



## DOCUMENTATION

---

For questions, please contact [support@h2o.ai](mailto:support@h2o.ai)

# Contents

<b>Documentation structure and philosophy</b>	<b>8</b>
Structure . . . . .	8
Philosophy . . . . .	8
<b>What is Enterprise h2oGPTE?</b>	<b>9</b>
<b>Access Enterprise h2oGPTE</b>	<b>10</b>
Overview . . . . .	10
Instructions . . . . .	10
<b>Traditional Enterprise h2oGPTE workflow</b>	<b>11</b>
Overview . . . . .	11
Step 1: Create a Collection . . . . .	11
Step 2: Add Document(s) to a Collection . . . . .	11
Step 3: Chat with the Collection . . . . .	12
<b>Use cases (examples)</b>	<b>14</b>
Overview . . . . .	14
Use case 1: Document search . . . . .	14
Use case 2: Website search . . . . .	14
Use case 3: Financial reporting . . . . .	14
Use case 4: Knowledge base search . . . . .	14
Use case 5: Compliance and regulatory search . . . . .	14
Use case 6: Human resources . . . . .	15
Use case 7: IT support . . . . .	15
Use case 8: Sales and marketing . . . . .	15
<b>Concepts</b>	<b>16</b>
LLM . . . . .	16
GPT . . . . .	16
RAG . . . . .	16
LLM Prompt . . . . .	16
Transformer Neural Networks . . . . .	16
Fine-Tuning . . . . .	16
Self-Reflection . . . . .	16
<b>Tutorials: UI</b>	<b>17</b>
Learning path . . . . .	17
<b>Tutorial 1: A quick introduction to Enterprise h2oGPTE</b>	<b>19</b>
Overview . . . . .	19
Prerequisites . . . . .	19
Step 1: Create a Collection . . . . .	19
Step 2: Add Documents (annual report) to the Collection . . . . .	19
Step 3: Chat with the Collection (annual report) . . . . .	20
Summary . . . . .	23
Next . . . . .	23
<b>Tutorial 2: Build an AI-powered chatbot (model) to enhance a website's search capabilities</b>	<b>24</b>
Overview . . . . .	24
Objectives . . . . .	24
Prerequisites . . . . .	24
Step 1: Crawl website . . . . .	24
Step 2: Create a Collection-specific API Key . . . . .	26
Step 3: Chat with the H2O Model Validation website . . . . .	27
Summary . . . . .	28
Next . . . . .	28

<b>Tutorial 3: Importing and interacting with audio</b>	<b>29</b>
Overview . . . . .	29
Objectives . . . . .	29
Prerequisites . . . . .	29
Step 1: Create a Collection . . . . .	29
Step 2: Add the audio lecture to the Collection . . . . .	29
Step 3: Chat with the Collection (audio lecture) . . . . .	30
Summary . . . . .	30
Next . . . . .	30
<b>Tutorial 4: Importing and interacting with images</b>	<b>31</b>
Overview . . . . .	31
Objectives . . . . .	31
Prerequisites . . . . .	31
Step 1: Create a Collection . . . . .	31
Step 2: Add the image to the Collection . . . . .	31
Step 3: Chat with the Collection (image) . . . . .	31
Summary . . . . .	32
<b>Tutorial 5: Dataset analysis with Enterprise h2oGPTE agents</b>	<b>33</b>
Overview . . . . .	33
Objectives . . . . .	33
Prerequisites . . . . .	33
Step 1: Create a new Chat with an agent enabled . . . . .	33
Step 2: Enter query . . . . .	33
Step 3: Observe results . . . . .	34
Summary . . . . .	35
<b>Tutorial 6: Model development and preparation with Enterprise h2oGPTE agents</b>	<b>36</b>
Overview . . . . .	36
Objectives . . . . .	36
Prerequisites . . . . .	36
Step 1: Create a new Chat with an agent enabled . . . . .	36
Step 2: Enter query . . . . .	36
Step 3: Observe results . . . . .	37
Summary . . . . .	38
<b>Tutorial 7: Guardrails and personally identifiable information (PII) in Enterprise h2oGPTE</b>	<b>39</b>
Overview . . . . .	39
Objectives . . . . .	39
Prerequisites . . . . .	39
Step 1: Create a Collection with guardrails enabled . . . . .	39
Step 2: Chat with a Collection with guardrails enabled . . . . .	42
Summary . . . . .	43
<b>Tutorial 8: With an Extractor, transform unstructured document content into structured JSON data</b>	<b>44</b>
Overview . . . . .	44
Objective . . . . .	44
Prerequisites . . . . .	44
Step 1: Discuss Extractors . . . . .	44
Step 2: Create a Collection . . . . .	44
Step 3: Create an Extractor . . . . .	44
Step 4: Generate a structured response . . . . .	45
Step 5: View extracted information . . . . .	46
Summary . . . . .	46
<b>Agents</b>	<b>47</b>
Overview . . . . .	47
Instructions . . . . .	47
Tools . . . . .	47

Data analysis capabilities . . . . .	47
Information retrieval and research . . . . .	48
Content generation and transformation . . . . .	48
Automation and scripting . . . . .	48
Specialized AI features . . . . .	48
Multimedia processing . . . . .	48
<b>Chats</b>	<b>49</b>
Overview . . . . .	49
AI-powered chatbots . . . . .	49
<b>Start a new Chat with/without a Collection or an imported Document(s)</b>	<b>50</b>
Overview . . . . .	50
Instructions . . . . .	50
Add a Collection . . . . .	50
Import a Document(s) . . . . .	50
<b>A Chat session</b>	<b>52</b>
Overview . . . . .	52
Components of a Chat session . . . . .	52
<b>Customize a Chat session</b>	<b>54</b>
Overview . . . . .	54
Instructions . . . . .	54
Documents . . . . .	54
Collection to use . . . . .	54
Description . . . . .	54
Documents . . . . .	54
Configuration . . . . .	54
LLM . . . . .	54
Enable vision . . . . .	54
Vision LLM . . . . .	55
Use agent . . . . .	55
Generation approach . . . . .	55
Show Automatic LLM Routing Cost Controls . . . . .	56
Show Expert Settings . . . . .	57
Prompts . . . . .	58
Prompt template to use . . . . .	58
<b>Share a Chat session</b>	<b>59</b>
Overview . . . . .	59
Instructions . . . . .	59
<b>Run an Evaluator on a Chat</b>	<b>60</b>
Overview . . . . .	60
Instructions . . . . .	60
Chat Evaluators . . . . .	60
Toxicity . . . . .	60
Hallucination . . . . .	60
Personally Identifiable Information (PII) leakage . . . . .	60
Sensitive data leakage . . . . .	60
Fairness bias . . . . .	60
<b>Delete a Chat session</b>	<b>61</b>
Overview . . . . .	61
Instructions . . . . .	61
<b>Collections usage overview</b>	<b>62</b>
Overview . . . . .	62
Types of questions . . . . .	62

