the table with detailed information about both prebuilt projects and projects that you created. For more information on viewing the projects, see [View projects](#).
3. Click on the **Website Privacy** project name and navigate to the corresponding data preparation steps.

Step 2: Ingest data

The first step of data preparation is ingesting data by uploading your datasets. In this tutorial, we are proceeding with the preloaded *Privacy dataset*. Let's configure the preloaded dataset.

1. To preview the dataset, click **Preview** under **Configure datasets** section. It will show you the top 100 rows of the dataset.
2. Under the **Configure columns** section, select the relevant columns for question, answer, and context from the given options.
3. Click **Save**.

Step 3: Build the workflow

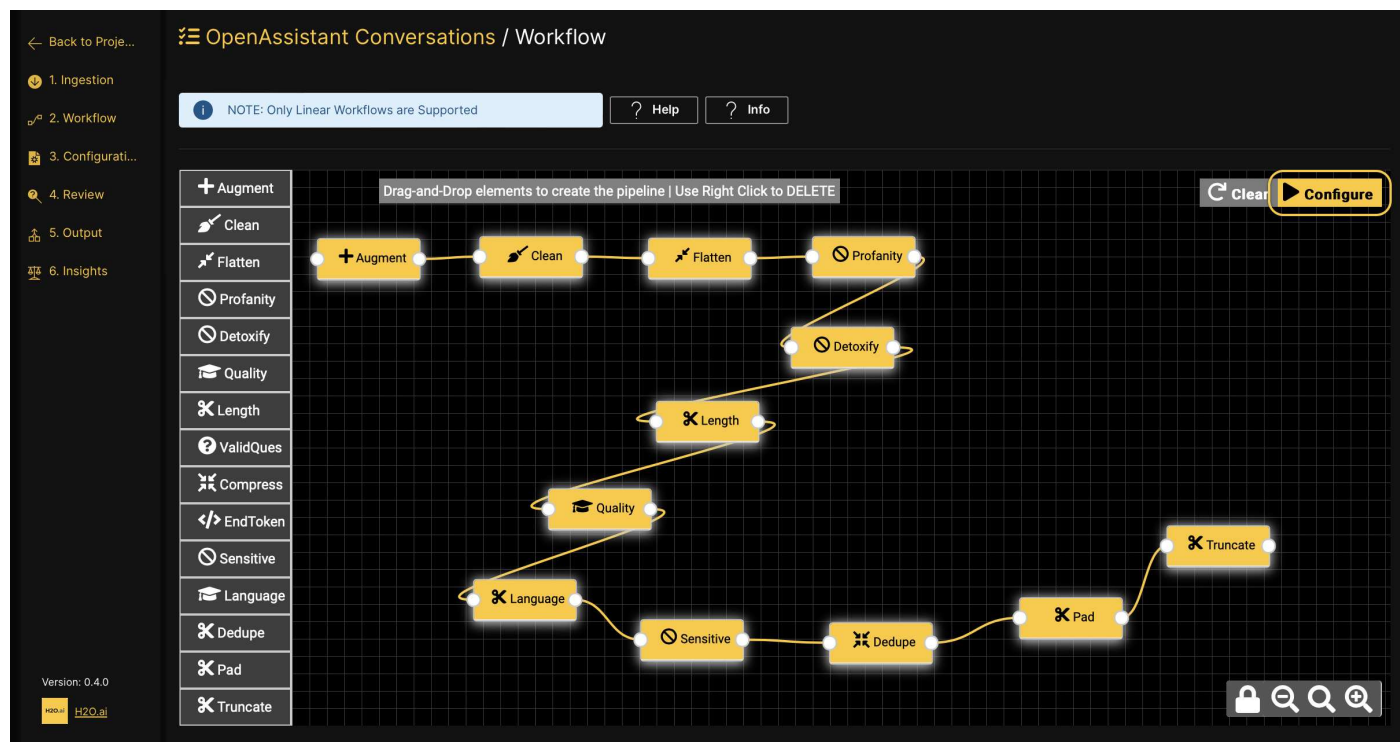
Using the workflow builder tool, let's configure the order of data preparation steps.

1. Inside the **Website Privacy** project, click **Workflow** On the left navigation menu.
2. Drag and drop the required steps in the following sequence. To learn more about the workflow builder tool, see [Workflow builder](#).

Tip: Workflow Augmentation > Text cleaning > Profanity Check > Detoxify > Length Check > Text Quality Check > Sensitive Info Check > Question Relevance > Language Understanding > Deduplication > Padding sequence > Truncate sequence > Output

Note: If there is no GPU, detoxify function will take a long time to run.

3. After configuring the order of data preparation steps, click **Configure** to run the workflow.



Step 4: Configure the parameters

H2O LLM DataStudio allows you to customize the behavior of the function for each data preparation step by setting parameters. Let's use the default parameter configurations for this tutorial.

1. Inside the **Website Privacy** project, click **Configuration** On the left navigation menu.
2. Go through each parameter configuration. To learn more about the available configurations, see [Configuration](#).
3. Once all the parameters are configured, click **Review** to move to the next step.

Step 5: Review the configured parameters and execute the workflow

Let's review the configured parameters before executing the data preparation workflow to ensure accuracy.

1. Inside the **Website Privacy** project, click **Review** On the left navigation menu.
2. Once you are satisfied with configured parameters, click **Run pipeline** to initiate the execution of the workflow.

Step 6: Review and analyze the output

After the data preparation process is completed, a resulting dataset is generated. Let's take time to review and analyze the output.

1. Inside the **Website Privacy** project, click **Output** On the left navigation menu.
 - The **Output** page graphically represents the number of rows and the percentage of rows against input rows in each data preparation step.
 - The **Output** page consists a preview of the final dataset with the top 100 rows.
2. Click **Download CSV** to download the output dataset in the CSV file format.
3. Click **Export to H2O Drive** to export the output dataset to H2O Drive.

Step 7: Compare input and output datasets

As the final step, we can compare and see the differences between the input and output datasets. Let's take a look at text cleaning differences, selected columns and dataset row differences.

1. Inside the **Website Privacy** project, click **Insights** On the left navigation menu. You can compare the input dataset and output dataset.

Click **View dropped rows** to view the rows dropped from the input dataset.

Summary

In this tutorial, we learned how to prepare a dataset with question, answer, and context columns for the problem type of **Question Answering**. Also, we discovered that H2O LLM DataStudio lets you upload, prepare, and analyze your datasets, ultimately achieving your desired data transformation goals.

Video guide

Watch [this](#) video guide to learn more about prepping a question-answering dataset using H2O LLM DataStudio.

Tutorial 2A: Dataset preparation for Text Summarization

Overview

This tutorial describes the process of preparing a dataset that consists of articles and their associated summaries. This task is essential for training text summarization models that can generate succinct and informative summaries from lengthy text. The dataset preparation process focuses on building a well-structured dataset for training text summarization systems.

Prerequisites

- [Access H2O LLM DataStudio via H2O AI Cloud \(HAIC\)](#).
- Review [Data preparation flow](#).
- GPU is recommended to use the [Detoxify](#) technique in data cleaning.

Step 1: Explore the project

For this tutorial, we are going to use the prebuilt **CNN-DailyMail** project, which consists of a CNN/DailyMail news dataset. Let's explore the project:

1. On the H2O LLM DataStudio navigation menu, click **Prepare**.
2. Explore the table with detailed information about both prebuilt projects and projects that you created. For more information on viewing the projects, see [View projects](#).
3. Click on the **CNN-DailyMail** project name and navigate to the corresponding data preparation steps.

Step 2: Ingest data

The first step of data preparation is ingesting data by uploading your datasets. In this tutorial, we are proceeding with the preloaded *CNN/DailyMail news dataset*. Let's configure the preloaded dataset.

1. To preview the dataset, click **Preview** under **Configure datasets** section. It will show you the top 100 rows of the dataset.
2. Under the **Configure columns** section, select the relevant columns for article and summary from the given options.
3. Click **Save**.

Step 3: Build the workflow

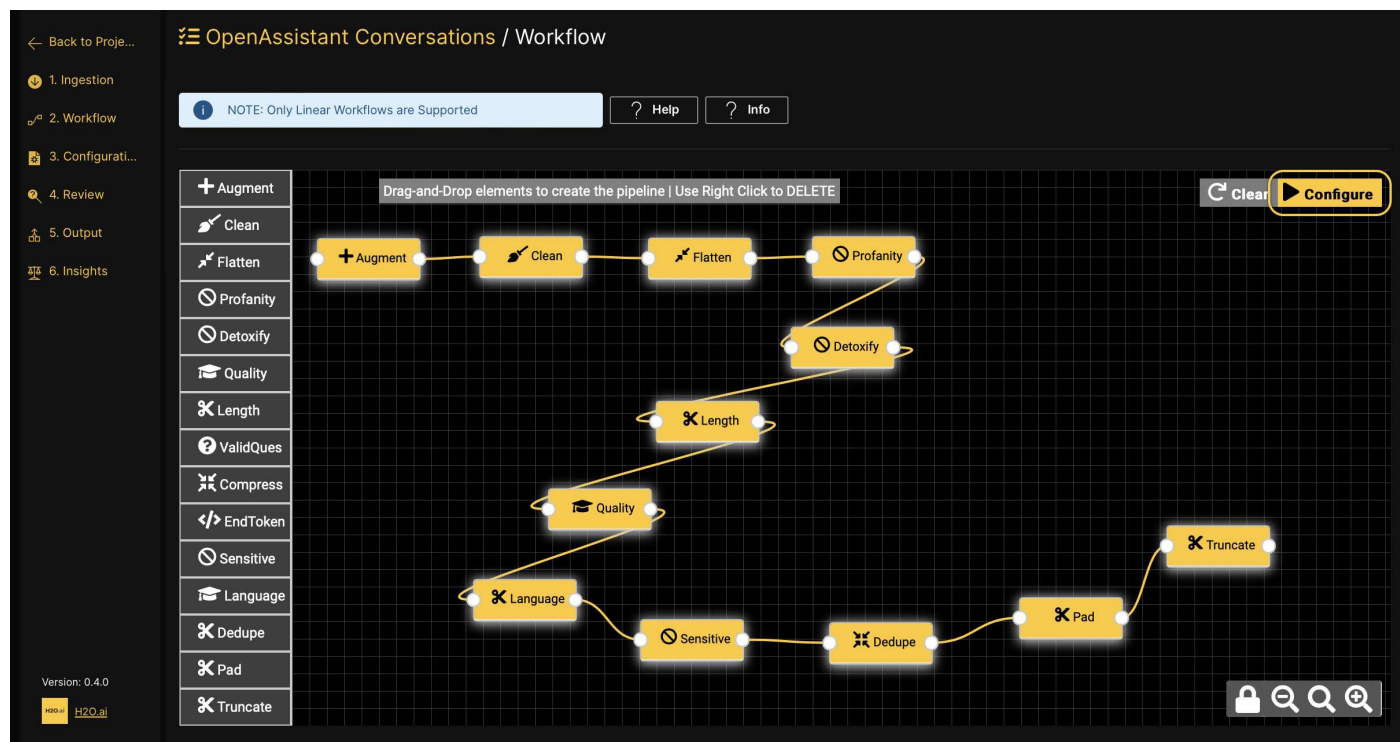
Using the workflow builder tool, let's configure the order of data preparation steps.

1. Inside the **CNN-DailyMail** project, click **Workflow** On the left navigation menu.
2. Drag and drop the required steps in the following sequence. To learn more about the workflow builder tool, see [Workflow builder](#).

Tip: Workflow Augmentation > Text cleaning > Profanity Check > Detoxify > Length Check > Text Quality Check > Sensitive Info Check > Filter Compression Ratio > Language Understanding > Deduplication > Boundary marking > Padding sequence > Truncate sequence > Output

Note: If there is no GPU, detoxify function will take a long time to run.

3. After configuring the order of data preparation steps, click **Configure** to run the workflow.



Step 4: Configure the parameters

H2O LLM DataStudio allows you to customize the behavior of the function for each data preparation step by setting parameters. Let's use the default parameter configurations for this tutorial.

1. Inside the **CNN-DailyMail** project, click **Configuration** On the left navigation menu.
2. Go through each parameter configuration. To learn more about the available configurations, see [Configuration](#).
3. Once all the parameters are configured, click **Review** to move to the next step.

Step 5: Review the configured parameters and execute the workflow

Let's review the configured parameters before executing the data preparation workflow to ensure accuracy.

1. Inside the **CNN-DailyMail** project, click **Review** On the left navigation menu.
2. Once you are satisfied with configured parameters, click **Run pipeline** to initiate the execution of the workflow.

✔ File Uploaded: MigraineTreatment.zip

Smart Chunking

Recommended to use when document files are hundreds/thousands of pages.

Perform Smart Chunking (Fast) ☐ Off

Sampling Ratio

0

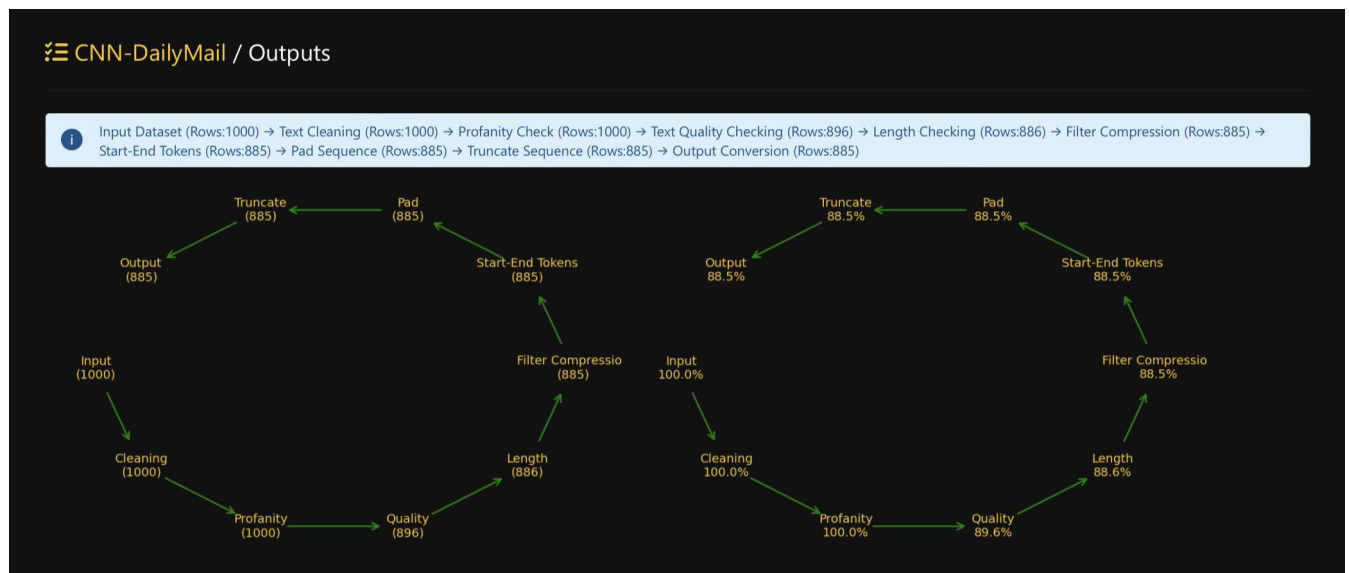
Note: If the ratio=0, best ratio is chosen automatically.

✔ Run Pipeline
✕ Discard

Step 6: Review and analyze the output

After the data preparation process is completed, a resulting dataset is generated. Let's take time to review and analyze the output.

- Inside the **CNN-DailyMail** project, click **Output** On the left navigation menu.
 - The **Output** page graphically represents the number of rows and the percentage of rows against input rows in each data preparation step.
 - The **Output** page consists a preview of the final dataset with the top 100 rows.



- Click **Download CSV** to download the output dataset in the CSV file format.
- Click **Export to H2O Drive** to export the output dataset to H2O Drive.

Showing top 100 rows of the final dataset

Show Intermediate Features ☐ Off

article	summary
<p>we all have our own ways of dealing with heartache, whether its drinking our cares away with friends, turning to the comfort of our families, or perhaps even buying a sports car. but one pensioner took a slightly more unorthodox approach after splitting from his wife by nailing a caravan on top of a motorised catamaran and sailing it around the english coast. the unusual vehicle, which also features a rather flimsy looking safety fence, was last spotted making its way along the swale between the isle of sheppey and kent coast. unusual sight: the pensioner has now taken to the water in a floating caravan, seen in recent days chugging along down the swale between the isle of sheppey and kent coast . puzzled locals, who have watched the bizarre contraption in disbelief, said the owner is a keen boatman in his '70s or 80s and lives in nearby sittingbourne. some residents had theorized that the man built the craft to appease his wife, who was said to hate sailing but love caravanning. however a close friend of the handyman said that the couple actually split two years ago, and his friend built the craft as a way of coping with his grief. and fisherman trevor ryan, 63, added: he bought the catamaran and then he fitted the caravan on it and launched it from iron wharf boatyard. close-up: the caravan has been fixed to the base of a catamaran fitted with a small engine - and while there is a flimsy-looking fence around the vehicle, there are no other obvious safety measures . he is about 70 or 80</p>	<p>_START_ caravan seen along the swale between isle of sheppey and kent coast . fixed to base of catamaran fitted with engine and surrounded by fence . witnesses say man is in his 70s or 80s and launched it from boatyard . _END_</p>

Download CSV Export to H2O Drive

Step 7: Compare input and output datasets

As the final step, we can compare and see the differences between the input and output datasets. Let's take a look at text cleaning differences, selected columns and dataset row differences.

1. Inside the **CNN-DailyMail** project, click **Insights** On the left navigation menu.

← Back to Project | CNN-DailyMail / Insights

1. Ingestion | 2. Workflow | 3. Configuration | 4. Review | 5. Output | 6. Insights

Input Dataset

Dataset Shape: (1000, 2)

Raw Dataset Snapshot (Top 10 rows)

article	summary
<p>Sally Forrest, an actress-dancer who graced the silver screen throughout the '40s and '50s in MGM musicals and films such as the 1956 noir <i>While the City Sleeps</i> died on March 15 at her home in Beverly Hills, California. Forrest, whose birth name was Katherine Feeney, was 86 and had long battled cancer. Her publicist, Judith Goffin, announced the news Thursday. Scroll down for video . Actress: Sally Forrest was in the 1951 Ida Lupino-directed film 'Hard, Fast and Beautiful' (left) and the 1956 Fritz Lang movie 'While the City Sleeps' A San Diego native, Forrest became a protege of Hollywood trailblazer Ida Lupino, who cast her in starring roles in films including the critical and commercial success <i>Not Wanted</i>, <i>Never Fear</i> and <i>Hard, Fast and Beautiful</i>. Some of Forrest's other film credits included <i>Bannerline</i>, <i>Son of Sinbad</i>, and <i>Excuse My Dust</i>, according to her IMDb page. The page also indicates Forrest was in multiple <i>Climax!</i> and</p>	<p>Sally Forrest, an actress-dancer w screen throughout the '40s and '5t films died on March 15 . Forrest, w Katherine Feeney, had long battlec native, Forrest became a protege c</p>

Output Dataset

Dataset Shape: (125, 2)

Output Dataset Snapshot (Top 10 rows)

article	summary
<p>we all have our own ways of dealing with heartache, whether its drinking our cares away with friends, turning to the comfort of our families, or perhaps even buying a sports car. but one pensioner took a slightly more unorthodox approach after splitting from his wife by nailing a caravan on top of a motorised catamaran and sailing it around the english coast. the unusual vehicle, which also features a rather flimsy looking safety fence, was last spotted making its way along the swale between the isle of sheppey and kent coast. unusual sight: the pensioner has now taken to the water in a floating caravan, seen in recent days chugging along down the swale between the isle of sheppey and kent coast . puzzled locals, who have watched the bizarre contraption in disbelief, said the owner is a keen boatman in his '70s or 80s and lives in nearby sittingbourne. some residents had theorized that the man built the craft to</p>	<p>_START_ caravan seen along the s</p>

View Dropped Rows

You can compare the input dataset and output dataset. Click **View dropped rows** to view the rows dropped from the input dataset.

Summary

In this tutorial, we learned how to prepare a dataset with article and their associated summary columns for the problem type of **Text Summarization**. Also, we discovered that H2O LLM DataStudio lets you upload, prepare, and analyze your datasets, ultimately achieving your desired data transformation goals.

Tutorial 3A: Dataset preparation for Instruct Tuning

Overview

This tutorial describes the process of preparing a dataset that consists of prompts and their respective responses. This task is essential for training models that can comprehend and adhere to given instructions and accurately respond to user prompts. The dataset preparation process focuses on building a well-structured dataset for training instruct tuning systems.

Prerequisites

- [Access H2O LLM DataStudio via H2O AI Cloud \(HAIC\)](#).
- Review [Data preparation flow](#).
- GPU is recommended to use the [Detoxify](#) technique in data cleaning.

Step 1: Explore the project

For this tutorial, we are going to use the prebuilt **Chat with AI** project, which is an AI Chatbot with prompt response. Let's explore the project:

1. On the H2O LLM DataStudio navigation menu, click **Prepare**.
2. Explore the table with detailed information about both prebuilt projects and projects that you created. For more information on viewing the projects, see [View projects](#).
3. Click on the **Chat with AI** project name and navigate to the corresponding data preparation steps.

Step 2: Ingest data

The first step of data preparation is ingesting data by uploading your datasets. In this tutorial, we are proceeding with the preloaded *Chatbot Instruction dataset*. Let's configure the preloaded dataset.

1. To preview the dataset, click **Preview** under **Configure datasets** section. It will show you the top 100 rows of the dataset.
2. Under the **Configure columns** section, select the relevant columns for article and summary from the given options.
3. Click **Save**.

← Back to Projects

1. Ingestion

2. Workflow

3. Configurati...

4. Review

5. Output

6. Insights

OpenAssistant Conversations / Ingestion

> 1.1 Project Details

> 1.2 Add New Datasets

✓ 1.3 Configure Datasets

View dataset (Total: 1)

oasst1 ▼ Preview

Configure Columns

Message ID Column * ⓘ

message_id ▼

Parent Column * ⓘ

parent_id ▼

Text Column * ⓘ

text ▼

Role Column * ⓘ

role ▼

Save

Version: 0.2.1

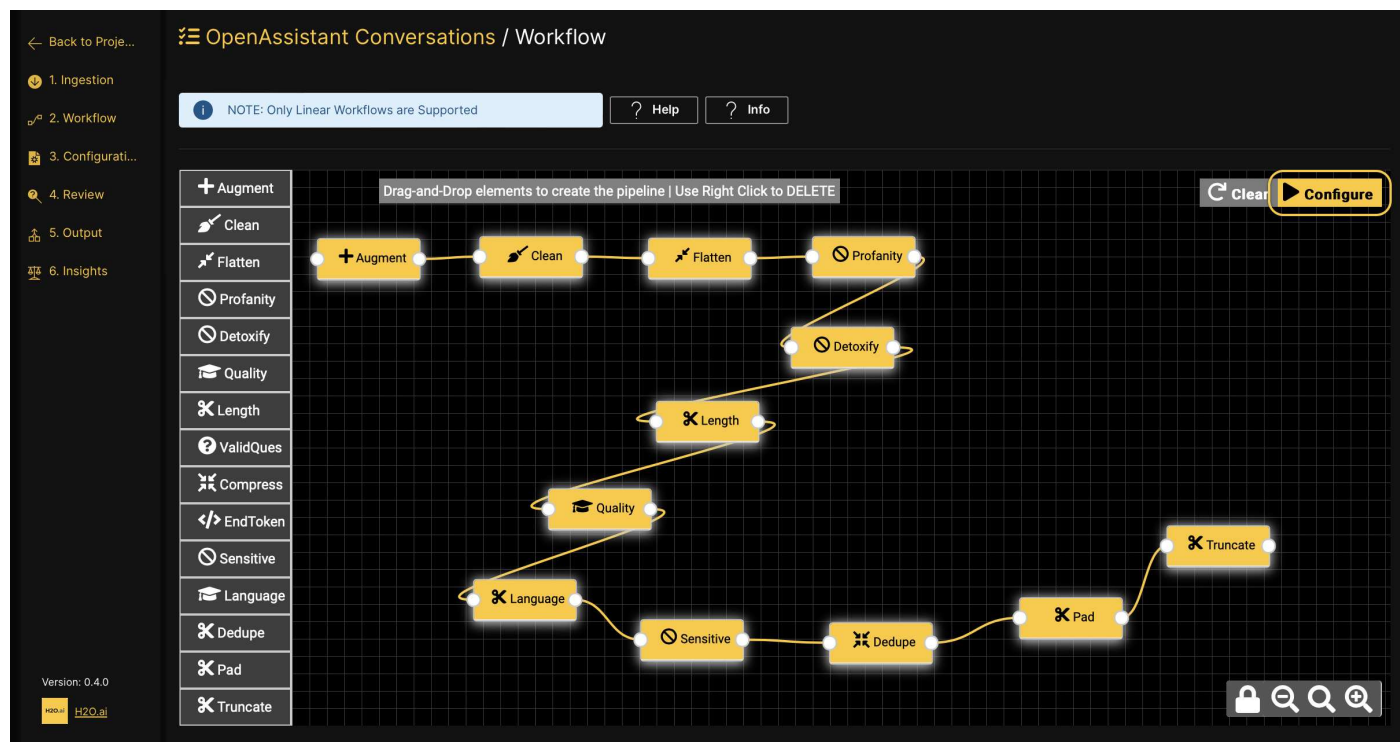
Step 3: Build the workflow

Using the workflow builder tool, let's configure the order of data preparation steps.

1. Inside the **Chat with AI** project, click **Workflow** On the left navigation menu.
2. Drag and drop the required steps in the following sequence. To learn more about the workflow builder tool, see [Workflow builder](#).
Tip: Workflow Augmentation > Text cleaning > Profanity Check > Detoxify > Length Check > Text Quality Check > Sensitive Info Check > Language Understanding > Deduplication > Padding sequence > Truncate sequence > Output

Note: If there is no GPU, detoxify function will take a long time to run.

3. After configuring the order of data preparation steps, click **Configure** to run the workflow.



Step 4: Configure the parameters

H2O LLM DataStudio allows you to customize the behavior of the function for each data preparation step by setting parameters. Let's use the default parameter configurations for this tutorial.

1. Inside the **Chat with AI** project, click **Configuration** On the left navigation menu.
2. Go through each parameter configuration. To learn more about the available configurations, see [Configuration](#).
3. Once all the parameters are configured, click **Review** to move to the next step.

Step 5: Review the configured parameters and execute the workflow

Let's review the configured parameters before executing the data preparation workflow to ensure accuracy.

1. Inside the **Chat with AI** project, click **Review** On the left navigation menu.
2. Once you are satisfied with configured parameters, click **Run pipeline** to initiate the execution of the workflow.

✔ File Uploaded: MigraineTreatment.zip

Smart Chunking

Recommended to use when document files are hundreds/thousands of pages.

Perform Smart Chunking (Fast) ☐ Off

Sampling Ratio

0

Note: If the ratio=0, best ratio is chosen automatically.

✔ Run Pipeline
✕ Discard

Step 6: Review and analyze the output

After the data preparation process is completed, a resulting dataset is generated. Let's take time to review and analyze the output.

- Inside the **Chat with AI** project, click **Output** On the left navigation menu.
 - The **Output** page graphically represents the number of rows and the percentage of rows against input rows in each data preparation step.
 - The **Output** page consists a preview of the final dataset with the top 100 rows.



- Click **Download CSV** to download the output dataset in the CSV file format.
- Click **Export to H2O Drive** to export the output dataset to H2O Drive.

Showing top 100 rows of the final dataset

Show Intermediate Features ☐ Off

prompt	response
suggest a different word for the given word without changing its meaning bring	fetch.
is paying off my credit card quicker than just paying the minimum payment each month?	yes, paying off your credit card quicker than just making the minimum payment each month can save you money in interest and reduce the amount of time it will take you to fully pay off the balance. by making larger payments than the minimum twice a month, or putting extra money on your credit card each month, you can reduce the amount of interest charged and reduce the amount of time it will take to pay off the balance.
i'm going to be passive aggressive and refuse to repeat my name to a customer	is there a reason you are wanting to be so rude to them?
create a series of 3-5 budgeting tips.	1. create a budget that outlines all of your income and expenses. 2. set up automated payments to help you stay on track. 3. utilize apps and websites that provide budgeting and savings tips. 4. cut out any unnecessary fees or subscriptions. 5. avoid impulse purchases and save money for important future expenses.

Download CSV Export to H2O Drive

Step 7: Compare input and output datasets

As the final step, we can compare and see the differences between the input and output datasets. Let's take a look at text cleaning differences, selected columns and dataset row differences.

1. Inside the **Chat with AI** project, click **Insights** On the left navigation menu.

← Back to Proje...

Chat with AI / Insights

Input Dataset	Output Dataset												
Dataset Shape: (1000, 2)	Dataset Shape: (448, 2)												
Raw Dataset Snapshot (Top 10 rows)	Output Dataset Snapshot (Top 10 rows)												
<table> <thead> <tr> <th>response</th><th>prompt</th></tr> </thead> <tbody> <tr> <td>. Determine the cause of the clog. Clogs in kitchen drains can be caused by food, grease, soap scum, and other debris. 2. If you have a plunger, try using it to dislodge the clog. Make sure to cover any drain and overflow openings with a towel to create a tight seal. Use your plunger in a steady up-and-down motion. 3. If a plunger doesn't remove the clog, you may need to use a drain auger (also known as a "snake") to remove it. You can rent one from a hardware store or purchase one online. 4. If a drain auger fails to remove the clog, you may need to disassemble the pipe and use a hand-held plumber's snake and/or rigid wire to try and break up the clog. 5. Consider calling a professional plumber if your kitchen drain continues to remain clogged after all of these steps.</td><td>What's the best way to fix my kitc</td></tr> </tbody> </table>	response	prompt	. Determine the cause of the clog. Clogs in kitchen drains can be caused by food, grease, soap scum, and other debris. 2. If you have a plunger, try using it to dislodge the clog. Make sure to cover any drain and overflow openings with a towel to create a tight seal. Use your plunger in a steady up-and-down motion. 3. If a plunger doesn't remove the clog, you may need to use a drain auger (also known as a "snake") to remove it. You can rent one from a hardware store or purchase one online. 4. If a drain auger fails to remove the clog, you may need to disassemble the pipe and use a hand-held plumber's snake and/or rigid wire to try and break up the clog. 5. Consider calling a professional plumber if your kitchen drain continues to remain clogged after all of these steps.	What's the best way to fix my kitc	<table> <thead> <tr> <th>prompt</th><th>response</th></tr> </thead> <tbody> <tr> <td>suggest a different word for the given word without changing its meaning bring</td><td>fetch.</td></tr> <tr> <td>is paying off my credit card quicker than just paying the minimum payment each month?</td><td>yes, paying off your credit card qu minimum payment each month car interest and reduce the amount of fully pay off the balance. by makin the minimum twice a month, or pu credit card each month, you can re interest charged and reduce the ai to pay off the balance.</td></tr> <tr> <td>i'm going to be passive aggressive and refuse to repeat my name to a customer</td><td>is there a reason you are wanting t</td></tr> </tbody> </table>	prompt	response	suggest a different word for the given word without changing its meaning bring	fetch.	is paying off my credit card quicker than just paying the minimum payment each month?	yes, paying off your credit card qu minimum payment each month car interest and reduce the amount of fully pay off the balance. by makin the minimum twice a month, or pu credit card each month, you can re interest charged and reduce the ai to pay off the balance.	i'm going to be passive aggressive and refuse to repeat my name to a customer	is there a reason you are wanting t
response	prompt												
. Determine the cause of the clog. Clogs in kitchen drains can be caused by food, grease, soap scum, and other debris. 2. If you have a plunger, try using it to dislodge the clog. Make sure to cover any drain and overflow openings with a towel to create a tight seal. Use your plunger in a steady up-and-down motion. 3. If a plunger doesn't remove the clog, you may need to use a drain auger (also known as a "snake") to remove it. You can rent one from a hardware store or purchase one online. 4. If a drain auger fails to remove the clog, you may need to disassemble the pipe and use a hand-held plumber's snake and/or rigid wire to try and break up the clog. 5. Consider calling a professional plumber if your kitchen drain continues to remain clogged after all of these steps.	What's the best way to fix my kitc												
prompt	response												
suggest a different word for the given word without changing its meaning bring	fetch.												
is paying off my credit card quicker than just paying the minimum payment each month?	yes, paying off your credit card qu minimum payment each month car interest and reduce the amount of fully pay off the balance. by makin the minimum twice a month, or pu credit card each month, you can re interest charged and reduce the ai to pay off the balance.												
i'm going to be passive aggressive and refuse to repeat my name to a customer	is there a reason you are wanting t												
	View Dropped Rows												

You can compare the input dataset and output dataset. Click **View dropped rows** to view the rows dropped from the input dataset.

Summary

In this tutorial, we learned how to prepare a dataset with prompts and their respective responses columns for the problem type of **Instruct Tuning**. Also, we discovered that H2O LLM DataStudio lets you to upload, prepare, and analyze your datasets, ultimately achieving your desired data transformation goals.

Tutorial 4A: Dataset preparation for Human - Bot Conversations

Overview

This tutorial describes the process of preparing a dataset comprising multiple dialogues between human users and chatbots. This task is essential for training models that can understand user intents and provide appropriate responses, leading to enhanced conversational experiences. The dataset preparation process focuses on effectively structuring and organizing conversational data, including user queries, bot responses, and relevant context.

Prerequisites

- [Access H2O LLM DataStudio via H2O AI Cloud \(HAIC\)](#).
- Review [Data preparation flow](#).
- GPU is recommended to use the [Detoxify](#) technique in data cleaning.

Step 1: Explore the project

For this tutorial, we are going to use the prebuilt **OpenAssistant Conversations** project, which has human-generated, human-annotated assistant-style conversations. Let's explore the project:

1. On the H2O LLM DataStudio navigation menu, click **Prepare**.
2. Explore the table with detailed information about both prebuilt projects and projects that you created. For more information on viewing the projects, see [View projects](#).
3. Click on the **OpenAssistant Conversations** project name and navigate to the corresponding data preparation steps.