Further reading . . . . .	63
<b>Create a Collection</b>	<b>64</b>
Overview . . . . .	64
Instructions . . . . .	64
Collection settings . . . . .	64
General . . . . .	64
Configuration . . . . .	65
Default chat settings . . . . .	67
<b>Add a Document(s) to a Collection</b>	<b>69</b>
Overview . . . . .	69
Instructions . . . . .	69
<b>Chat with a Collection</b>	<b>71</b>
Overview . . . . .	71
Instructions . . . . .	71
<b>Update a Collection's settings</b>	<b>72</b>
Overview . . . . .	72
Instructions . . . . .	72
<b>Supported file types for a Collection</b>	<b>73</b>
<b>Make a Collection public</b>	<b>81</b>
Overview . . . . .	81
Instructions . . . . .	81
<b>Run an Evaluator on a Collection</b>	<b>82</b>
Overview . . . . .	82
Instructions . . . . .	82
Collection Evaluators . . . . .	82
Toxicity . . . . .	82
Hallucination . . . . .	82
Personally Identifiable Information (PII) leakage . . . . .	82
Sensitive data leakage . . . . .	82
Fairness bias . . . . .	82
<b>Create a topic model for a Collection</b>	<b>83</b>
Overview . . . . .	83
Instructions . . . . .	83
<b>Delete a Collection</b>	<b>84</b>
Overview . . . . .	84
Delete a specific Collection . . . . .	84
Delete multiple Collections . . . . .	84
<b>Generate a Document summary</b>	<b>85</b>
Overview . . . . .	85
Instructions . . . . .	85
Document AI settings . . . . .	85
LLM . . . . .	85
Enable vision . . . . .	85
Vision LLM . . . . .	86
Max. number of chunks and approximate cost range . . . . .	86
Prompt Template . . . . .	86
Personality (System Prompt) . . . . .	86
Prompts before/after document context . . . . .	86
Prompt per image batch for vision models . . . . .	86
Prompt for final image batch reduction for vision models . . . . .	86

JSON Schema . . . . .	86
PII Detection . . . . .	86
Document Metadata to include . . . . .	86
Keep intermediate results . . . . .	87
<b>View a Document summary</b>	<b>88</b>
Overview . . . . .	88
Instructions . . . . .	88
<b>Delete a Document summary</b>	<b>89</b>
Overview . . . . .	89
Instructions . . . . .	89
<b>Delete a Document(s)</b>	<b>90</b>
Overview . . . . .	90
Instructions . . . . .	90
<b>Models</b>	<b>91</b>
Overview . . . . .	91
Supported LLMs . . . . .	91
Run self-tests . . . . .	91
Overview . . . . .	91
Instructions . . . . .	91
Tabs . . . . .	92
LLMs . . . . .	92
Charts . . . . .	92
Usage and performance . . . . .	93
<b>Prompts</b>	<b>95</b>
Create a prompt template . . . . .	95
Overview . . . . .	95
Instructions . . . . .	95
Create a prompt template for a specific language . . . . .	95
Clone a prompt template . . . . .	95
Overview . . . . .	95
Instructions . . . . .	96
Prompt template settings . . . . .	96
General . . . . .	96
Prompts . . . . .	96
Self-reflection . . . . .	99
Sample questions . . . . .	100
<b>Extractors</b>	<b>102</b>
Overview . . . . .	102
Extractor flow . . . . .	102
Create an Extractor . . . . .	102
Run an Extractor . . . . .	103
View a completed Extractor . . . . .	103
<b>Evaluation</b>	<b>105</b>
Overview . . . . .	105
Instructions . . . . .	105
Add a new leaderboard . . . . .	105
Overview . . . . .	105
Instructions . . . . .	105
<b>Feedback</b>	<b>106</b>
View submitted feedback (Upvote and Downvote responses) . . . . .	106
Overview . . . . .	106
Instructions . . . . .	106

<b>Impersonate another user</b>	<b>107</b>
Overview	107
Instructions	107
<b>APIs</b>	<b>108</b>
Create an API Key	108
Overview	108
Instructions	108
Delete an API Key	108
Overview	108
Instructions	108
Types of API Keys	109
Global API Keys	109
Collection-specific API Keys	109
<b>Jobs</b>	<b>110</b>
View a Job	110
Overview	110
Instructions	110
Terminate a Job	110
Overview	110
Instructions	110
<b>Connectors</b>	<b>112</b>
Overview	112
Instructions	112
Connectors	112
<b>Admin center</b>	<b>114</b>
Overview	114
Instructions	114
RBAC and Jobs	115
<b>h2oGPTe REST API: OpenAPI specification file</b>	<b>116</b>
Overview	116
How we generated the SDKs	116
General CLI structure	116
<b>Python SDK</b>	<b>117</b>
Overview	117
Steps to generate and test the Python SDK	117
<b>JavaScript SDK</b>	<b>119</b>
Overview	119
Steps to generate and test the JavaScript SDK	119
<b>Go SDK</b>	<b>121</b>
Overview	121
Steps to generate and test the Go SDK	121
<b>Key terms</b>	<b>123</b>
Collection	123
Job	123
Document	123
Chat	123
API Key	123
Extractors	123
PII Detection	123
Evaluators	124

<b>Vector database</b>	<b>125</b>
Embedded Vex vector database . . . . .	125
External vector databases . . . . .	125
<b>Major releases</b>	<b>126</b>
<b>FAQs</b>	<b>127</b>
General . . . . .	127
How does Enterprise h2oGPTe handle embedded images in PDFs? . . . . .	127
Does Enterprise h2oGPTe have an API? . . . . .	127
Which Embedding Models are supported in Enterprise h2oGPTe? . . . . .	127
What is the Vector Database for Enterprise h2oGPTe (1) and can I customize it (2)? . . . . .	127
Collections . . . . .	128
Can Enterprise h2oGPTe refer to several Collections to answer a query? . . . . .	128
What file types (documents) does Enterprise h2oGPTe Collections support? . . . . .	128
Documents . . . . .	128
Are uploaded documents cleared after the user session, or do they persist? . . . . .	128
How are entitlements managed so that users are not querying each other's documents or the model is not picking chunks from other documents? . . . . .	128
Chats . . . . .	128
Can I chat with a Collection in any language? . . . . .	128



# Documentation structure and philosophy

## Structure

The Enterprise h2oGPTe documentation is organized into the following main sections:

- **Get started:** This section introduces Enterprise h2oGPTe.
- **Concepts:** This section provides fundamental concepts surrounding Enterprise h2oGPTe.
- **Tutorials: UI:** This section offers tutorials on Enterprise h2oGPTe.
- **Guide:** This section encompasses several subsections:
  - **Chats:** This subsection covers Chats in Enterprise h2oGPTe.
  - **Collections:** This subsection covers Collections in Enterprise h2oGPTe.
  - **Documents:** This subsection covers Documents in Enterprise h2oGPTe.
  - **Models:** This subsection covers how to compare and test large language models (LLMs) used across Enterprise h2oGPTe.
  - **Prompts:** This subsection covers how to create a prompt templates in Enterprise h2oGPTe.
  - **Evaluation:** This subsection covers how to assess model performance through leaderboards created in H2O Eval Studio.
  - **Feedback:** This subsection covers how you can view past submitted feedback on certain Upvote and Updown responses.
  - **Impersonate another user:** This subsection covers the user impersonation feature that lets you temporarily sign in as a different user.
  - **APIs:** This subsection provides guidance on creating API Keys for integrating external applications to Enterprise h2oGPTe.
  - **Jobs:** This subsection covers Jobs in Enterprise h2oGPTe.
  - **Admin center:** This subsection covers the admin center page that offers administrators an overview of recent Jobs.
- **h2oGPTe Python Client Library:** This section offers several examples of how to use the h2oGPTe Python Client Library.
- **Changelog:** This subsection blogs about the major Enterprise h2oGPTe releases.
- **Key terms:** This section defines key terms relevant to Enterprise h2oGPTe.
- **Architecture:** This subsection covers the Enterprise h2oGPTe architecture and vector database.
- **Major releases:** This subsection covers the major Enterprise h2oGPTe releases.
- **FAQs:** This section answers frequently asked questions about Enterprise h2oGPTe.

## Philosophy

- **Philosophy of alignment:** The documentation for Enterprise h2oGPTe is designed to match the product's structure. This means that information in the documentation is organized to align with the features and workflows of Enterprise h2oGPTe. Our goal is to make learning the product easy and intuitive for users.
- **Simplicity and modularity principles:** Our documentation follows a simple and modular approach. Most pages have two main sections: "Overview" and "Instructions." A page can also have other sections that heavily relate to the page's "Overview" and "Instructions." This makes it easier for users to understand and use Enterprise h2oGPTe effectively.

## What is Enterprise h2oGPTe?

Enterprise h2oGPTe is an AI-powered search assistant that helps you find answers to questions about your documents, websites, and workplace content - powered by H2O LLM.

- With Enterprise h2oGPTe, finding what you need has never been easier
- Ask a question, and Enterprise h2oGPTe will provide relevant results from across provided materials
- Whether you're looking for a specific document or trying to understand a complex concept, Enterprise h2oGPTe can help you save time and effort

# Access Enterprise h2oGPTE

## Overview

You can utilize Enterprise h2oGPTE through your **H2O Generative AI** app store account, with access available at a freemium tier.

**Caution:** Freemium The freemium version of Enterprise h2oGPTE is designed for users who want to explore the product's comprehensive feature set without committing to a paid subscription. This tier grants full access to the product's functionalities but with certain limitations focusing on sharing capabilities and resource utilization.

In the freemium version, users are restricted from sharing **Collections** or **Chats**. Additionally, there are defined thresholds for the volume/number of Collections, **Documents**, and Chats that can be created, as well as usage constraints on Large Language Models (LLMs).

It's important to note that the freemium version operates on infrastructure that may exhibit slightly reduced performance compared to the premium version. This may make it less suitable for high-throughput operations or large-scale projects where performance is critical. However, for individual endeavors or small-scale projects where performance is not a top priority, the freemium version remains a cost-effective and accessible solution.

To upgrade Enterprise h2oGPTE, please get in touch with **sales@h2o.ai**.

## Instructions

The following steps describe how to access Enterprise h2oGPTE:

1. Access your **H2O Generative AI** app store account.
2. In the **What do you want to search?** box, enter: **Enterprise h2oGPTE**
3. Click the **Enterprise h2oGPTE** card.

# Traditional Enterprise h2oGPTe workflow

## Overview

The traditional workflow in Enterprise h2oGPTe to Chat with one of your documents consists of these 3 sequential steps:

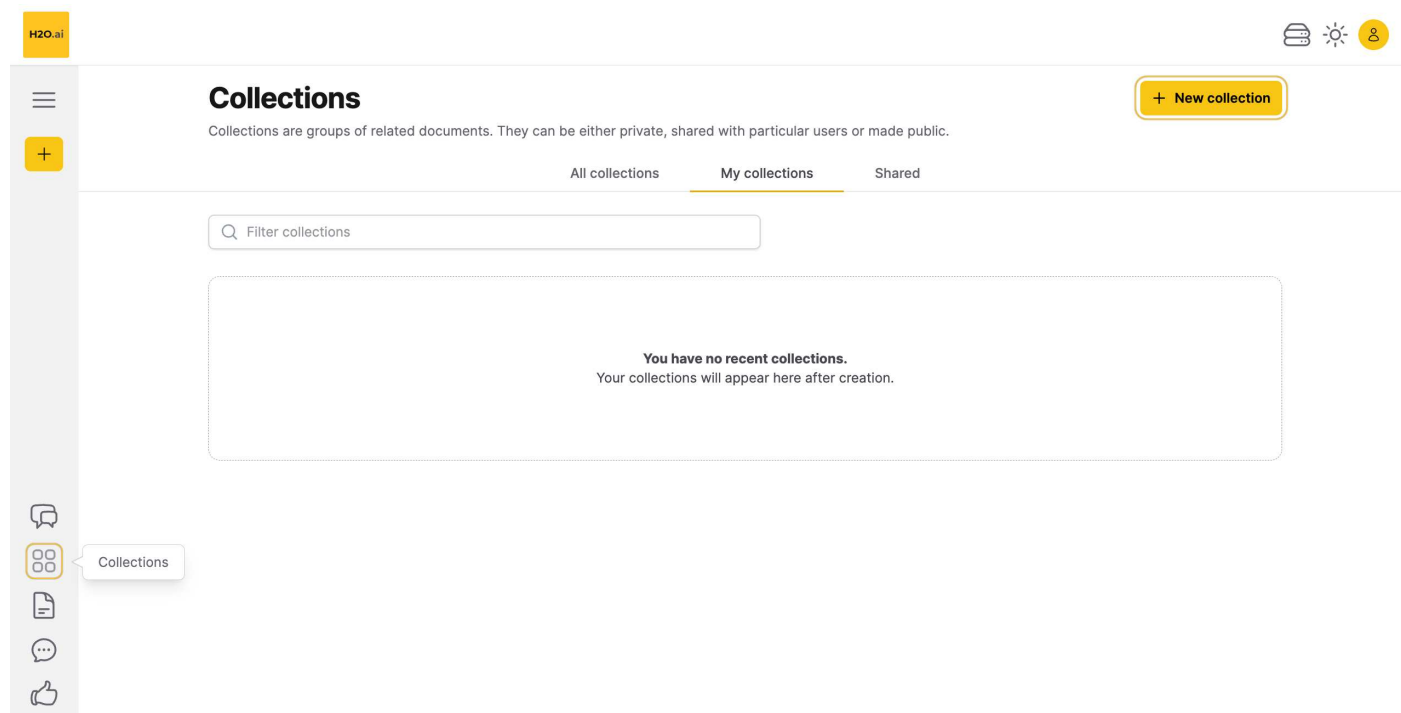
- [Step 1: Create a Collection](#)
- [Step 2: Add Document\(s\) to a Collection](#)
- [Step 3: Chat with the Collection](#)

**Note:** The discussed flow on this page is implemented in the following tutorial: [Tutorial 1A: A quick introduction to Enterprise h2oGPTe](#).