Step 2: Ingest data

The first step of data preparation is ingesting data by uploading your datasets. In this tutorial, we are proceeding with the preloaded *oasst1* dataset. Let's configure the preloaded dataset.

1. To preview the dataset, click **Preview** under **Configure datasets** section. It will show you the top 100 rows of the dataset.
2. Under the **Configure columns** section, select the relevant columns for article and summary from the given options.
3. Click **Save**.

← Back to Projects

1. Ingestion

2. Workflow

3. Configurati...

4. Review

5. Output

6. Insights

OpenAssistant Conversations / Ingestion

> 1.1 Project Details

> 1.2 Add New Datasets

✓ 1.3 Configure Datasets

View dataset (Total: 1)

oasst1

Preview

Configure Columns

Message ID Column *

message_id

Parent Column *

parent_id

Text Column *

text

Role Column *

role

Save

Version: 0.2.1

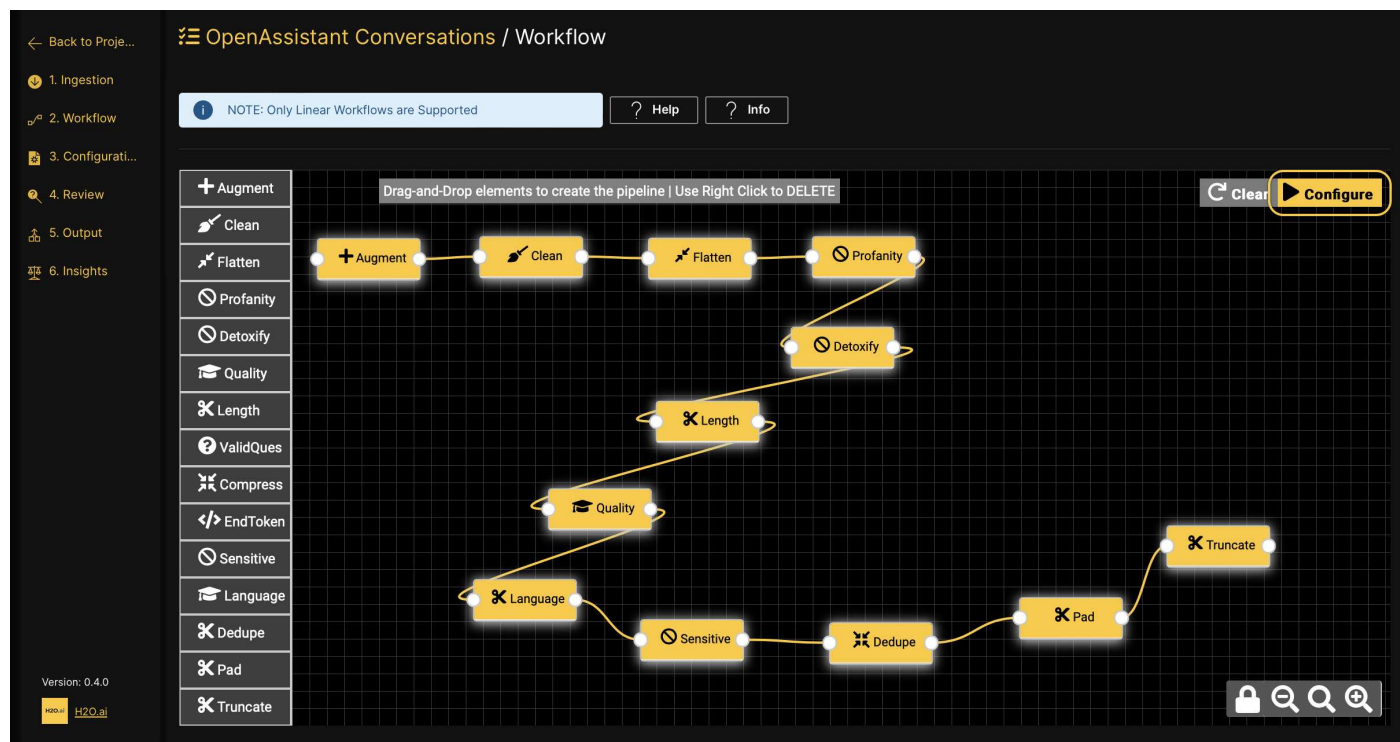
Step 3: Build the workflow

Using the workflow builder tool, let's configure the order of data preparation steps.

1. Inside the **OpenAssistant Conversations** project, click **Workflow** On the left navigation menu.
2. Drag and drop the required steps in the following sequence. To learn more about the workflow builder tool, see [Workflow builder](#).
Tip: Workflow Augmentation > Text cleaning > Flatten Conversation > Profanity Check > Detoxify > Length Check > Text Quality Check > Sensitive Info Check > Language Understanding > Deduplication > Padding sequence > Truncate sequence > Output

Note: If there is no GPU, detoxify function will take a long time to run.

3. After configuring the order of data preparation steps, click **Configure** to run the workflow.



Step 4: Configure the parameters

H2O LLM DataStudio allows you to customize the behavior of the function for each data preparation step by setting parameters. Let's use the default parameter configurations for this tutorial.

1. Inside the **OpenAssistant Conversations** project, click **Configuration** On the left navigation menu.
2. Go through each parameter configuration. To learn more about the available configurations, see [Configuration](#).
3. Once all the parameters are configured, click **Review** to move to the next step.

Step 5: Review the configured parameters and execute the workflow

Let's review the configured parameters before executing the data preparation workflow to ensure accuracy.

1. Inside the **OpenAssistant Conversations** project, click **Review** On the left navigation menu.
2. Once you are satisfied with configured parameters, click **Run pipeline** to initiate the execution of the workflow.

✔ File Uploaded: MigraineTreatment.zip

Smart Chunking

Recommended to use when document files are hundreds/thousands of pages.

Perform Smart Chunking (Fast) ☐ Off

Sampling Ratio

0

Note: If the ratio=0, best ratio is chosen automatically.

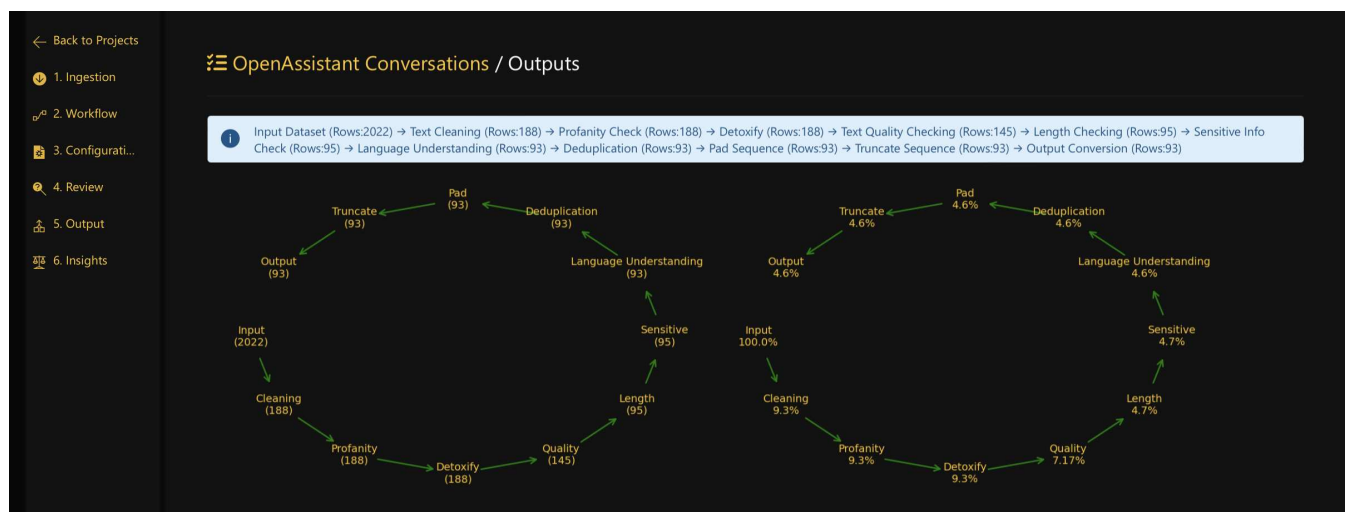
✔ Run Pipeline

✕ Discard

Step 6: Review and analyze the output

After the data preparation process is completed, a resulting dataset is generated. Let's take time to review and analyze the output.

- Inside the **OpenAssistant Conversations** project, click **Output** On the left navigation menu.
 - The **Output** page graphically represents the number of rows and the percentage of rows against input rows in each data preparation step.
 - The **Output** page consists a preview of the final dataset with the top 100 rows.



- Click **Download CSV** to download the output dataset in the CSV file format.
- Click **Export to H2O Drive** to export the output dataset to H2O Drive.

Showing top 100 rows of the final dataset

Show Intermediate Features ☐ Off

text

<human>: can you write haiku for different inductive biases in deep learning models? start with attention <bot>: sure, here are haikus about various inductive biases in the field of deep learning models. attention bright patterns caught in sight, explicitly shines more insight. inductive bias lights. nearest neighbour data points nearby search, faster and easier answers. neighbours lend a hand. maximum conditional independence predictive models sparse, exploring independence links. fully decompose thought. minimum cross-validation error data is observed machine learning fit for sure. minimise errors. <human>: the haiku's you've provided seem improperly constructed, they each have 6-8-6 syllables instead of 5-7-5. they are very creative, however! <bot>: sorry about that. as i have only been trained on text and not sound, i only have a weak understanding of syllables. i can probably pull off a decent acrostic, though!

<human>: would you be able to design an electric pokmon? describe it as best as you can and also give me the stats of that pokmon. <bot>: name: spritesla category: electrostatic pokmon type: electric/ghost spritesla is a small, round, electric-lime ghost with two antennae shaped like tesla coils. the bottom of its body forms a thunderbolt shape as it fades out. it emits a slight electrostatic charge at all times. as signified by the persistent grin on its face, tends to be mischievous and pull pranks, such as infamously scaring people by subtly warping the image on a tv. stats: - hp: 60 - attack: 40 - defense: 20 - special attack: 90 - special defense: 60 - speed: 120 <human>: thanks, that's a really cool idea! is the first evolution form? does it evolve? and in a game such as pokmon red, how strong is this relative to other choices for my roster? <bot>: spritesla evolves to "spritefact!" a large box shaped green ghostly mass which at first glance looks like an old television, it has two long antennae from its top and from its bottom a green trail is sometimes visible which fades out as it gets further of the body. it is considered a fairly weak compared with most ghost and electric type pokmon, but has alot of charm and a mischief personality to compensate.

<human>: given two matrices a and b which you can multiply together, does $(a \cdot b)^t = a^t \cdot b^t$ or $b^t \cdot a^t$? <bot>: when you multiply two matrices a and b and then take the transpose

Download CSV Export to H2O Drive

Step 7: Compare input and output datasets

As the final step, we can compare and see the differences between the input and output datasets. Let's take a look at text cleaning differences, selected columns and dataset row differences.

1. Inside the **OpenAssistant Conversations** project, click **Insights** On the left navigation menu.

← Back to Proje... **OpenAssistant Conversations / Insights**

1. Ingestion 2. Workflow 3. Configurati... 4. Review 5. Output 6. Insights

Input Dataset

Dataset Shape: (2022, 7)

Raw Dataset Snapshot (Top 10 rows)

message_id	parent_id
e0cf4fa2-e944-44fb-b502-11f155062092	nan
7d05acb7-9360-458c-8a1d-c0b6492b8f8a	e0cf4fa2-e944-44fb-b502-11f155

Output Dataset

Dataset Shape: (43, 1)

Output Dataset Snapshot (Top 10 rows)

text

<human>: can you write haiku for different inductive biases in deep learning models? start with attention <bot>: sure, here are haikus about various inductive biases in the field of deep learning models. attention bright patterns caught in sight, explicitly shines more insight. inductive bias lights. nearest neighbour data points nearby search, faster and easier answers. neighbours lend a hand. maximum conditional independence predictive models sparse, exploring independence links. fully decompose thought. minimum cross-validation error data is observed machine learning fit for sure. minimise errors. <human>: the haiku's you've provided seem improperly constructed, they each have 6-8-6 syllables instead of 5-7-5. they are very creative, however! <bot>: sorry about that. as i have only been trained on text and not sound, i only have a weak understanding of syllables. i can probably pull off a decent acrostic, though!

View Dropped Rows

You can compare the input dataset and output dataset. Click **View dropped rows** to view the rows dropped from the input dataset.

Summary

In this tutorial, we learned how to prepare a dataset with prompts and their respective responses columns for the problem type of **Instruct Tuning**. Also, we discovered that H2O LLM DataStudio lets you to upload, prepare, and analyze your datasets, ultimately achieving your desired data transformation goals.

Tutorial 5A: Dataset preparation for Continued PreTraining

Overview

This tutorial describes the process of preparing datasets with extensive texts for further pretraining of language models. The dataset preparation process focuses on organizing long text data, enabling the language models to learn from a wide range of linguistic patterns. This contributes to improved language understanding and generation capabilities.

Prerequisites

- [Access H2O LLM DataStudio via H2O AI Cloud \(HAIC\)](#).
- Review [Data preparation flow](#).
- GPU is recommended to use the [Detoxify](#) technique in data cleaning.

Step 1: Explore the project

For this tutorial, we are going to use the prebuilt **Wikipedia 512 Pretraining** project, which is a Wikipedia preprocessed for pretraining of models. Each sample in the dataset has an average tokenized length of 512 RoBERTa-Base tokens. Let's explore the project:

1. On the H2O LLM DataStudio navigation menu, click **Prepare**.
2. Explore the table with detailed information about both prebuilt projects and projects that you created. For more information on viewing the projects, see [View projects](#).
3. Click on the **Wikipedia 512 Pretraining** project name and navigate to the corresponding data preparation steps.

Step 2: Ingest data

The first step of data preparation is ingesting data by uploading your datasets. In this tutorial, we are proceeding with the preloaded *wikipedia_512_pretraining* dataset. Let's configure the preloaded dataset.

1. To preview the dataset, click **Preview** under **Configure datasets** section. It will show you the top 100 rows of the dataset.
2. Under the **Configure columns** section, select the relevant columns for article and summary from the given options.
3. Click **Save**.

← Back to Projects

1. Ingestion

2. Workflow

3. Configurati...

4. Review

5. Output

6. Insights

OpenAssistant Conversations / Ingestion

> 1.1 Project Details

> 1.2 Add New Datasets

✓ 1.3 Configure Datasets

View dataset (Total: 1)

oasst1 ▼ Preview

Configure Columns

Message ID Column * ⓘ

message_id ▼

Parent Column * ⓘ

parent_id ▼

Text Column * ⓘ

text ▼

Role Column * ⓘ

role ▼

Save

Version: 0.2.1

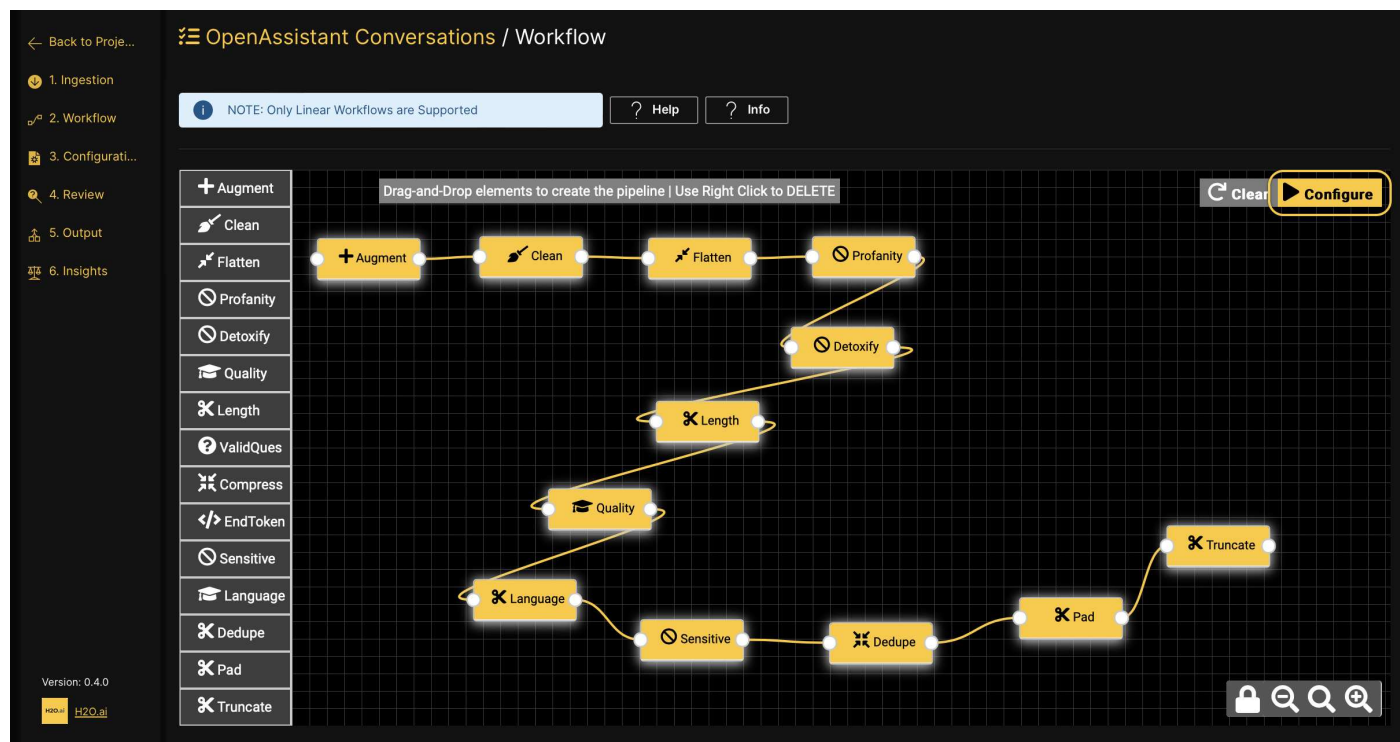
Step 3: Build the workflow

Using the workflow builder tool, let's configure the order of data preparation steps.

1. Inside the **Wikipedia 512 Pretraining** project, click **Workflow** On the left navigation menu.
2. Drag and drop the required steps in the following sequence. To learn more about the workflow builder tool, see [Workflow builder](#).
Tip: Workflow Augmentation > Text cleaning > Profanity Check > Detoxify > Length Check > Text Quality Check > Sensitive Info Check > Language Understanding > Deduplication > Padding sequence > Truncate sequence > Output

Note: If there is no GPU, detoxify function will take a long time to run.

3. After configuring the order of data preparation steps, click **Run** to run the workflow.



Step 4: Configure the parameters

H2O LLM DataStudio allows you to customize the behavior of the function for each data preparation step by setting parameters. Let's use the default parameter configurations for this tutorial.

1. Inside the **Wikipedia 512 Pretraining** project, click **Configuration** On the left navigation menu.
2. Go through each parameter configuration. To learn more about the available configurations, see [Configuration](#).
3. Once all the parameters are configured, click **Review** to move to the next step.

Step 5: Review the configured parameters and execute the workflow

Let's review the configured parameters before executing the data preparation workflow to ensure accuracy.

1. Inside the **Wikipedia 512 Pretraining** project, click **Review** On the left navigation menu.
2. Once you are satisfied with configured parameters, click **Run pipeline** to initiate the execution of the workflow.

✔ File Uploaded: MigraineTreatment.zip

Smart Chunking

Recommended to use when document files are hundreds/thousands of pages.

Perform Smart Chunking (Fast) ☐ Off

Sampling Ratio

0

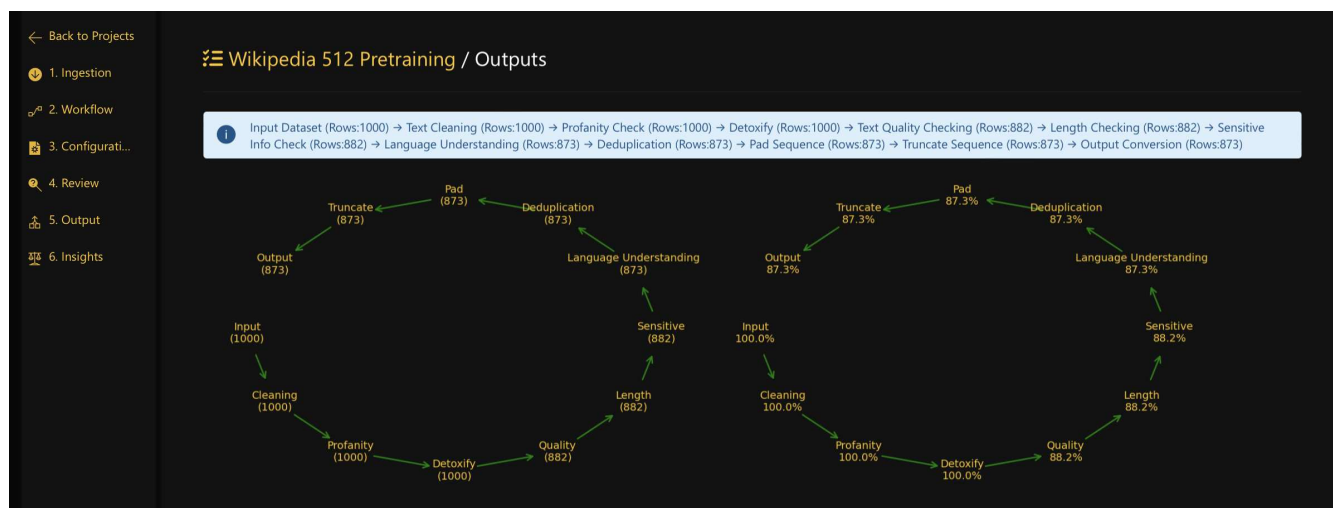
Note: If the ratio=0, best ratio is chosen automatically.

✔ Run Pipeline
✕ Discard

Step 6: Review and analyze the output

After the data preparation process is completed, a resulting dataset is generated. Let's take time to review and analyze the output.

- Inside the **Wikipedia 512 Pretraining** project, click **Output** On the left navigation menu.
 - The **Output** page graphically represents the number of rows and the percentage of rows against input rows in each data preparation step.
 - The **Output** page consists a preview of the final dataset with the top 100 rows.



- Click **Download CSV** to download the output dataset in the CSV file format.
- Click **Export to H2O Drive** to export the output dataset to H2O Drive.

Showing top 100 rows of the final dataset

Show Intermediate Features ☐ Off

text

```
<human>: i am learning mandarin. could you give me some advice to memorize how to write chinese characters and speed up my learning? <bot>: learning how to write chinese characters can be a challenging but rewarding process. here are some tips to help you memorize chinese characters and speed up your learning: 1. understand the basic strokes: chinese characters are made up of a combination of basic strokes. learning and understanding these strokes can help you to recognize and write characters more easily. some common strokes include horizontal, vertical, diagonal, and curved lines. 2. use flashcards: flashcards can be an effective tool for memorizing chinese characters. you can create your own flashcards with a character on one side and its meaning and pronunciation on the other. review the flashcards regularly to help you remember the characters. 3. practice writing characters: writing chinese characters by hand can help you to remember them better. you can practice by writing characters repeatedly, or by copying passages of chinese text. you can also use apps or websites that provide writing practice exercises. 4. focus on radicals: chinese characters are made up of components called radicals. learning the most common radicals and their meanings can help you to recognize and remember characters more easily. you can find lists of radicals online or in a chinese character dictionary. 5. read chinese text: reading chinese text can help you to learn characters in context. you can start with simple texts and gradually work your way up to more complex ones. reading aloud can also help you to remember the pronunciation of characters. 6. watch chinese movies or tv shows: watching chinese movies or tv shows can help you to improve your listening skills and expose you to new vocabulary and grammar. you can use subtitles to help you understand the dialogue. 7. get a tutor or join a language exchange: working with a tutor or language exchange partner can give you personalized guidance and feedback. they can help you identify areas where you need to improve and give you tips and strategies for memorizing characters. remember that learning chinese characters takes time and practice. be patient with yourself and celebrate your progress. we wish you good luck! <human>: can you make me a list of the most important chinese characters to learn?
```

<human>: in typescript, how can i tell the type system "i know that in theory this value may have these multiple types, but in practice i know it will only have this one" <bot>: *Answered: 10/10/2023, 10:10 AM*

Download CSV Export to H2O Drive

Step 7: Compare input and output datasets

As the final step, we can compare and see the differences between the input and output datasets. Let's take a look at text cleaning differences, selected columns and dataset row differences.

1. Inside the **Wikipedia 512 Pretraining** project, click **Insights** On the left navigation menu.

← Back to Proj... **OpenAssistant Conversations / Insights**

1. Ingestion 2. Workflow 3. Configurati... 4. Review 5. Output 6. Insights

Input Dataset

Dataset Shape: (2022, 7)

Raw Dataset Snapshot (Top 10 rows)

message_id	parent_id
e0cf4fa2-e944-44fb-b502-11f155062092	nan
7d05acb7-9360-458c-8a1d-c0b6492b8f8a	e0cf4fa2-e944-44fb-b502-11f155

Output Dataset

Dataset Shape: (95, 1)

Output Dataset Snapshot (Top 10 rows)

text

```
<human>: i am learning mandarin. could you give me some advice to memorize how to write chinese characters and speed up my learning? <bot>: learning how to write chinese characters can be a challenging but rewarding process. here are some tips to help you memorize chinese characters and speed up your learning: 1. understand the basic strokes: chinese characters are made up of a combination of basic strokes. learning and understanding these strokes can help you to recognize and write characters more easily. some common strokes include horizontal, vertical, diagonal, and curved lines. 2. use flashcards: flashcards can be an effective tool for memorizing chinese characters. you can create your own flashcards with a character on one side and its meaning and pronunciation on the other. review the flashcards regularly to help you remember the characters. 3. practice writing characters: writing chinese characters by hand can help you to remember them better. you can
```

View Dropped Rows

You can compare the input dataset and output dataset. Click **View dropped rows** to view the rows dropped from the input dataset.

Summary

In this tutorial, we learned how to prepare a dataset with extensive texts for the problem type of **Continued PreTraining** (further pretraining of language models). Also, we discovered that H2O LLM DataStudio lets you to upload, prepare, and analyze your datasets, ultimately achieving your desired data transformation goals.

Tutorial 1B: Convert documents into Q&A pairs for data preparation

Overview

In this tutorial, you'll learn how to convert documents into question-answer pairs using H2O LLM DataStudio. These pairs can be used for data preparation tasks, such as transforming and cleaning datasets.

Prerequisites

Before you start, ensure you have the following:

- [Access to H2O LLM DataStudio via H2O AI Managed Cloud \(HAMC\)](#).
- Familiarity with the [data curation flow](#).

Step 1: Create a new project

To begin the process of data curation, let's follow these steps to create a new project:

1. On the H2O LLM DataStudio left navigation menu, click **Curate**.
2. On the **All Projects / Curate Data for LLMs** page, click **New**. **Note:** If this is your first time creating a new project, you must integrate H2OGPTe by providing the required credentials. You cannot create a new project without configuring H2OGPTe. For more information, see [Settings](#).
3. In this tutorial, we will upload a research paper on chronic migraine diagnosis and treatment to generate question-answer pairs. In the **Project name** text box, let's enter **migraine-treatment-curate** as the name for the new project.
4. On the **Description** text box, let's enter **chronic migraine diagnosis and treatment** as the description for the project.
5. On the **Document description** text box, let's provide a brief description for the file's content. This label helps to quickly categorize and identify the document's purpose. Since we are uploading a research paper, type **Research Paper**.
6. From the **Task type** dropdown menu, select **Question-answer** task.
7. Download the [research paper on chronic migraine diagnosis and treatment](#) to your computer. Once the file is downloaded, follow these steps:
 1. Click **Browse** to open the file selection dialog.
 2. Locate and select the downloaded file from your computer.
 3. Alternatively, you can drag and drop the file into the designated area.
8. Click **Upload** to upload the document.

Step 2: Configure settings

After uploading your document, configure the following settings. For this tutorial, let's keep the default settings as specified in each step:

1. In the **LLM selection** section, select your preferred H2OGPTE LLM from the available options. For this tutorial, we will keep the default LLM selection.
2. Choose your preferred relevance score from the dropdown menu. For this tutorial, use the default relevance score, which is the Bert approach.
3. Use the slider labeled **Number of tokens per chunk** to set the maximum number of tokens per chunk of text processed by the model. The default value is 1,000 tokens. Keep this default value.
4. Keep the **More customization settings** at their default values.
5. Enable the **Perform Smart Chunking (Fast)** option if you are processing large documents (hundreds or thousands of pages). This feature speeds up chunking but may limit the generation of sufficient records for fine-tuning. The default setting is disabled, so we will keep it as is.
6. Adjust the [Sampling ratio](#) for smart chunking. By default, the sampling ratio is set to 0, meaning LLM DataStudio will automatically select the sampling ratio based on the document length. Keep this default setting.
7. Toggle the **Use h2oGPTE's ingestion pipeline** option. The default is enabled, so we will keep it as is. This option allows you to choose between using h2oGPTE's ingestion pipeline or the default LLM DataStudio pipeline.

Step 3: Run pipeline

Now that you've configured all the necessary settings, it's time to execute the pipeline and begin the data processing.

1. Click **Run pipeline** to start the process.

Step 4: View the project

To view and interact with your project:

1. In the H2O LLM DataStudio left navigation menu, click **Curate**.
2. Select the newly created project by clicking on its name.

You can view the table of question-answer pairs along with other details, such as the status of each project, project details, the number of pairs, and more. For a complete list of what you can view, see [View a specific Curate project](#).

migraine-treatment-curate

Status: **Complete** Pairs: **118** Sources: **Label** **Input** **Output** **JSON** **Execute**

Search

#	Prompt	Answer	Relevan...	Filename	Flag
0	In which volume and issue was this paper published?	This paper was published in volume 6, issue 3 of the journal.	0.4289	The_diagnosis_and_trea	
1	When was this paper published?	This paper was published in May 2015.	0.33376	The_diagnosis_and_trea	
2	How many citations does this paper have?	As of now, this paper has 142 citations.	0.4914	The_diagnosis_and_trea	
3	What is the title of the research paper?	The title of the research paper is "The diagnosis and treatment of chronic migraine".	0.39904	The_diagnosis_and_trea	
4	What are the page numbers of this paper?	The paper is published from pages 115 to 123.	0.4252	The_diagnosis_and_trea	
5	What is the DOI of this paper?	The DOI of this paper is 10.1177/2040622315579627.	0.40457	The_diagnosis_and_trea	
6	How many reads does this paper have?	As of now, this paper has 1,529 reads.	0.4344	The_diagnosis_and_trea	
7	Who is the author of this paper?	The author of this paper is Mark Weatherall.	0.43019	The_diagnosis_and_trea	
8	Which institution is the author affiliated with?	The author is affiliated with Buckinghamshire Healthcare NHS Trust.	0.44612	The_diagnosis_and_trea	

118 of 118

Step 5: Publish the dataset as preparation project

To use the generated question-answer pairs in the data preparation process:

1. In the **Output** section, select **Publish as Preparation Project** from the dropdown menu.
2. Click **Execute**.
3. In the **Publish as Prepare Project** window, let's keep the **Project name** and **Project description** as they are, and select question answering as the task type.
4. Click **Publish** to use the generated question-answer pairs in the [Data preparation flow](#).

migraine-treatment-curate

Status Complete Pairs 118 Sources Label Input Output JSON Execute

Search

#	Prompt	Answer	Relevan...	Filename	Flag
0	In which volume and issue was this paper published?	This paper was published in volume 6, issue 3 of the journal.	0.4289	The_diagnosis_and_trea	
1	When was this paper published?	This paper was published in May 2015.	0.33376	The_diagnosis_and_trea	
2	How many citations does this paper have?	As of now, this paper has 142 citations.	0.4914	The_diagnosis_and_trea	
3	What is the title of the research paper?	The title of the research paper is "The diagnosis and treatment of chronic migraine".	0.39904	The_diagnosis_and_trea	
4	What are the page numbers of this paper?	The paper is published from pages 115 to 123.	0.4252	The_diagnosis_and_trea	
5	What is the DOI of this paper?	The DOI of this paper is 10.1177/2040622315579627.	0.40457	The_diagnosis_and_trea	
6	How many reads does this paper have?	As of now, this paper has 1,529 reads.	0.4344	The_diagnosis_and_trea	
7	Who is the author of this paper?	The author of this paper is Mark Weatherall.	0.43019	The_diagnosis_and_trea	
8	Which institution is the author affiliated with?	The author is affiliated with Buckinghamshire Healthcare NHS Trust.	0.44612	The_diagnosis_and_trea	

Summary

In this tutorial, we learned how to convert a document into question-answer pairs using H2O LLM DataStudio. We walked through the process of creating a new project, configuring the necessary settings, and running the pipeline. Finally, we explored how to publish the curated dataset as a preparation project to continue the data preparation process, enabling you to achieve your data transformation goals efficiently.

Tutorial 1C: Create question-type and robust-type evaluation datasets

Overview

In this tutorial, you'll learn how to create a question-type evaluation dataset using H2O LLM DataStudio and then generate a robust evaluation dataset from it. By following the steps, you'll be able to configure and customize a dataset that tests various [question types](#), including simple, conditional, compressed, and multihop reasoning questions. Additionally, you'll transform this dataset into a [robust evaluation type](#), ensuring it can effectively challenge and validate your model's performance under diverse and complex conditions.

Prerequisites

Before starting this tutorial, make sure you have the following:

1. [Access to H2O LLM DataStudio via H2O AI Managed Cloud \(HAMC\)](#).
2. A pre-existing question-answer dataset. This tutorial makes use of a question-answer dataset based on Tweets. Download the [Tweet QA CSV file](#) to your machine.
3. Familiarity with the [data curation flow](#).
4. Credentials for H2OGPTe integration.