## Step 1: Create a Collection

Create a Collection (a group of related documents) as the first step in the flow. In this case, “related Documents” refer to the Document(s) you want to ask questions about in Enterprise h2oGPTe.

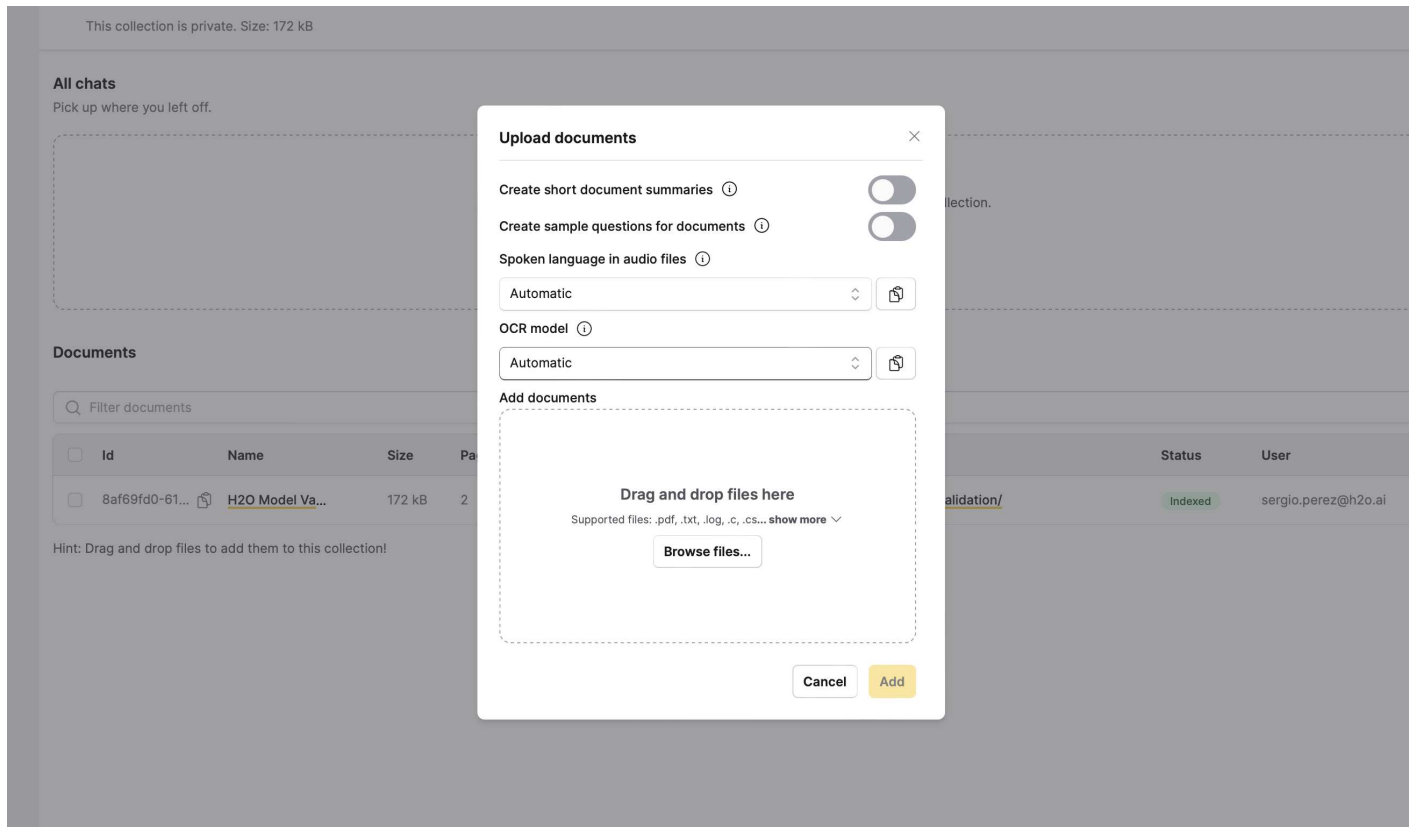
- To learn how to create a Collection, see [Create a Collection](#).



## Step 2: Add Document(s) to a Collection

Add Document(s) to the Collection as the second step in the flow. Added Documents are indexed and stored in a database.

- To learn how to add Document(s) to a Collection, see [Add a Document\(s\) to a Collection](#).



### Step 3: Chat with the Collection

Create a Chat with the Collection as the final step in the flow. Chatting with the Collection lets you ask h2oGPTe questions about the Document(s) in the Collection. When answering a question, Enterprise h2oGPTe provides the Document(s) references it utilized to respond.

When you ask a question about the Document(s), Enterprise h2oGPTe crawls through the indexed Document(s) in the Collection to find relevant content to answer the question while utilizing the H2O LLM to summarize a concise question response.

- To learn how to Chat with a Collection, see [Chat with a Collection](#).

H2O.ai

**You**

Summarize Driverless AI (DAI)

**H2O**

Driverless AI (DAI) is an automated machine learning platform developed by H2O.ai that aims to achieve high predictive accuracy comparable to expert data scientists in a shorter time. It automates various data science and machine learning workflows, including feature engineering, model validation, model tuning, model selection, and model interpretation. DAI supports both Python and Java standalone scoring pipelines, making it suitable for various industries and applications. Additionally, DAI offers a user-friendly experiment setup wizard, ensuring optimal configuration for specific use cases. It also supports multi-GPU workstations and servers for faster processing. Users can install DAI via Docker, native, or cloud installation methods, and configure it using various options and environment variables. DAI supports datasets in various formats and offers automatic visualization and machine learning interpretability features.

Hide references ▾

**Overview Using Driverless AI 1107 documentation.pdf**

Page 6 Highest Score 0.993



Page 1 Highest Score 0.990



Page 7 Highest Score 0.980

## Use cases (examples)

### Overview

This page provides an overview of the various example use cases for Enterprise h2oGPTE, a powerful AI-powered search and answer tool. The page highlights the benefits of using Enterprise h2oGPTE for document search, website search, financial reporting, knowledge base search, compliance and regulatory search, human resources, IT support, sales and marketing, and other business applications.

The page also provides examples of how Enterprise h2oGPTE can be used in each of these use cases, such as searching for company policies, finding information about a product launch, or troubleshooting a common issue in a company's software.

Overall, the page provides a comprehensive overview of the capabilities and benefits of Enterprise h2oGPTE, making it an ideal resource for anyone looking to improve their search and answer capabilities.