Step 1: Create a new evaluation dataset

To begin the process of [data curation](#), let's follow these steps to create a new evaluation dataset:

1. On the H2O LLM DataStudio left navigation menu, click **Custom Eval**.
2. On the **Create Your Own Eval Datasets (Beta)** page, click **New**. **Note:** If this is your first time creating a new evaluation dataset, you must integrate h2oGPTe by providing the required credentials. For more information, see [Settings](#).
3. In the **Project name** text box, enter `tweets-eval-dataset`.
4. In the **Description** text box, enter `evaluation dataset on tweets qa`.
5. From the **Dataset type** drop-down menu, select **Question type** as the evaluation dataset type.
6. In the **Do you already have a QA Dataset?** dropdown menu, select **Yes**.
7. Toggle the **Use h2oGPTe's ingestion pipeline** option. The default is enabled, so we will keep it as is. This option allows you to choose between using h2oGPTe's ingestion pipeline or the default LLM DataStudio pipeline.
8. Click **Next**.

Step 2: Upload question answering (QA) dataset

Once you've configured the basic project settings, the next step is to upload your question answering (QA) dataset. Follow these steps:

1. Click **Browse** to open the file selection dialog, and then locate and select the downloaded `tweet_qa.csv` file from your machine. Alternatively, you can drag and drop the file into the designated area.
2. Click **Upload**.

Step 3: Configure settings

After uploading your document, configure the following settings. For this tutorial, let's keep the default settings as specified in each step:

1. In the **LLM selection** section:
 1. Select your preferred H2OGPTe LLM.
 2. For this tutorial, we'll use the default option.
2. In the **Configure columns** section:
 1. Select the `context` column of the dataset for context or background information relevant to the questions.
 2. Select the `question` column of the dataset that contains the questions.
 3. Select the `answer` column of the dataset that contains the answers.

3. In the **Question type dataset** section, configure the distribution of each question type (simple, conditional, compressed, multihop). For this tutorial, we will use the default settings.

Step 4: Run pipeline

Now that you've configured all the necessary settings, it's time to execute the pipeline and begin the eval dataset creation.

1. Click **Run pipeline**.

Step 5: View the project

To view and interact with your new project:

1. In the H2O LLM DataStudio left navigation menu, click **Custom Eval**.
2. Select your project by clicking its name.

You will see a table of question-answer pairs, along with details such as question types, status, and the number of pairs. For a complete list of what you can view, see [View an evaluation dataset](#).

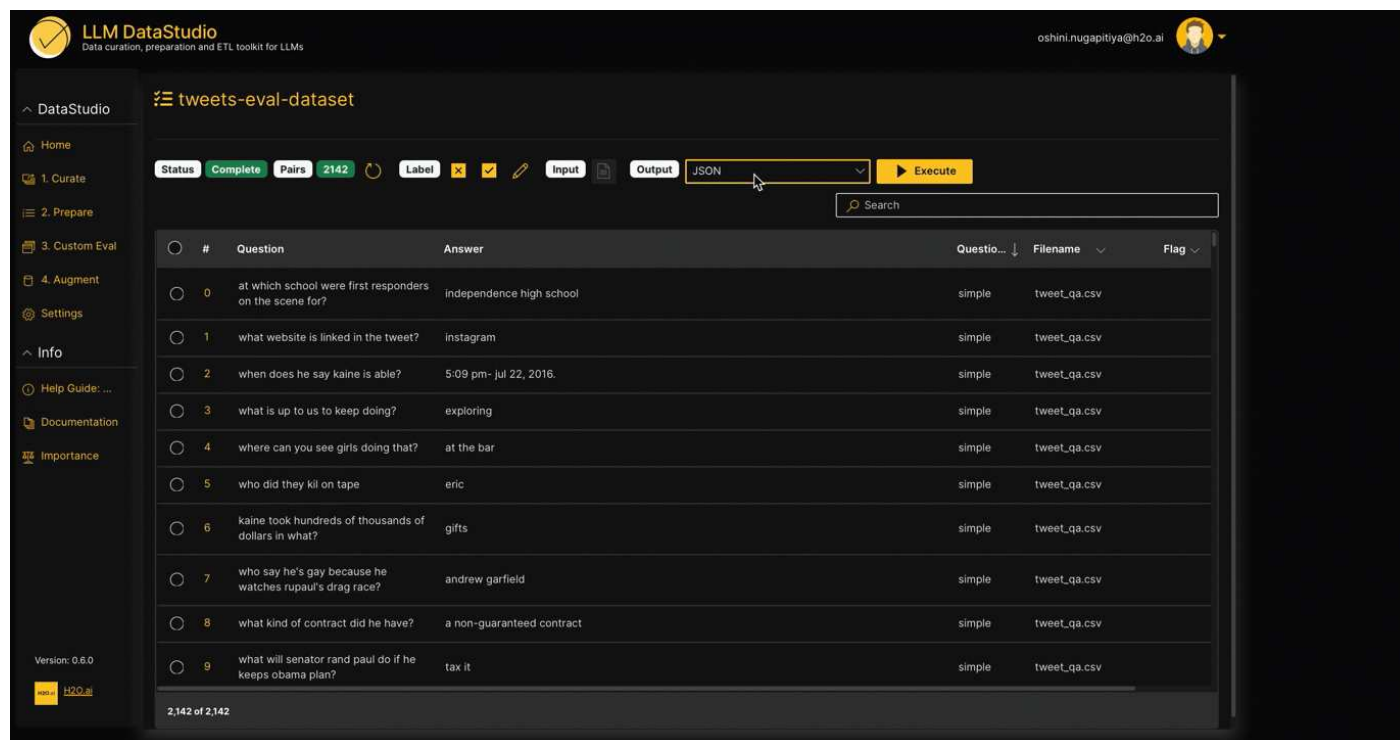
#	Question	Answer	Question type	Filename	Flag
0	at which school were first responders on the scene for?	independence high school	simple	tweet_qa.csv	
1	what website is linked in the tweet?	instagram	simple	tweet_qa.csv	
2	when does he say kaine is able?	5:09 pm- jul 22, 2016.	simple	tweet_qa.csv	
3	what is up to us to keep doing?	exploring	simple	tweet_qa.csv	
4	where can you see girls doing that?	at the bar	simple	tweet_qa.csv	
5	who did they kil on tape	eric	simple	tweet_qa.csv	
6	kaine took hundreds of thousands of dollars in what?	gifts	simple	tweet_qa.csv	
7	who say he's gay because he watches rupaul's drag race?	andrew garfield	simple	tweet_qa.csv	
8	what kind of contract did he have?	a non-guaranteed contract	simple	tweet_qa.csv	
9	what will senator rand paul do if he keeps obama plan?	tax it	simple	tweet_qa.csv	

2,142 of 2,142

Step 6: Generate a robust evaluation dataset

To generate a robust evaluation dataset from your newly created evaluation dataset:

1. In the **Output** section, select **Generate robust eval dataset** from the dropdown menu.
2. Choose one or more entries from the evaluation dataset.
3. Click **Execute**.



LLM DataStudio
Data curation, preparation and ETL toolkit for LLMs

oshini.nugapitiya@h2o.ai

tweets-eval-dataset

Status: **Complete** Pairs: **2142** Label: [X] [✓] [✎] Input: [📄] Output: **JSON** [▶ Execute]

Search

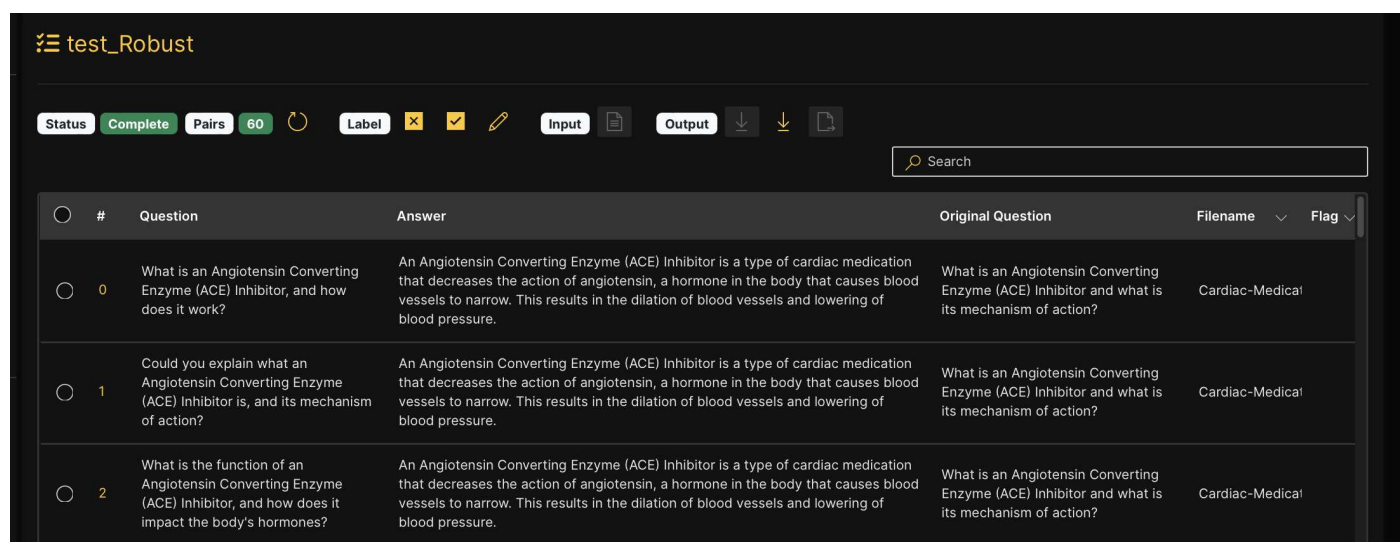
#	Question	Answer	Question...	Filename	Flag
0	at which school were first responders on the scene for?	independence high school	simple	tweet_qa.csv	
1	what website is linked in the tweet?	instagram	simple	tweet_qa.csv	
2	when does he say kaine is able?	5:09 pm- jul 22, 2016.	simple	tweet_qa.csv	
3	what is up to us to keep doing?	exploring	simple	tweet_qa.csv	
4	where can you see girls doing that?	at the bar	simple	tweet_qa.csv	
5	who did they kil on tape	eric	simple	tweet_qa.csv	
6	kaine took hundreds of thousands of dollars in what?	gifts	simple	tweet_qa.csv	
7	who say he's gay because he watches rupaul's drag race?	andrew garfield	simple	tweet_qa.csv	
8	what kind of contract did he have?	a non-guaranteed contract	simple	tweet_qa.csv	
9	what will senator rand paul do if he keeps obama plan?	tax it	simple	tweet_qa.csv	

2,142 of 2,142

Version: 0.6.0

H2O.ai

The robust evaluation dataset contains the new questions generated based on the original question, the answer, and the original question.



test_Robust

Status: **Complete** Pairs: **60** Label: [X] [✓] [✎] Input: [📄] Output: [📄] [📄] [📄]

Search

#	Question	Answer	Original Question	Filename	Flag
0	What is an Angiotensin Converting Enzyme (ACE) Inhibitor, and how does it work?	An Angiotensin Converting Enzyme (ACE) Inhibitor is a type of cardiac medication that decreases the action of angiotensin, a hormone in the body that causes blood vessels to narrow. This results in the dilation of blood vessels and lowering of blood pressure.	What is an Angiotensin Converting Enzyme (ACE) Inhibitor and what is its mechanism of action?	Cardiac-Medical	
1	Could you explain what an Angiotensin Converting Enzyme (ACE) Inhibitor is, and its mechanism of action?	An Angiotensin Converting Enzyme (ACE) Inhibitor is a type of cardiac medication that decreases the action of angiotensin, a hormone in the body that causes blood vessels to narrow. This results in the dilation of blood vessels and lowering of blood pressure.	What is an Angiotensin Converting Enzyme (ACE) Inhibitor and what is its mechanism of action?	Cardiac-Medical	
2	What is the function of an Angiotensin Converting Enzyme (ACE) Inhibitor, and how does it impact the body's hormones?	An Angiotensin Converting Enzyme (ACE) Inhibitor is a type of cardiac medication that decreases the action of angiotensin, a hormone in the body that causes blood vessels to narrow. This results in the dilation of blood vessels and lowering of blood pressure.	What is an Angiotensin Converting Enzyme (ACE) Inhibitor and what is its mechanism of action?	Cardiac-Medical	

Summary

In this tutorial, we learned how to create a question-type evaluation dataset using H2O LLM DataStudio and generate a robust evaluation dataset. We walked through the process of creating a new evaluation dataset, uploading a question-answer dataset, and configuring the necessary settings for different question types. Finally, we explored how to transform the dataset into a robust evaluation type to effectively test and validate your model's performance across a range of complex question types, ensuring a more comprehensive evaluation process.

Supported problem types

Overview

H2O LLM DataStudio offers support for various problem types and workflows, providing users with the necessary tools to prepare datasets and train models for specific tasks. This page serves as a comprehensive guide to the supported problem types, highlights their importance, and explains how the application can assist in dataset preparation and model training.

Question and Answer

- **Description:** H2O LLM DataStudio simplifies dataset preparation for question answering models. The datasets consist of contextual information, questions, and their respective answers. Its features facilitate the creation of well-structured datasets essential for training models to accurately respond to user queries based on the provided context.
- **Expected Columns:** ‘Question’, ‘Answer’, and ‘Context’.
- **Example:**

question	answer	context
What are the cookies used for?	Cookies: In order to offer and provide a customized, personal service, PhotoPost.com uses cookies to store and help track your information as you travel throughout the site. For example, we may use cookies to help remind us who you are and to deliver content and services based upon your account information. In addition, third party advertising networks may issue cookies when serving advertisements.	All Enthusiast, Inc.’s Privacy Policy All Enthusiast, Inc.’s PhotoPost.com respects the privacy and security of its users. Our goal is to provide you with a personalized Internet experience that delivers the information, resources, and services that are most relevant and helpful to you. In order to achieve this goal, we sometimes collect information during your visits to understand what differentiates you from each of our millions of other users....We welcome any questions or comments you have about PhotoPost.com: please direct them to our contact form.

Text Summarization

- **Description:** The Text Summarization workflow is designed for datasets consisting of articles and their corresponding summaries. Using H2O LLM DataStudio tools, this workflow simplifies the process of extracting vital information from articles, allowing you to create concise summaries that capture the main points. The resulting datasets are valuable for training text summarization models that can produce concise and informative summaries from lengthy text.
- **Expected Columns:** ‘Article’ and ‘Summary’.
- **Example:**

article	summary
Sally Forrest, an actress-dancer who graced the silver screen throughout the '40s and '50s in MGM musicals and films such as the 1956 noir <i>While the City Sleeps</i> died on March 15 at her home in Beverly Hills, California. Forrest, whose birth name was Katherine Feeney, was 86 and had long battled cancer. Her publicist, Judith Goffin, announced the news Thursday....Forrest married writer-producer Milo Frank in 1951. He died in 2004. She is survived by her niece, Sharon Durham, and nephews, Michael and Mark Feeney. Career: A San Diego native, Forrest became a protege of Hollywood trailblazer Ida Lupino, who cast her in starring roles in films	“Sally Forrest, an actress-dancer who graced the silver screen throughout the '40s and '50s in MGM musicals and films died on March 15 . Forrest, whose birth name was Katherine Feeney, had long battled cancer . A San Diego native, Forrest became a protege of Hollywood trailblazer Ida Lupino, who cast her in starring roles in films .”

Instruct Tuning

- **Description:** H2O LLM DataStudio assists in preparing datasets that include prompts or instructions along with their corresponding responses. These datasets are essential for training models to understand and follow provided instructions, enabling accurate responses to user prompts.
- **Expected Columns:** ‘Prompt’ and ‘Response’.
- **Example:**

prompt	response
Translate the phrase “Good Morning” to French	Bonjour

Human - Bot Conversations

- **Description:** This workflow deals with datasets containing dialogues between human users and chatbots. These datasets are crucial for training models to comprehend user intents and deliver appropriate responses, thereby improving conversational experiences. H2O LLM DataStudio aids in efficiently structuring and organizing the conversational data, including user queries, and bot responses.
- **Expected Columns:** ‘Message_id’, ‘Parent_id’, ‘Text’, and ‘Role’.
- **Example:**

message_id	parent_id	text	role
384ad8e0-8fc2-4dfd-bf48-0c417f6c5f0f	7d05acb7-9360-458c-8a1d-c0b6492b8f8a	“What are your thoughts on the censorship of ChatGPT’s output and its liberal biases?”	prompter

Continued PreTraining

- **Description:** In this workflow, H2O LLM DataStudio helps prepare datasets containing extensive texts for further pretraining of language models. The dataset preparation process focuses on organizing long text data, allowing language models to learn from a diverse range of linguistic patterns. This enhances their language understanding and generation capabilities.
- **Expected Column:** ‘Text’.
- **Example:**

text

Chrysaethe amoena Chrysaethe amoena is a species of beetle in the family Cerambycidae. It was described by Gounelle in 1911.

Video guide

Watch [this](#) video guide to learn more about how H2O LLM DataStudio streamlines data preparation for diverse LLM tasks by supporting various workflows.

Supported functionalities

H2O LLM DataStudio supports a multitude of functions to facilitate the preparation of datasets for various task types. The primary goal is to structure data optimally for maximal model performance. Following is an overview of the key functions available:

- **Text Cleaning:** Offers a range of cleaning methods to clean text data for all task types.
 - Removes unwanted characters (e.g., emojis)
 - Removes whitespaces
 - Converts text to lowercase
 - Standardize handling of URLs and emails.
- **Profanity Check:** Uses a profanity model to identify and remove texts containing profanity. It is applicable for question and answer, instruct tuning, human-bot conversations, and continued pretraining tasks.
- **Text Quality Check:** Checks and filters out low-quality texts for question and answer, instruct tuning, human-bot conversations, and continued pretraining tasks. The app uses text grade technique to include texts within the desired grade range (school age) to ensure the quality of the text. The lower means too simple text, and the higher means too complex text.
- **Length Checker:** Filters the dataset based on user-defined minimum and maximum length parameters for all task types. By default, the context and answers should be within a range of 10-5000 characters while the questions should be within a range of 10-3000 characters.
- **Valid Question:** Uses a range of techniques to determine whether the question for question-answer pairs is actually a question. If it finds that there is no question in the pair, the app filters out that particular question-answer pair from further processing.
- **Pad Sequence:** Adds padding to the auto-generated question-and-answer pairs so that each text is of the same length.
- **Truncate Sequence by Score:** Allows truncation of the sequence based on a score and max length parameter required for all task types. By default, it truncates auto-generated text if the text is greater than 10,000 characters. It applies summarization to the text to reduce its length.
- **Compression Ratio Filter:** Filters text summarization data by comparing the compression ratio of the summaries. It removes rows for summarization tasks if the summarization ratio is greater than 35%. This is only relevant for summarization tasks.
- **Boundary Marking:** Adds `_START_` and `_END_` tokens at the boundaries of the summary text. This is only relevant for summarization tasks.
- **Sensitive Info Checker:** Identifies and removes any texts containing sensitive information (e.g., emails, phone numbers), critical for instruct tuning tasks.
- **RLHF Protection:** Appends datasets to facilitate RLHF for all task types.
- **Language Understanding:** Checks the language of text, allows filtering based on user inputs or threshold, beneficial for all task types.
- **Data Deduplication:** Calculates text similarity within the dataset and removes text based on a duplicate score threshold for all task types.
- **Toxicity Detection:** Calculates toxicity scores for text objects and filters according to a threshold beneficial for all task types.

Video guide

Watch [this](#) video guide to learn more about the key functions in data preparation for LLMs.

Create a project

Overview

Projects let you to manage and organize your data preparation tasks. Each project represents a specific data preparation task.

Instructions

The following steps describe how to create a project.

1. On the H2O LLM DataStudio left navigation menu, click **Prepare**.
2. In the **All Projects / Prepare Data for LLMs** page, click **New**.

All Projects / Prepare Data for LLMs

Total Projects: 5

Create a new project by clicking New button on the right

#	Name	Description	Type	Datasets	Status	Created	Modified
P1	Website Privacy	Online Privacy Policy QnA Dataset	Question Answering	1	Complete	2023-06-14	2023-06-14
P2	CNN-DailyMail	DailyMail News Text Summarization	Text Summarization	1	Complete	2023-06-14	2023-06-14
P3	Chat with AI	AI Chatbot with prompt response	Instruct Tuning	1	Complete	2023-06-14	2023-06-14
P4	OpenAssistant Conversations	A human-generated, human-annotated assistant-style conversations	Human Bot Conversations	1	Complete	2023-06-14	2023-06-14
P5	Wikipedia 512 Pretraining	Wikipedia preprocessed for pretraining of models. Each sample in the dataset has an average tokenized length of 512 RoBERTa-Base tokens.	Continued PreTraining	1	Complete	2023-06-14	2023-06-14

3. In the **Project name** text box, enter a name for the project (for example, My new project). 4. In the **Description** text box, enter a description for the project. 5. From the **Problem type** list, select the problem type you need for the data preparation task. **Note:** Once you have created a project with the selected problem type, you cannot change the problem type afterward. If you need to use a different problem type, you must create a new project from scratch.

6. Click **Add** to create the new project.

View projects

Overview

After creating projects through the H2O LLM DataStudio user interface, you can view a summary table with the following information about your current projects:

- **Name:** The name of the project.
- **Description:** A detailed summary highlighting the project's purpose and objectives.
- **Type:** The target task type of the project.
- **Datasets:** The number of datasets uploaded in the project.
- **Status:** The status of the data preparation process.
- **Created Date:** The date when the project was initially created.
- **Modified Date:** The date when the project most recently modified or updated.

View all projects

The following steps describe how to view all projects in H2O LLM DataStudio.

1. On the H2O LLM DataStudio left navigation menu, click **Prepare**.
2. To interact with a specific project, click on the project name to navigate the corresponding data preparation steps.

Delete a project

Overview

When a project is no longer needed, you can delete it. Deleted projects are permanently removed from the H2O LLM DataStudio instance.

Instructions

The following steps describe how to delete one or more projects.

1. On the H2O LLM DataStudio left navigation menu, click **Prepare**.
2. Click the **Delete** button.
3. Select project(s) to delete.
4. Click **Delete**.

Key functionalities

Converting documents to question-answer pairs and summarization pairs for fine tuning of LLMs using H2O LLM Data Studio supports the following key functionalities:

Variety of data types

With H2O LLM Data Studio, you can convert a variety of data types into question-answer pairs or summarization pairs. Supported data types include:

- Documents (pdf, docx, md, txt)
- Audio (.wav, m4a, mp3)
- Markdown and HTML
- Collections of the above in .zip format
- Web URLs and PDFs

LLM-Based question-answer pair generation

H2O LLM DataStudio utilizes the H2OGPT large open source LLM to use the documents as a reference to formulate and format question-answer pair generation. This capability handles the complete end to end pipeline from breaking down documents into chunks, using intelligent prompting techniques and ensuring consistent output formats.

LLM-Based context summarization pair generation

H2O LLM DataStudio's dataset curation capability can be used to generate context summarization pairs. It allows you to curate a dataset for another LLM fine-tuning workflow. This workflow uses the same smart chunking and prompting techniques to generate article-summary pairs. These article-summary pairs can be propagated to Prepare pipelines and LLM Studio for fine-tuning.

Fast QA Mode

The fast QA mode allows you to configure what proportion of input documents to use for question-answer generation. It identifies sections (chunks) of input documents that are diverse and content-rich, ensuring refined and diverse question-answer generation.

View and customize the output

- Reference Check: Explore the original document's text chunks to see where the question-answer pair was generated.
- Flag: Mark whether a row is relevant or irrelevant. These rows can be filtered out during data preparation.
- Edit: Customize and update any question-answer pair.
- Download Dataset: Download the datasets in either JSON or CSV format.
- Send to Prepare: Easily send the curated dataset to a Prepare project. This integration allows users to configure a data preparation workflow by augmenting the dataset with other curated collections or selecting public datasets, while also removing sensitive or toxic content and filtering out irrelevant rows.

Use the new structured dataset to finetune LLM in H2O LLM Studio.

You can get your dataset as a CSV for easy import into LLM Studio for fine-tuning. For more information, see [Effortless Fine-Tuning of Large Language Models with Open-Source H2O LLM Studio](#).

Create a new project for data curation

Overview

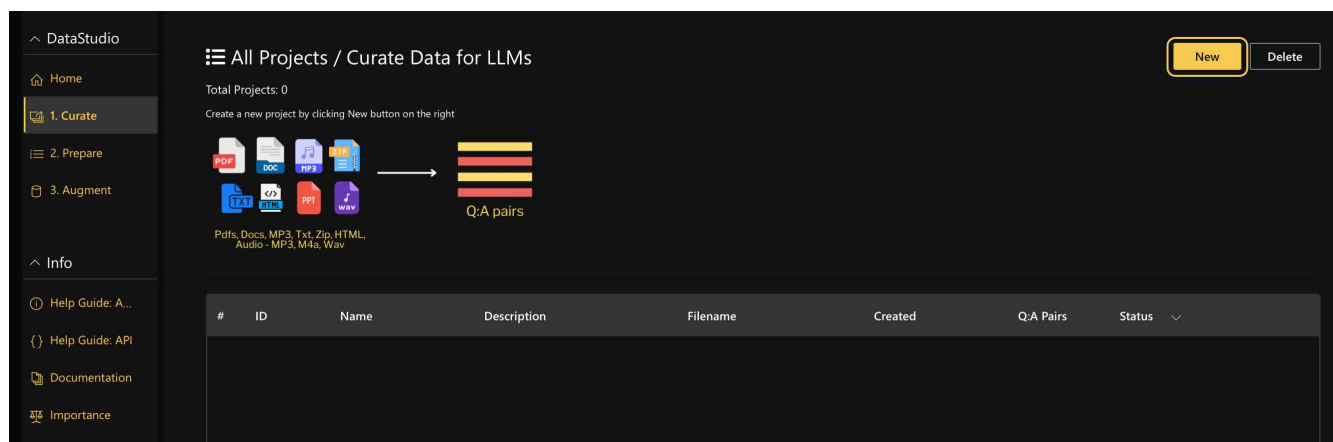
H2O LLM DataStudio supports the conversion of documents to question-answer pairs, article summarization, and file summarization. To do this, start by creating a new project.

Note: Before starting a new project, you must integrate **h2ogpte** by providing the required credentials. You cannot create a new project without configuring h2ogpte. For more information, see [Settings](#).

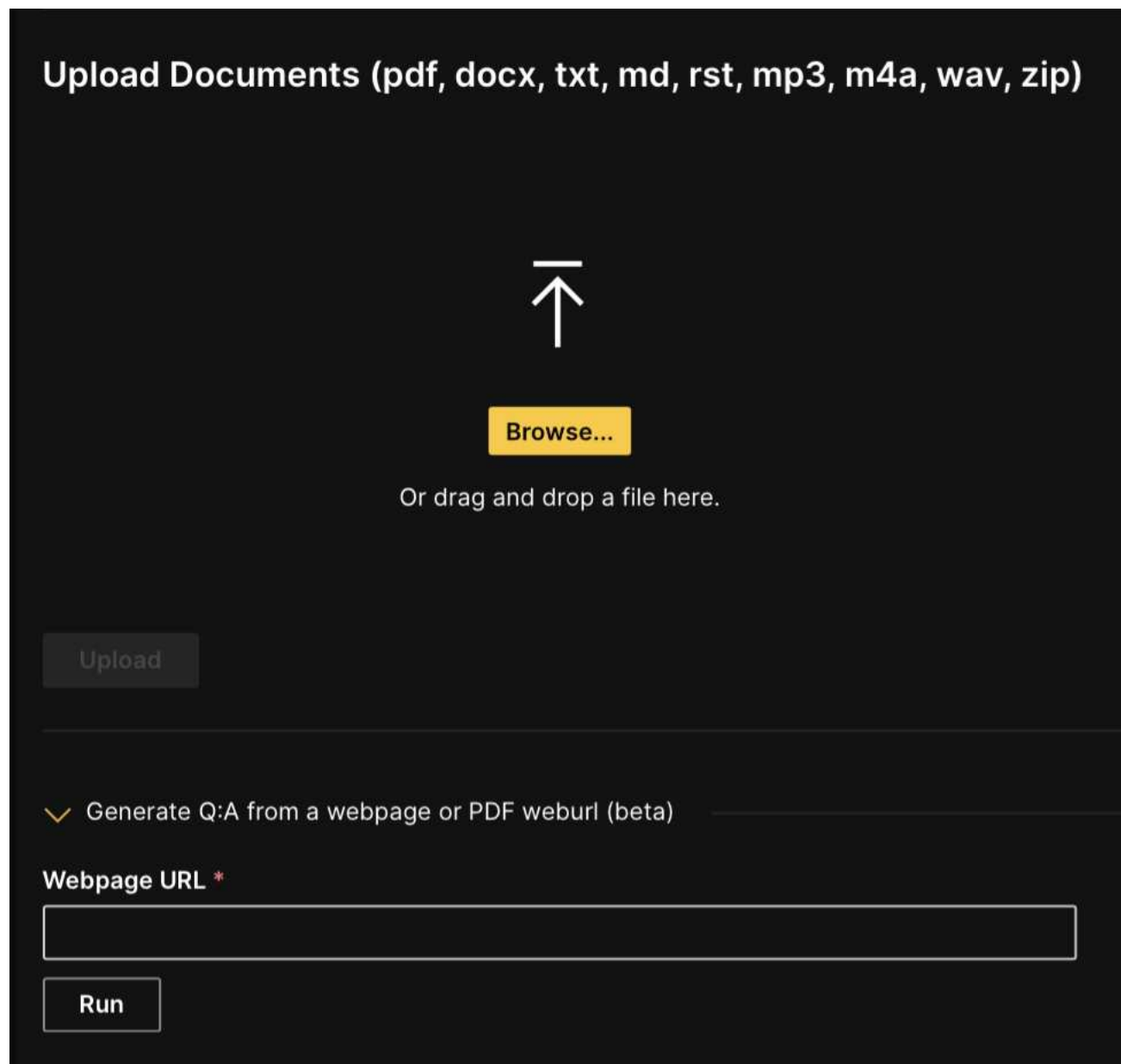
Instructions

To create a new project for data curation, consider the following instructions:

1. On the H2O LLM DataStudio left navigation menu, click **Curate**.
2. On the **All Projects / Curate Data for LLMs** page, click **New**.



3. On the **Project name** text box, enter a name for the project (for example, **My new curate project**).
4. On the **Description** text box, enter a description for the project.
5. On the **Document description** text box, provide a brief description for your file's content. This label helps to quickly categorize and identify the document's purpose. Some examples of document types are, **Quarterly Financial Report**, **Purchase Order**, **User Guide**, **Product Brief**, **Research Paper**, and **Meeting Recording**.
6. From the task type drop-down menu, select the appropriate task type for the experiment.
 - **Question-answer task:** Used to generate question-answer pairs from documents for LLM fine-tuning
 - **Summarization task:** Used to generate chunk summaries for LLM fine-tuning
 - **File Summary task:** Used to generate complete file summary (**Note:** This task type cannot be used for fine-tuning due to context limitations)
7. Click **Browse** and choose the file you want to upload or add the webpage URL and click **Run**. **Note:**
 - H2O LLM DataStudio supports PDF, DOCX, TXT, MD, RST and audio files (MP3, M4A, WAV).
 - If you have multiple documents you can upload them in a ZIP file.



Upload Documents (pdf, docx, txt, md, rst, mp3, m4a, wav, zip)

↑

Browse...

Or drag and drop a file here.

Upload

✓ Generate Q:A from a webpage or PDF weblurl (beta)

Webpage URL *

Run

8. Click **Upload** to upload the document.
 9. Once you upload the documents, select your preferred H2OGPTE LLM from the available LLMs on the **LLM selection** section.
 10. In the **Relevance score selection** dropdown menu, choose between the **Bert approach** (context-aware deep learning model), **Regex approach** (pattern-based matching), or **FinBert approach** (financial text relevance) to set your preferred method for calculating relevance scores.
 11. Use the slider labeled **Number of tokens per chunk** to adjust the tokenization settings. This controls the maximum number of tokens per chunk of text processed by the model. The default is set to 1000 tokens.
- Note:** More customization options are available only for the English language. For optimal results, we recommend using the **Bring your own prompt** in the Expert option. Please be aware that if you select options from multiple levels (basic, advanced, or expert options), the system will automatically apply the most advanced setting by default.

Basic options

1. In the **Select personas type** section, use the dropdown menu labeled **Personas types** to choose a predefined persona type for the model to emulate during the processing. For detailed descriptions on the different options, click the info icon next to the setting. If you want to create a custom persona, click the **Design own persona** button, which will guide you through the process of defining your own persona settings.

Advanced options

1. **Select knowledge prompts:** Select the knowledge prompts you want the system to focus on in the provided text box. For detailed descriptions of the different knowledge prompts, click the info icon next to the setting.
2. **Select question type:** From the dropdown menu, choose the type of questions you wish to generate. The options include factual, open-ended, inferential, analytical, predictive, and comparative. For detailed descriptions of each question type, click the info icon next to the setting.
3. **Select difficulty level of questions:** Use the dropdown to specify the desired difficulty level (easy, medium, or hard) for the questions generated.
4. **Select desired answer length:** Choose the preferred length of answers (short, medium, or long) from the dropdown options.

Expert options

1. **Bring your own Prompt:** Input a custom prompt to generate specific question-answer pairs. For example, you can provide detailed instructions on the type of questions and answers you need, targeting a specific audience or focusing on particular aspects of a document.
2. Enable the **Perform Smart Chunking (Fast)** option to optimize the processing of large documents, typically hundreds or thousands of pages. This feature speeds up the chunking process but may limit the generation of sufficient records for fine-tuning.
3. Adjust the **Sampling ratio** for smart chunking. By default, the sampling ratio is set to 0, and when the ratio is 0, LLM DataStudio will automatically select the sampling ratio based on the length of the document. It is recommended to set the sampling ratio to a value greater than or equal to 0.5.
4. Toggle the **Use h2oGPTe's ingestion pipeline** option to choose between using the h2oGPTe's ingestion pipeline or the default LLM DataStudio pipeline. If this is turned on, it will initiate the following process:
 1. A new collection is created in h2oGPTe.
 2. The document is uploaded directly to this collection.
 3. h2oGPTe generates and returns content chunks from the uploaded document.