### Use case 1: Document search

- **Overview:** Enterprise h2oGPTE can be used to search through a large collection of documents, such as company policies, procedures, and guidelines, to find specific information. Users can ask a question or provide a keyword, and Enterprise h2oGPTE will return relevant results from the documents.
- **Example:** A user wants to find the company's policy on remote work. They can ask a question or provide the keyword "remote work" and Enterprise h2oGPTE will return the relevant policy document.
- **Benefits:** This feature helps users quickly find the information they need, reducing the time spent searching through documents and improving productivity.

### Use case 2: Website search

- **Overview:** Enterprise h2oGPTE can be integrated with a company's website to provide a more efficient search experience. Users can ask a question or provide a keyword, and Enterprise h2oGPTE will return relevant results from the website's content.
- **Example:** A user wants to find information about a product offered by the company. They can ask a question or provide the keyword "product" and Enterprise h2oGPTE will return the relevant product page.
- **Benefits:** This feature helps users quickly find the information they need on the company's website, improving the user experience and reducing the bounce rate.

### Use case 3: Financial reporting

- **Overview:** Enterprise h2oGPTE can be used to search through a company's financial reports, such as balance sheets, income statements, and cash flow statements, to find specific information. Users can ask a question or provide a keyword, and Enterprise h2oGPTE will return relevant results from the reports.
- **Example:** A user wants to find information about the company's revenue for the past quarter. They can ask a question or provide the keyword "revenue" and Enterprise h2oGPTE will return the relevant revenue values from the financial report.
- **Benefits:** Enterprise h2oGPTE can help users quickly find the information they need, reducing the time spent searching through financial reports and improving productivity.

### Use case 4: Knowledge base search

- **Overview:** Enterprise h2oGPTE can be used to search through a company's knowledge base, which is a centralized repository of information that contains answers to frequently asked questions, troubleshooting guides, and other helpful resources. Users can ask a question or provide a keyword, and Enterprise h2oGPTE will return relevant results from the knowledge base.
- **Example:** A user wants to find information about a common issue that occurs in the company's software. They can ask a question or provide the keyword "issue" and Enterprise h2oGPTE will return the relevant troubleshooting guide.
- **Benefits:** Enterprise h2oGPTE can help users quickly find solutions to common issues, reducing the time spent searching through the knowledge base and improving productivity.

### Use case 5: Compliance and regulatory search

- **Overview:** Enterprise h2oGPTE can be used to search through a company's compliance and regulatory documents, such as contracts, policies, and regulations, to find specific information. Users can ask a question or provide a

keyword, and Enterprise h2oGPTe will return relevant results from the documents.

- **Example:** A user wants to find information about the company’s data privacy policy. They can ask a question or provide the keyword “privacy” and Enterprise h2oGPTe will return the relevant policy document.
- **Benefits:** Enterprise h2oGPTe can help users quickly find the information they need, reducing the time spent searching through compliance and regulatory documents and improving productivity.

## Use case 6: Human resources

- **Overview:** Enterprise h2oGPTe can be used to search through a company’s human resource (HR) policies, procedures, and guidelines, to find specific information. Users can ask a question or provide a keyword, and Enterprise h2oGPTe will return relevant results from the HR materials.
- **Example:** A user wants to find information about the company’s maternity leave policy. They can ask a question or provide the keyword “maternity leave” and Enterprise h2oGPTe will return the relevant HR policy document.
- **Benefits:** Enterprise h2oGPTe can help users quickly find the information they need, reducing the time spent searching through HR materials and improving productivity.

## Use case 7: IT support

- **Overview:** Enterprise h2oGPTe can be used to search through a company’s IT documentation, such as user guides, technical specifications, and troubleshooting guides, to find specific information. Users can ask a question or provide a keyword, and Enterprise h2oGPTe will return relevant results from the IT documentation.
- **Example:** A user wants to find information about the company’s network infrastructure. They can ask a question or provide the keyword “network” and Enterprise h2oGPTe will return the relevant technical specification document.
- **Benefits:** Enterprise h2oGPTe can help users quickly find the information they need, reducing the time spent searching through IT documentation and improving productivity.

## Use case 8: Sales and marketing

- **Overview:** Enterprise h2oGPTe can be used to search through a company’s sales and marketing materials, such as product descriptions, pricing information, and customer testimonials, to find specific information. Users can ask a question or provide a keyword, and Enterprise h2oGPTe will return relevant results from the sales and marketing materials.
- **Example:** A user wants to find information about the company’s new product launch. They can ask a question or provide the keyword “product launch” and Enterprise h2oGPTe will return the relevant marketing material.
- **Benefits:** Enterprise h2oGPTe can help users quickly find the information they need, reducing the time spent searching through sales and marketing materials and improving productivity.



## Concepts

Enterprise h2oGPTe uses several key terms across its documentation, and each, in turn, is explained in the following sections.

### LLM

A Large Language Model (LLM) is a type of AI model that uses deep learning techniques and uses massive datasets to analyze and generate human-like language. For example, many AI chatbots or AI search engines are powered by LLMs.

Generally speaking, LLMs can be characterized by the following parameters:

- Size of the training dataset
- Cost of training (computational power)
- Size of the model (parameters)
- Performance after training (or how well the model is able to respond to a particular question)

### GPT

GPT, short for **Generative Pre-Trained Transformer**, is an advanced open-source language model that utilizes transformer architectures to generate human-like text. It is trained on vast amounts of unlabeled text data from the internet, enabling it to understand and generate coherent and contextually relevant text. Unlike rule-based systems, GPT learns patterns and structures in text data to generate human-like responses.

For more information, see [GPT \(Generative Pre-Trained Transformer\)](#).

### RAG

**Retrieval-augmented generation (RAG)** is an AI framework for improving the quality of responses generated by Large Language Models (LLMs) by grounding the model on external sources of knowledge. RAG-equipped chatbots absorb their information from a variety of sources, including databases, documents, and the internet, to provide accurate and contextually relevant responses. This is particularly useful when users have complex or multi-step queries. Using a RAG system contributes significantly towards making the business more agile, especially if the company has a customer-facing chatbot.

For more information, see [Boosting LLMs to New Heights with Retrieval Augmented Generation](#).

### LLM Prompt

A Large Language Model (LLM) Prompt is a question or request you send to an LLM to generate a desired response. This can be a question you want the LLM to answer or a request for the LLM to complete. The goal of using an LLM Prompt is to elicit a specific response from the model, whether it be a piece of information, a summary, or a creative work.

### Transformer Neural Networks

Neural networks are an efficient way to solve machine learning problems and can be used in various situations. Neural networks offer precision and accuracy. Finding the correct neural network for each project can increase efficiency. Recurrent neural networks (RNNs) remember previously learned predictions to help make future predictions with accuracy. Unlike RNNs, **Transformer Neural Networks** do not have a concept of timestamps. This enables them to pass through multiple inputs at once, making them a more efficient way to process data.

For more information, see [Transformer Architecture](#).

### Fine-Tuning

Fine-Tuning refers to the process of taking a pre-trained language model and further training it on a specific task or domain to improve its performance on that task. It is an important technique used to adapt Large Language Models (LLMs) to specific tasks and domains.