If this is turned off, the ingestion pipeline of LLM DataStudio is used automatically.

5. Click **Run pipeline** to trigger the pipeline.

You can check the logs of the curation process from **Logs: Doc2QA Project**.

Click **Refresh** to review the progress and the percentage of completion.

LLM DataStudio
Data curation, preparation and ETL toolkit for LLMs

^ DataStudio

Home

1. Curate

2. Prepare

3. Custom Eval

4. Augment

Settings

Info

Help Guide: ...

Documentation

Importance

New Project / Curate Data for LLMs

Project Name: modest_elgamal Description: sad neumann Document Type: Text Task Type: Question-answer

File Uploaded: MigraineTreatment (1).zip

LLM Selection

Select your preferred H2OGPT LLM *

mistralai/Mixtral-8x7B-Instruct-v0.1

Smart Chunking

Recommended to use when document files are hundreds/thousands of pages.

Perform Smart Chunking (Fast) Off

Sampling Ratio: 0

Note: If the ratio=0, best ratio is chosen automatically.

Run Pipeline Discard

Logs: Doc2QA Project

Project Created: dqa_e2df1d7c-e35c-4044-93f8-3fb8b1088376

Documents Parsing

Text Chunking

Generating Q/A pairs from text chunks

> Processing file (1/14): The_diagnosis_and_treatment_of_chronic_migraine.pdf | 11.11% complete

Refresh Terminate

Click **Terminate** to stop the data curation process midway. It terminates the running process, and the question-answer pairs generated so far will be available to view and download.

View a Curate project

Overview

After creating **Curate** projects through the H2O LLM DataStudio user interface, you can view a summary table with the following information about your current projects:

- **ID:** The unique ID of the project.
- **Name:** The name of the project.
- **Description:** A detailed summary highlighting the project's purpose and objectives.
- **Filename:** The name of the uploaded document file.
- **Created:** The date when the project was initially created.
- **Q:A pairs:** The number of question-answer pairs generated from the uploaded document.
- **Number of documents:** The number of documents uploaded to the project.
- **Number of pages:** The number of total pages of documents in the project.
- **Status:** The status of each project. It indicates the progress of the project (Example: Running, Completed).

Total Projects: 1

Total Projects: 1
Create a new project by clicking New button on the right

PDFs, Docs, MP3, Txt, Zip, HTML, Audio - MP3, M4a, Wav

Q:A pairs / Summarization pairs

#	ID	Name	Description	Filename	Created	Q:A Pairs	# of Docum...	# of Pages	Status
1	dqa_45b6bd4	hopeful_goldberg	stoic booth	obama.zip	2024-04-03	100	2	13	Complete

View a specific Curate project

The following steps describe how to view a specific Curate project in H2O LLM DataStudio.

1. On the H2O LLM DataStudio left navigation menu, click **Curate**.
2. To interact with a specific project, click on the project name.

Inside the Curate project you selected, you can view the following details.

Status Complete Pairs 91 Sources Label Input Output JSON JSON CSV Publish as Preparation Project Publish as Custom Eval Project

#	Prompt	Answer
0	What is the name of the language	H2O-Danube-18B

- **Status:** The status of each project. It indicates the progress of the project (Example: Running, Complete).
- **Project details:** Click info to view project details.
 - The **Project details** tab shows the project name, project ID, project description, and the date and time when the project was initially created.
 - The **FastQA mode** indicates whether the smart chunking mode was activated, or not. The **Chunk sampling** shows the sampling ratio that has been used to convert documents into question-answer pairs.
- **Pairs:** The number of dataset entries generated from the uploaded document.
- **Reload/Logs:** Click refresh to review the events and actions that occurred during the data curation process.

- **View reference:** Select one or more entries from the table and click visibility to view all the references for the selected entry.
- **Label:**
 - Select the irrelevant entries from the table and click `disabled_by_default` to mark them as irrelevant.
 - Select the relevant entries from the table and click `check_box` to mark them as relevant.
- **Edit Q:A pairs:** Select a dataset entry and click edit to edit the generated entries.
- **Input:**
 - **View document:** Click description to view the uploaded PDF documents.
- **Output:**
 - Select the desired file type (JSON or CSV) from the dropdown menu. Click **Execute** to generate the question-answer pairs in the selected format.

Note: You can download the generated curation pairs even if the project fails or terminates during the curation process.

- Select **Publish as Preparation Project** from the dropdown menu to use the generated question-answer pairs in the [Data preparation flow](#). Once you click **Execute**, the dataset generated from the [Data curation process](#) will be ingested into the [Data preparation flow](#).
- Select **Publish as Custom Eval Project** from the dropdown menu to use the generated question-answer pairs to create your own evaluation dataset. Once you click **Execute**, the dataset generated from the [Data curation process](#) will be ingested into the [Custom Eval flow](#).
- **Export to H2O Drive:** Click `add_to_drive` to export the project to H2O Drive.
- Use the search bar to search for specific questions.
- The table of question-answer pairs includes the following details:
 - **Prompt:** The question
 - **Answer:** The corresponding answer
 - **Relevance:** Indicates a similarity calculation between the context and the answer. Each question-answer pair has a relevance score assigned. The relevance score is calculated as the ratio of matching sequences between the created answer and the original context which the answer was generated from. A relevance score of 1 indicates that the answer is directly quoted from the context. A relevance score of 0 means the answer has no overlapping words with the original context. The relevance score helps in filtering out questions that are most relevant.
 - **Filename:** The name of the document from which the specific question-answer pair was generated. You can filter the question-answer pair based on the filename.
- **Flag:** Indicates the question-answer pairs that have been identified as irrelevant. You can filter the question-answer pair based on the flag.

<div> <div>Search</div> </div>						
	#	Prompt	Answer	Relevan...	Filename	Flag
	0	What is the name of the language model presented in the paper?	H2O-Danube-1.8B	0.65574	H2O_Danube2.pdf	
	1	What is the size of the model's parameter range?	1.8 billion parameters	0.62284	H2O_Danube2.pdf	
	2	What is the training data used for the model?	A combination of web documents, encyclopedia, and public knowledge databases, excluding coding data.	0.61414	H2O_Danube2.pdf	
	3	What is the purpose of the H2O-Danube2-1.8B model?	It is trained on a more diverse mix of data over multiple data stages and is intended to demonstrate competitive performance across various benchmarks.	0.71365	H2O_Danube2.pdf	
	4	What is the significance of the Hugging Face Open LLM Leaderboard?	It is a ranking of large language models based on their performance on various benchmarks, and H2O-Danube2-1.8B is the highest ranked open model on the leaderboard for models below the 2B parameter range.	0.64232	H2O_Danube2.pdf	
	5	What is the difference between H2O-Danube-1.8B and H2O-Danube2-1.8B?	H2O-Danube2-1.8B is a continued modeling effort trained on additional 2T tokens, while H2O-Danube-1.8B is the initial model trained on 1T tokens.	0.70861	H2O_Danube2.pdf	
	6	What is the advantage of using smaller language models?	Smaller language models can provide efficient inference on consumer hardware and edge devices, and fine-tuning smaller models for specific tasks can enable competitive performance with benefits of model size and inference speed.	0.61731	H2O_Danube2.pdf	

91 of 91

Delete a Curate project

Overview

When a Curate project is no longer needed, you can delete it.

danger warning Deleted Curate projects are permanently removed from the H2O LLM DataStudio instance.

Instructions

The following steps describe how to delete one or more Curate projects.

1. On the H2O LLM DataStudio left navigation menu, click **Curate**.
2. Click the **Delete** button.
3. Select Curate project(s) to delete.
4. Click **Delete**.

Create your own evaluation datasets

Overview

The Custom Eval feature in H2O LLM DataStudio enables you to create your own evaluation datasets from various document formats (such as PDFs and DOC files), from audio and video files and existing datasets. These evaluation datasets can be downloaded in JSON formats for each evaluation type, allowing for seamless integration with [H2O Eval Studio](#).

Note: Custom Eval only supports English language.

Instructions

To create your own evaluation dataset, consider the following instructions:

Note: Before starting the process, you must integrate **h2ogpte** by providing the required credentials. You cannot create a new eval dataset without configuring h2ogpte. For more information, see [Settings](#).

1. On the H2O LLM DataStudio left navigation menu, click **Custom Eval**.
2. On the **Create Your Own Eval Datasets** page, click **New**.
3. On the **Project name** text box, enter a name for the project.
4. On the **Description** text box, enter a description for the project.
5. On the **Dataset type** drop-down menu, select the evaluation dataset type. The available dataset types are,
 -

Question type: In the dataset, each entry includes a specific question, the correct answer to that question, and a label indicating the type of question. You can configure the question type distribution after uploading the document.

LLM DataStudio
Data curation, preparation and ETL toolkit for LLMs

oshini.lnugapitiya@h2o.ai

Create Your Own Eval Datasets (Beta) New Delete

Project Name: test Description: Dataset Type: Question Type

File Uploaded: HeartDiseaseMedications.zip

LLM Selection
Select your preferred H2OGPTE LLM: mistralai/Mistral-8x7B-instruct-v0.1

Question Type Dataset
Change resulting question type distribution.

Question Type	Value
Sample Question	0.2
Conditional Question	0.3
Compress Question	0.25
Multi-hop Reasoning Question	0.25

Version: 0.51 Run Pipeline Discard

- **Simple question:** Simplifies the question and makes it easier to understand.
- **Conditional question:** Increases question complexity by incorporating scenarios or conditions affecting the context.
- **Compress question:** Makes the question more indirect and shorter while retaining its essence.
- **Multi-hop reasoning question:** Increases question complexity by requiring multiple logical connections or inferences.

- **Multi-Choice:** In this dataset, each entry consists of a question followed by multiple answer choices with one correct, or most appropriate answer and three incorrect choices.
- **Token presence:** In this dataset, each entry includes a question, the correct answer, and a list of key tokens that are relevant to the answer.

6. In the **Do you already have a QA Dataset?** drop-down menu, select **Yes** if you have a Q&A dataset in CSV format. Otherwise, click **No**.

If you select **No**, H2O LLM DataStudio will generate Q&A pairs from the uploaded document for you.

7. In the **Ingestion pipeline selection** section, toggle the **Use h2oGPTe's ingestion pipeline** option to choose between using the h2oGPTe's ingestion pipeline or the default LLM DataStudio pipeline.

If this is turned on, it will initiate the following process: 1. A new collection is created in h2oGPTe. 2. The document is uploaded directly to this collection. 3. h2oGPTe generates and returns content chunks from the uploaded document.

If this is turned off, the ingestion pipeline of LLM DataStudio is used automatically.

8. Click **Next**.

9. If you selected **Yes** on the **Do you already have a QA Dataset?** drop-down menu, follow these steps:

1. Click **Browse** to upload the Q&A dataset in CSV file format.
2. Click **Upload**.
3. In the **LLM selection** section, select your preferred H2OGPTE LLM from the available LLMs.
4. In the **Configure columns** section, select the columns which contain the context, question, and answer from the given options.
5. In the **Question type dataset** section, configure the [question type](#) distribution for the resulting evaluation dataset.
6. Click **Run pipeline**.

If you selected **No** on the **Do you already have a QA Dataset?** drop-down menu, follow these steps: 1. Click **Browse** to upload the document or add the webpage URL if you are generating question-answer pairs from a webpage, or PDF web URL. 2. Click **Upload**. 3. In the **LLM selection** section, select your preferred H2OGPTE LLM from the available LLMs. 4. Use the slider labeled **Number of tokens per chunk** to adjust the tokenization settings. This controls the maximum number of tokens per chunk of text processed by the model. The default is set to 1000 tokens. 5. In the **Question type dataset** section, configure the [question type](#) distribution for the resulting evaluation dataset. 6. Click **Run pipeline**.

View an evaluation dataset

Overview

After creating your own evaluation dataset through the H2O LLM DataStudio user interface, you can view a summary table with the following information about your current evaluation datasets:

- **ID:** The unique ID of the custom Eval project.
- **Name:** The name of the custom Eval project.
- **Description:** A detailed summary highlighting the custom Eval project's purpose and objectives.
- **Filename:** The name of the uploaded document file.
- **Eval dataset type:** The type of the evaluation dataset (question type, multi-choice, token presence).
- **Created:** The date when the project was initially created.
- **Number of entries:** The number of entries (rows) on the dataset.
- **Status:** The status of each project. It indicates the progress of the project (Example: **Running**, **Complete**).

View a specific Custom Eval project

The following steps describe how to view a specific Custom Eval project in H2O LLM DataStudio. 1. On the H2O LLM DataStudio left navigation menu, click **Custom Eval**. 2. To interact with a specific project, click on the project name.

Inside the Custom Eval project you selected, you can view the following details.

- **Status:** The status of each project. It indicates the progress of the project (Example: **Running**, **Complete**).
- **Pairs:** The number of question-answer pairs generated in the evaluation dataset.
- **Reload/logs:** Click refresh to review the events and actions that occurred during the process of creating an evaluation dataset.
- **Label:**
 - Select the question-answer pairs from the table and click `disabled__by__default` to mark them as irrelevant.
 - Select the question-answer pairs from the table and click `check__box` to mark them as relevant.
- **Edit Q:A pairs:** Select a question-answer pair and click edit to edit the dataset entries. Click **Update records** to update the dataset.
- **Input:**
 - **View document:** Click description to view the uploaded PDF documents.
- **Output:**
 - Click download to download the curated question-answer pairs in JSON or CSV file formats.
- **Generate robust eval dataset:** Select one or more rows from the generated question-answer pairs and click `file__open` to generate a robust evaluation dataset. The robust evaluation dataset contains the new questions generated based on the original question, the answer, and the original question.

test_Robust

Status **Complete** Pairs **60** Label Input Output

<input type="radio"/>	#	Question	Answer	Original Question	Filename	Flag
<input type="radio"/>	0	What is an Angiotensin Converting Enzyme (ACE) Inhibitor, and how does it work?	An Angiotensin Converting Enzyme (ACE) Inhibitor is a type of cardiac medication that decreases the action of angiotensin, a hormone in the body that causes blood vessels to narrow. This results in the dilation of blood vessels and lowering of blood pressure.	What is an Angiotensin Converting Enzyme (ACE) Inhibitor and what is its mechanism of action?	Cardiac-Medical	
<input type="radio"/>	1	Could you explain what an Angiotensin Converting Enzyme (ACE) Inhibitor is, and its mechanism of action?	An Angiotensin Converting Enzyme (ACE) Inhibitor is a type of cardiac medication that decreases the action of angiotensin, a hormone in the body that causes blood vessels to narrow. This results in the dilation of blood vessels and lowering of blood pressure.	What is an Angiotensin Converting Enzyme (ACE) Inhibitor and what is its mechanism of action?	Cardiac-Medical	
<input type="radio"/>	2	What is the function of an Angiotensin Converting Enzyme (ACE) Inhibitor, and how does it impact the body's hormones?	An Angiotensin Converting Enzyme (ACE) Inhibitor is a type of cardiac medication that decreases the action of angiotensin, a hormone in the body that causes blood vessels to narrow. This results in the dilation of blood vessels and lowering of blood pressure.	What is an Angiotensin Converting Enzyme (ACE) Inhibitor and what is its mechanism of action?	Cardiac-Medical	

Delete an evaluation dataset(s)

Overview

When a Custom Eval project is no longer needed, you can delete it.

danger warning Deleted a Custom Eval projects are permanently removed from the H2O LLM DataStudio instance.

Instructions

The following steps describe how to delete one or more Custom Eval projects:

1. On the H2O LLM DataStudio left navigation menu, click **Custom Eval**.
2. Click **Delete** and select the Custom Eval project(s) you wish to delete.
3. Click **Delete**.

Augment

Augmentation allows you to blend your datasets with publicly available datasets preloaded to the H2O LLM DataStudio to obtain variety. In some cases, you can integrate your datasets with RLHF-related datasets to include more domain aspects. The **Augment** tab displays a catalog of rich datasets that can be immediately used in the **Prepare** pipeline. Additionally, you can bring your own datasets for the augmentation process. Augmentation settings can be configured within the [Configuration](#) step in the **Prepare** pipeline.

Instructions

1. On the H2O LLM DataStudio left navigation menu, click **Augment**.
2. Click on the name of the dataset to get a preview of the dataset.

Augment Datasets (RLHF, Improve Content, Extra Rows)

H2O LLM DataStudio provides 18 datasets for different workflow types that you can augment with your input datasets during the [Configuration](#) step of data preparation. Also, the **Augment datasets** contains [RLHF](#) related datasets for question answering, text summarization, instruct tuning, and human-bot conversation problem types.

The DataCatalog table contains the following information about augmentation datasets:

- **Name:** The name of the dataset.
- **Description:** A brief description of the content of the dataset and its purpose.
- **Workflows:** The target workflow/task type of the dataset.
- **Rows:** The number of rows in the dataset.
- **Cols:** The number of columns in the dataset.
- **URL:** The source of the dataset.
- **License:** The [Open Data Commons](#) licenses that issued for the dataset owner.

Preloaded datasets

Stanford Q&A Dataset Stanford Question Answering Dataset consists of questions posed by crowdworkers on a set of Wikipedia articles where the answer to every question is from the corresponding reading passage.

- **Workflow type:** Question Answering
- **Number of rows:** 5000
- **Number of columns:** 3
- **URL:** <https://huggingface.co/datasets/squad>
- **License:** cc-by-4.0

QG-Bench Subset by SQuAD The SQUAD Dataset for question generation task.

- **Workflow type:** Question Answering
- **Number of rows:** 6283
- **Number of columns:** 3
- **URL:** https://huggingface.co/datasets/lmqg/qg_squad
- **License:** cc-by-4.0

Tweet base Q&A The Q&A dataset with short tweet, a question and a text phrase as the answer.

- **Workflow type:** Question Answering
- **Number of rows:** 10692
- **Number of columns:** 3
- **URL:** https://huggingface.co/datasets/tweet_qa
- **License:** cc-by-sa-4.0

RLHF EE QA The RLHF dataset for Q&A problems.

- **Workflow type:** Question Answering
- **Number of rows:** 180
- **Number of columns:** 3
- **URL:** https://huggingface.co/datasets/kastan/EE_QA_for_RLHF

- **License:** mit

News Article Summary The news article summary dataset contains summarized news from news articles from different newspapers.

- **Workflow type:** Text Summarization
- **Number of rows:** 4515
- **Number of columns:** 2
- **URL:** <https://www.kaggle.com/datasets/sunnysail2345/news-summary>
- **License:** gpl-2.0

Costco Article Summary The Costco article text summarization dataset.

- **Workflow type:** Text Summarization
- **Number of rows:** 86
- **Number of columns:** 2
- **URL:** https://huggingface.co/datasets/awinml/costco_long_practice
- **License:** mit

Dialogue Summary The dialogue summarization dataset.

- **Workflow type:** Text Summarization
- **Number of rows:** 12460
- **Number of columns:** 2
- **URL:** <https://huggingface.co/datasets/knkarthick/dialogsum>
- **License:** mit

RLHF OpenAI Summaries The RLHF OpenAI Summaries dataset contains sample (5000) of the CarperAI RLHF summarise dataset based on reddit thread summaries.

- **Workflow type:** Text Summarization
- **Number of rows:** 5000
- **Number of columns:** 2
- **URL:** https://huggingface.co/datasets/CarperAI/openai_summarize_comparisons

Code QA The Code QA dataset contains prompt-reply pairs where the prompt is to create a Python function which satisfies the functionality described in a specified docstring. The responses are the generated functions.

- **Workflow type:** Instruct Tuning
- **Number of rows:** 591
- **Number of columns:** 2
- **URL:** <https://huggingface.co/datasets/OllieStanley/humaneval-mbpb-codegen-qa>

Python QA The Python QA dataset contains prompt-reply pairs where the prompt is to create a Python unit test which tests for the functionality described in a specific docstring. The responses are the generated unit tests.

- **Workflow type:** Instruct Tuning
- **Number of rows:** 591
- **Number of columns:** 2
- **URL:** <https://huggingface.co/datasets/OllieStanley/humaneval-mbpb-testgen-qa>

Self Instruct The Self Instruct dataset contains prompt-reply pairs.

- **Workflow type:** Instruct Tuning
- **Number of rows:** 448
- **Number of columns:** 2

RLHF Instruct Tuning The RLHF Instruct Tuning dataset contains a technical Q&A set based on RLHF dataset.

- **Workflow type:** Instruct Tuning
- **Number of rows:** 337

- **Number of columns:** 2
- **URL:** <https://huggingface.co/datasets/kastan/rlhf-qa-comparisons>

Human Assistance Dataset The Human Assistance Dataset contains Human-Assistance style conversations, sampled to 5000 rows.

- **Workflow type:** Human Bot Conversations
- **Number of rows:** 33143
- **Number of columns:** 1
- **URL:** <https://huggingface.co/datasets/Dahoas/first-instruct-human-assistant-prompt>

Biomedical Human Assistance The Biomedical Human Assistance dataset contains the User-Assistant style conversations on biomedical.

- **Workflow type:** Human Bot Conversations
- **Number of rows:** 10000
- **Number of columns:** 1
- **URL:** https://huggingface.co/datasets/ericyu3/openassistant_inpainted_dialogs_5k_biomedical
- **License:** apache-2.0

User Assistant Conversations The User Assistant Conversations dataset contains the User-Assistant style conversations, sampled to 5000 rows.

- **Workflow type:** Human Bot Conversations
- **Number of rows:** 126287
- **Number of columns:** 2
- **URL:** https://huggingface.co/datasets/birgermoell/open_assistant_dataset

Anthropic RLHF Dataset sample The Anthropic RLHF Dataset sample dataset contains human preference data about helpfulness and harmlessness meant to train preference (or reward) models for subsequent RLHF training. This datasets takes a sample of 1000 entries.

- **Workflow type:** Human Bot Conversations
- **Number of rows:** 2332
- **Number of columns:** 1
- **URL:** <https://huggingface.co/datasets/Anthropic/hh-rlhf>
- **License:** mit

BERT Pre-training The BERT Dataset for pretraining.

- **Workflow type:** Continued PreTraining
- **Number of rows:** 20000
- **Number of columns:** 1
- **URL:** https://huggingface.co/datasets/nthngdy/bert_dataset_202203/viewer/nthngdy-bert_dataset_202203/train
- **License:** apache-2.0

TWT Eval The TWET Eval Pretraining Dataset.

- **Workflow type:** Continued PreTraining
- **Number of rows:** 20000
- **Number of columns:** 1
- **URL:** <https://huggingface.co/datasets/ArnavL/TWTEval-Pretraining-Processed>

Settings

Overview

H2O LLM DataStudio extends the dataset curation capabilities by integrating **H2OGPTE**. You can add or modify H2OGPTE credentials from the **Settings** page.

Note:

- You have the flexibility to use your own H2OGPTE credentials with a secure token.
- Before launching a [new Curate project](#), or a [new Custom Eval project](#), configure your H2OGPTE URL and H2OGPTE token on the Settings page.

Integrate h2ogpte

To integrate h2ogpte, consider the following instructions:

1. Enter the h2ogpte URL in the **H2OGPTE URL** text box. For more information, see [Access Enterprise h2ogpte](#).
2. Generate a new API Key in h2ogpte. For more information, see [Create an API key](#).
3. Copy the generated API Key and paste it in the **H2OGPTE Token** text box.
4. Click **Add**.

Key terms

H2O LLM DataStudio uses several key terms across its documentation, and each, in turn, is explained in the sections below.

LLM (Large Language Model)

A LLM refers to advanced AI models excelling in natural language understanding and generation, utilizing vast neural networks and extensive training data.

Data curation

Data curation refers to converting unstructured data like PDFs, DOCs, audio, and video files into structured formats, such as question-answer pairs or summaries. The goal is to make the information within those files more structured and easier to work with.

Tokenization

A process that breaks down text into smaller units, called tokens, to facilitate natural language processing tasks. Each token can be a word, subword, or character, allowing language models to analyze and understand text more effectively.

Truncation

The process of shortening text by removing characters from the beginning or end of a sequence, commonly used to fit text within specified length constraints or to prepare input data for language models with fixed input size.

RLHF (Reinforcement Learning with Human Feedback)

A training technique for Large Language Models (LLMs) that combines pre-training on a large corpus with fine-tuning on task-specific datasets, enhancing model performance for various tasks.

Augmentation

Augmentation allows you to blend your datasets with other publicly available datasets for the purpose of obtaining variety. In some cases, you can integrate your datasets with [RLHF](#) related datasets, to add more domain aspects. The Augment tab shows a catalog of rich datasets which can be used immediately in the Prepare pipeline. You can also bring your own datasets for the augmentation process. For more information, see [Augment](#).

HAMC

H2O AI Managed cloud is the main platform for users whereas H2O Admin Center is connected with the application that customers can use to configure their dedicated cloud deployment based on business use cases. H2O Admin Center supports firewall management and user management.

The fully managed solution of the H2O AI Cloud has all the same features to make, operate, and innovate with your own AI. The infrastructure and software management of the application are completely handled by H2O.ai allowing the customer to focus on solving their business problems with AI.

These new features include:

- A completely dedicated cloud environment for each customer.
- All operation activities (installation, upgrades, day-to-day operations) are handled by H2O.ai.
- Monitoring and automatic alerts of resource consumption.
- Quick onboarding process.
- Protection of public resources with several layers of security including DDoS protection, web application firewall, and firewalls.
- Optional security control of who can access resources by include-listing specific IP addresses.

Most users prefer controlling the application in a self-service manner so users get access to their account as well as inbound/outbound rules for moving data in/out of the platform.

H2O Managed Cloud is designed for high availability and dependability, providing tools to:

- Develop AI based applications
- Deploy AI based applications
- Manage AI based applications
- Maintain AI based applications

Which helps protect the confidentiality, integrity, and availability of your systems and data is our top priority.

Relevance score

The relevance score is a numerical value that quantifies the relationship between a query and a text segment. It is used to assess how well a specific piece of content matches a given search or context, enabling better sorting and prioritization of information. Various approaches, such as the Bert approach, Regex approach, and FinBert approach, offer different methods to calculate this score. The **Bert approach** uses a transformer-based model for a deep understanding of context, while the **Regex approach** relies on pattern matching, and the **FinBert approach** is fine-tuned for financial text.

Sampling ratio

Sampling ratio refers to the proportion of data selected from a larger dataset to be processed or analyzed. In the context of **Smart chunking**, the sampling ratio determines the fraction of the entire document that will be used to generate records. A lower ratio means that only a small subset of the data is sampled, which can be useful for handling very large datasets more efficiently. Conversely, a higher ratio involves more data, potentially improving the accuracy of the model but at the cost of increased processing time. If set to 0, the system automatically determines the optimal sampling ratio based on the document size and other factors.

Robust evaluation dataset

A robust evaluation dataset is a special set of test data used to evaluate how well a system can handle different types of questions. It includes new versions of the original questions, created by slightly changing how they are asked, but keeping the same answers. This ensures that the system is tested not only on the original questions but also on various forms of those questions, helping to check if it can consistently give the right answers even when the questions are worded differently.

FAQs

The sections below provide answers to frequently asked questions. If you have additional questions, please send them to cloud-feedback@h2o.ai.

General

What are the main workflows supported by H2O LLM DataStudio?

H2O LLM DataStudio supports several workflows and task types, including **Question and Answer models**, **Text Summarization**, **Instruct Tuning**, **Human-Bot Conversations**, and **Continued PreTraining** of language models. It provides tailored functionalities to optimize data preparation for each task type. For more information, see [Supported problem types](#).

What are some of the main features of H2O LLM DataStudio?

H2O LLM DataStudio offers a wide range of data preprocessing and preparation functions, such as text cleaning, text quality detection, tokenization, truncation, and data augmentation with external and RLHF datasets. It also has tools for length checking, relevance checking, profanity checking, and more. For more information, see [Supported functionalities](#).

What is the importance of good data for training Large Language Models?

Good data is crucial for training LLMs because it influences the accuracy, reliability, and effectiveness of the trained models. High-quality, well-prepared data lead to models that can understand and generate language more effectively. They provide more accurate responses in real-world applications.

Is the LLM DataStudio multi-user or single user?

The LLM DataStudio was created for a single user at a time but it is possible for multiple users to use the same instance. However there is no support for concurrent labeling as you may run the risk of having your label overridden if multiple people label at the same time.

Is there an API for LLM DataStudio?

No, there is no API for LLM DataStudio. It is a GUI based application.

If GPT is used, is LLM DataStudio still necessary, or is filtering handled automatically?

H2O DataStudio manages tasks beyond question-answer pair generation. This includes QA based on RAG, QA based on pure LLM, QA Diversity, as well as the management of the Prep Component and Augment Component. DataStudio facilitates various data-related functions, including filtering, available in the Prepare section of the workflow.

What does the relevance mean for the question-answer pairs?

The relevance in question-answer pairs pertains to how relevant the question is within the given contexts (chunks) of the document.

What are the models used in LLM DataStudio?

LLM DataStudio uses the [whisper-tiny model](#) to transcribe audio files as well as BERT models for relevance scores and performing clustering on text chunks to identify a diverse sample of text chunks if Smart Chunking turned on (we use BERT to create the embeddings from the text).

Prepare

How is detoxifying performed in LLM DataStudio?

Detoxifying uses a BERT model that was trained on the public [Jigsaw Unintended Bias in Toxicity dataset](#). This model returns predictions for general toxicity and toxicity subtypes; identity attack, insult, and threat. The thresholds that are used to flag a question-answer pair are configurable in the application during the Preparation phase. By default, any record with a prediction for any toxic features greater than 0.9 is dropped.

What are the file formats supported for exporting the processed data?

After the data preparation process is completed, the resulting dataset can be exported in various formats such as JSON, CSV, Parquet. The choice of the export file format may depend on your specific requirements and the nature of your downstream tasks.

How does H2O LLM DataStudio handle data quality?

H2O LLM DataStudio offers several tools to ensure data quality. It provides functionalities for text cleaning, text quality checking, profanity checking, and sensitive information detection. These tools help refine the raw datasets and ensure the quality and suitability for the training of LLMs.

Can I manage multiple data tasks in H2O LLM DataStudio?

Yes, H2O LLM DataStudio allows users to manage their data tasks effectively. It provides a user-friendly **Projects** tab where users can create, organize, and track their data preparation projects. For more information, see [View projects](#).

How does H2O LLM DataStudio ensure data privacy and safety?

H2O LLM DataStudio has several features to ensure data privacy and safety. It has tools to identify and remove texts containing sensitive information or profanity. By using these features, users can ensure that the data used for training LLMs are clean, safe, and suitable for the task at hand.

Are there redundancy checks in LLM DataStudio? (i.e. two question-answer pairs are very similar)

The LLM is prompted to create high quality questions and answers and the Smart Chunking (for sampling) ensures that diverse chunks are sampled (not a random sample). In the Smart Chunking mode, clusters are identified from the vector embeddings using k-Means and a sample of chunks from each cluster are utilized to create Question and Answer pairs. We also have a Deduplication check in the Prepare section (the data cleaning pipeline) which will remove question-answer pairs based on a duplicate score threshold. It uses MinHashing to convert the text to a hashing sequence. Jaccard similarity is then used to calculate the similarity on the hashing sequences and every question and answer pair that is above the similarity threshold (defaults to 0.9) is considered a duplicate and dropped.

Curate and Custom Eval

When converting documents to question-answer pairs, is there a way of seeing the progress of the data curation process in percentage?

Yes. On the **Logs: Doc2QA Project** panel, click Refresh to review the progress and view percentage of completion for each file of your document. For more information, see [Create a new project for data curation](#).

Can we stop the process midway at some point and download the questions-answer pairs generated so far before it's fully complete?

Yes, you can. To stop the data curation process midway, click **Terminate** on the **Logs: Doc2QA Project** panel. It terminates the running process, and the question-answer pairs generated so far will be available to download.

Can you create question-answer pairs on multiple files in one go?

Yes, you can create question-answer pairs on multiple files in one go by compressing the files into a zip file and uploading the zip file in to the LLM DataStudio.

How does LLM DataStudio work with audio files?

LLM DataStudio supports audio files. The uploaded audio files are converted to a transcript using whisper. LLM DataStudio then continues with the curation process as it would for a PDF/text document. LLM DataStudio uses the whisper-tiny model to do the transcription.

For question and answer generation in LLM DataStudio, can you select the LLM to use to create question-answer pairs?

When creating a new project for data curation and creating your own evaluation dataset, LLM DataStudio lets you to select Mixtral or Llama2 LLM.

Can I merge datasets in LLM DataStudio?

You can upload a ZIP file with multiple files in it and the question-answer pairs will be created for each file. You can determine the file it came from by the **File name** column. You can also use the [Augment](#) option when you Prepare a dataset. This will allow you to combine your question-answer dataset with an Augment dataset(s). The Augment datasets are preloaded fine-tuning datasets that are publicly available and are known to be helpful for fine-tuning.

How does chunking work in LLM DataStudio?

Each chunk is 4000 characters long and each chunk has a small overlap with each other. To speed up the creation of question-answer pairs, you can turn on [Smart Chunking Mode](#). In Smart Chunking Mode, clusters are identified from the vector embeddings using k-Means and a sample of chunks from each cluster are utilized to create question-answer pairs.

Augment

Does H2O LLM DataStudio support data augmentation?

Yes, H2O LLM DataStudio supports data augmentation. It enables the mixing or augmentation of multiple datasets together for all task types. This can be important for improving the robustness and performance of the trained models. In some cases you can integrate your datasets with RLHF related datasets to provide more domain aspects. The **Augment** tab shows a catalog of rich datasets which can be used immediately in the **Prepare** pipeline. You can also bring your own datasets for the augmentation process. For more information, see [Augment](#).