### Self-Reflection

In Enterprise h2oGPTe, Self-Reflection asks another Large Language Model (LLM) to reflect on the answer given to the question based on the provided context. Self-reflection can be used to evaluate the LLM's performance.

## Tutorials: UI

Learn how to find information 10x faster

### Learning path

graph LR;

```
User[User] --> A1[Tutorial 1: A quick introduction to Enterprise h2oGPTe];
User --> B1[Tutorial 2: Build an AI-powered chatbot to enhance a website's search capabilities];
User --> C1[Tutorial 3: Importing and interacting with audio];
User --> D1[Tutorial 4: Importing and interacting with images];
User --> Agents[Agents]
Agents --> E1[Tutorial 5: Dataset analysis with Enterprise h2oGPTe agents];
Agents --> F1[Tutorial 6: Model development and Regex preparation with Enterprise h2oGPTe agents];
User --> GuardrailsPII[Guardrails and PII]
GuardrailsPII --> G1[Tutorial 7: Guardrails and personally identifiable information in Enterprise h2oGPTe]
User --> Extractors[Extractors]
Extractors --> H1[Tutorial 8: With an Extractor, transform unstructured document content into structured]
%% 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 Agents fill:#FEC925;
style E1 fill:#FEC925;
style F1 fill:#FEC925;
style G1 fill:#FEC925;
style GuardrailsPII fill:#FEC925;
style H1 fill:#FEC925;
style Extractors fill:#FEC925;
%% Add links to each node
click A1 "/enterprise-h2ogpte/tutorials/tutorial-1" "Go to Tutorial 1"
click B1 "/enterprise-h2ogpte/tutorials/tutorial-2" "Go to Tutorial 2"
click C1 "/enterprise-h2ogpte/tutorials/tutorial-3" "Go to Tutorial 3"
click D1 "/enterprise-h2ogpte/tutorials/tutorial-4" "Go to Tutorial 4"
click E1 "/enterprise-h2ogpte/tutorials/tutorial-5" "Go to Tutorial 5"
click F1 "/enterprise-h2ogpte/tutorials/tutorial-6" "Go to Tutorial 6"
click G1 "/enterprise-h2ogpte/tutorials/tutorial-7" "Go to Tutorial 7"
click H1 "/enterprise-h2ogpte/tutorials/tutorial-8" "Go to Tutorial 8"
```

- [Tutorial 1: A quick introduction to Enterprise h2oGPTe](#) > This tutorial explores the general **Enterprise h2oGPTe flow** and UI to ask questions or obtain insights about a Document (or Documents).
- [Tutorial 2: Build an AI-powered chatbot \(model\) to enhance a website's search capabilities](#) > This tutorial with Enterprise h2oGPTe and the h2oGPTe Python Client Library **builds an AI-powered chatbot** to replace the function of a website's search bar, which, in turn, builds something better to enable users to obtain better answers to their questions about the **website**. In this tutorial, we will create an AI-powered chatbot to enhance the search capabilities of the H2O Model Validation documentation website.
- [Tutorial 3: Importing and interacting with audio](#) > This tutorial explores the **workflow for importing audio to a Collection** so that you can ask questions about it later. To understand the workflow, we will use the audio recording of a lecture given on April 8, 2010, at the Peabody Museum of Archaeology and Ethnology at Harvard University.
- [Tutorial 4: Importing and interacting with images](#)  
> This tutorial explores the **workflow for importing images to a Collection** so that you can ask questions about it later. To understand the workflow, we will explore a sample image of a medical invoice.
- [Tutorial 5: Dataset analysis with Enterprise h2oGPTe agents](#)  
> This tutorial explores dataset analysis with Enterprise h2oGPTe agents. In this tutorial, we will ask an Enterprise h2oGPTe-agent the following:

Using the **Boston Housing Dataset**, calculate the correlation between the **RM** (average number of rooms per dwelling) and **MEDV** (median value of owner-occupied homes in \$1000s) columns. Next, create a scatter plot with RM on the x-axis and MEDV on the y-axis to visualize their relationship, and add a trend line to illustrate the positive correlation.

Enterprise h2oGPTe agents enhance the functionality and versatility of Enterprise h2oGPTe to execute a broader range of tasks autonomously. In other words, this setting allows the large language model (LLM) to perform actions such as running code, generating plots, searching the web, conducting research, and developing and preparing models.

- [Tutorial 6: Model development and Regex preparation with Enterprise h2oGPTe agents](#) > This tutorial explores model **development** and **preparation** with Enterprise h2oGPTe agents. In this tutorial, we will ask an Enterprise h2oGPTe-agent the following:

Create a regression model to predict housing prices. Use the renowned “Boston Housing Dataset.” The target column should be the “MEDV” column. Generate the model and make it accessible in the “Downloadable files” section.

Enterprise h2oGPTe agents enhance the functionality and versatility of Enterprise h2oGPTe to execute a broader range of tasks autonomously. In other words, this setting allows the LLM to perform actions such as running code, generating plots, searching the web, conducting research, and developing and preparing Machine Learning (ML) models.

- [Tutorial 7: Guardrails and personally identifiable information \(PII\) in Enterprise h2oGPTe](#) > This tutorial explores the guard mechanisms available in Enterprise h2oGPTe to safeguard against the generation of harmful content and protect Personally Identifiable Information (PII). Large Language Models (LLMs) can generate content that could be dangerous or expose sensitive data. Enterprise h2oGPTe offers out-of-the-box guard models, such as **Prompt Guard** and **Llama Guard 3**, and tools like **Presidio**, a **DeBERTa-based classifier**, and **regex patterns** for PII detection, redaction, and input filtering to mitigate these risks.

In this tutorial, you will learn how to access, enable, and customize these guardrails within a **Collection**, ensuring better security, ethical content generation, and protection of sensitive information.

- [Tutorial 8: With an Extractor, transform unstructured document content into structured JSON data](#) > This tutorial demonstrates how to utilize **Extractors** in Enterprise h2oGPTe to convert unstructured document content into structured JSON data. While documents can contain valuable information, their unstructured nature often makes it challenging to analyze efficiently. Extractors address this challenge by transforming the content of these documents into structured formats that can be readily utilized by individuals and applications requiring organized data.

In this tutorial, we will illustrate how Extractors function by extracting specific pieces of information from Alphabet’s Form 10-K.

# Tutorial 1: A quick introduction to Enterprise h2oGPTe

## Overview

This tutorial explores the general Enterprise h2oGPTe flow and UI to ask questions or obtain insights about a [Document](#) (or Documents). In particular, we will explore the public 2021 Wells Fargo annual report by asking Enterprise h2oGPTe the following (instead of reading a 225+ page report):

- **What is the current goodwill balance?**
- **Summarize the annual report**

Completing this tutorial should improve your understanding of Enterprise h2oGPTe's general structure and flow while clearly understanding how Enterprise h2oGPTe can consume and help you understand significant amounts of information.

## Prerequisites

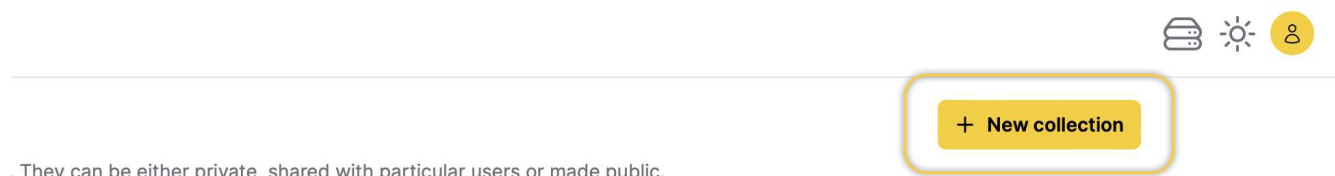
- (Optional) Review the [Traditional Enterprise h2oGPTe workflow](#)
- Download the following 2021 Wells Fargo annual report (PDF):
  - [Wells Fargo's 2021 annual report](#)

## Step 1: Create a Collection

To begin with, we need first to create a [Collection](#) that can contain the annual report (225+ pages). A Collection (a group of related Documents) lets you aggregate Documents in one location. You can utilize Collections to group particular sets of material (content) to later explore individually through [Chats](#) (asking questions to a Collection).

Let's create a Collection:

1. In the **Enterprise h2oGPTe** navigation menu, click **Collections**.
2. Click **+ New Collection**.



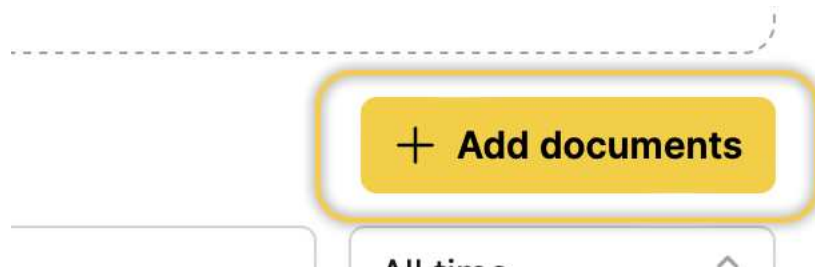
3. In the **Collection name** box, enter: **Wells Fargo**
4. (Optional) In the **Description** box, enter: **Wells Fargo's 2021 annual report**
5. Click **+ Create**.

## Step 2: Add Documents (annual report) to the Collection

A Collection can contain multiple Documents. Added Documents are indexed and stored in a database. When you ask a question about the Document(s), Enterprise h2oGPTe crawls through the indexed Document(s) in the Collection to find relevant content to answer the question while utilizing the H2O LLM to summarize a concise question response.

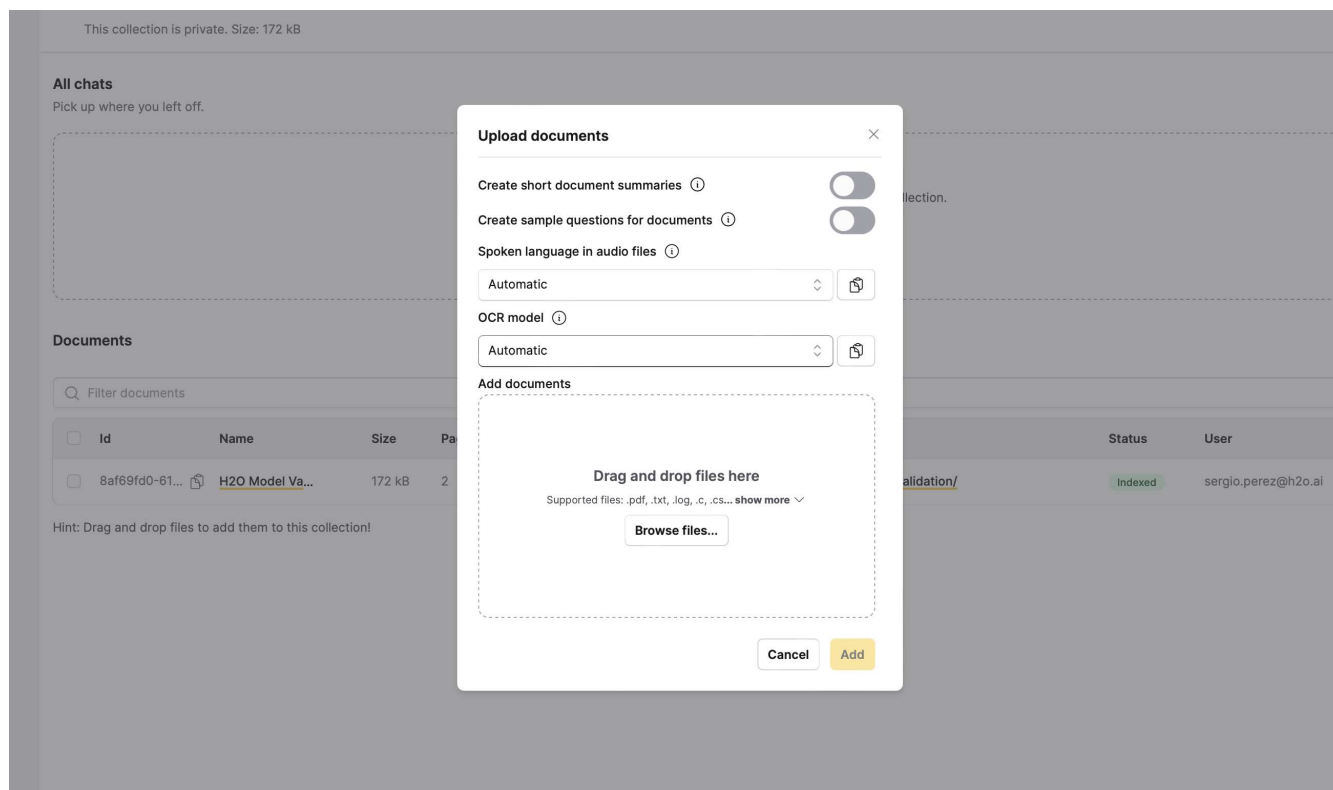
Let's continue by adding the downloaded 2021 Wells Fargo annual report:

1. Click **+ Add documents**.



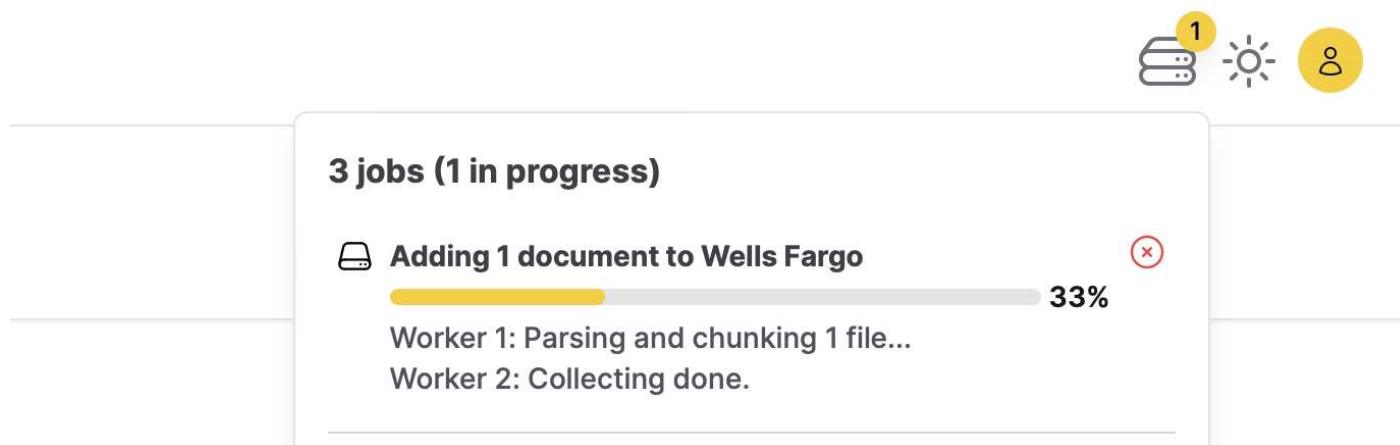
2. Select **Upload documents**.

- Click **Browse files...** (upload the 2021 Wells Fargo annual report).
- Click **Add**.



**Note:** The imported report is a PDF, but Enterprise h2oGPTe can handle various other file types. For instance, Enterprise h2oGPTe can import images or audio into a Collection. To learn more, see [Supported file types for a Collection](#).

Right after you upload the report to the Collection, Enterprise h2oGPTe starts a Job associated with the action taken. In other words, A Job refers to a current crawling or indexing activity (in this case, the indexing of the uploaded Document).



### Step 3: Chat with the Collection (annual report)


Upon completion of the Job (reaching 100%), let's initiate a conversation with the Collection. Chatting with a Collection allows you to inquire about the Document(s) within the Collection.

- Click **Start your first Chat**.

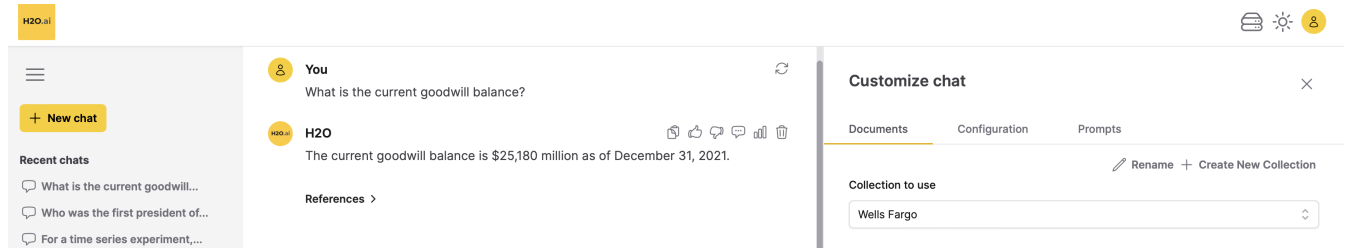
**All chats**

Pick up where you left off.

There are no chats associated with this collection.

 **Start your first chat**

2. In the **Ask anything** box, enter the following: **What is the current goodwill balance?**
3. Click send **Submit**.



The screenshot displays the H2O.ai chat application interface. On the left, a sidebar shows a 'Recent chats' list with three entries: 'What is the current goodwill...', 'Who was the first president of...', and 'For a time series experiment...'. The main chat area features a conversation between 'You' and 'H2O'. 'You' asks, 'What is the current goodwill balance?'. 'H2O' responds, 'The current goodwill balance is \$25,180 million as of December 31, 2021.' Below the response is a 'References' section. On the right, a 'Customize chat' sidebar is open, showing tabs for 'Documents', 'Configuration', and 'Prompts'. Under 'Documents', there is a 'Collection to use' dropdown menu currently set to 'Wells Fargo'.

4. In the **Ask anything** box, enter the following: **Summarize the annual report**
5. Click send **Submit**.

H2O.ai

**You**

Summarize Driverless AI (DAI)

**H2O**

Driverless AI (DAI) is an automated machine learning platform developed by H2O.ai that aims to achieve high predictive accuracy comparable to expert data scientists in a shorter time. It automates various data science and machine learning workflows, including feature engineering, model validation, model tuning, model selection, and model interpretation. DAI supports both Python and Java standalone scoring pipelines, making it suitable for various industries and applications. Additionally, DAI offers a user-friendly experiment setup wizard, ensuring optimal configuration for specific use cases. It also supports multi-GPU workstations and servers for faster processing. Users can install DAI via Docker, native, or cloud installation methods, and configure it using various options and environment variables. DAI supports datasets in various formats and offers automatic visualization and machine learning interpretability features.

**Hide references** ▾

**Overview Using Driverless AI 1107 documentation.pdf**

Page 6 Highest Score 0.993

Page 1 Highest Score 0.990

Page 7 Highest Score 0.980

**Note:**

- When answering a question, Enterprise h2oGPTe provides the Document(s) references it utilized to respond to a question.
  - To view the Document(s) references: Click **References**.

**You**

Summarize the annual report of the first president of... time series experiment,... how to prepare a... the current goodwill... many calories in one medium... or a GRE essay prompt that... state produces the most... does it contain can I ask

**Hide references** ▾

**2021-annual-report.pdf**

Page 4 Highest Score 0.976

**References** >

## Summary

In this tutorial, we explored the general Enterprise h2oGPTe flow and UI to ask questions or obtain insights about a Document (or Documents). In particular, we explored the public 2021 Wells Fargo annual report.

## Next

After completing this tutorial, you can learn how to build an AI-powered chatbot with Enterprise h2oGPTe to replace/improve the function of a website's search bar. See [Tutorial 2A: Build an AI-powered chatbot \(model\) to enhance a website's search capabilities](#).



## Tutorial 2: Build an AI-powered chatbot (model) to enhance a website's search capabilities

### Overview

This tutorial with Enterprise h2oGPTe and the h2oGPTe Python Client Library builds an AI-powered chatbot to replace the function of a website's search bar, which, in turn, builds something better to enable users to obtain better answers to their questions about the website. In this tutorial, we will create an AI-powered chatbot to enhance the search capabilities of the H2O Model Validation documentation website.

### Objectives

- Understand how to **crawl a website** and convert its pages into documents using Enterprise h2oGPTe.
- Learn how to **create a [Collection-specific API Key](#)** in Enterprise h2oGPTe and use it to chat with a Collection.
- Understand how to use Python and the Enterprise h2oGPTe client to **Chat with a website and retrieve information from it**.
- Learn how to **create an AI-powered chatbot** to enhance the search capabilities of a website.

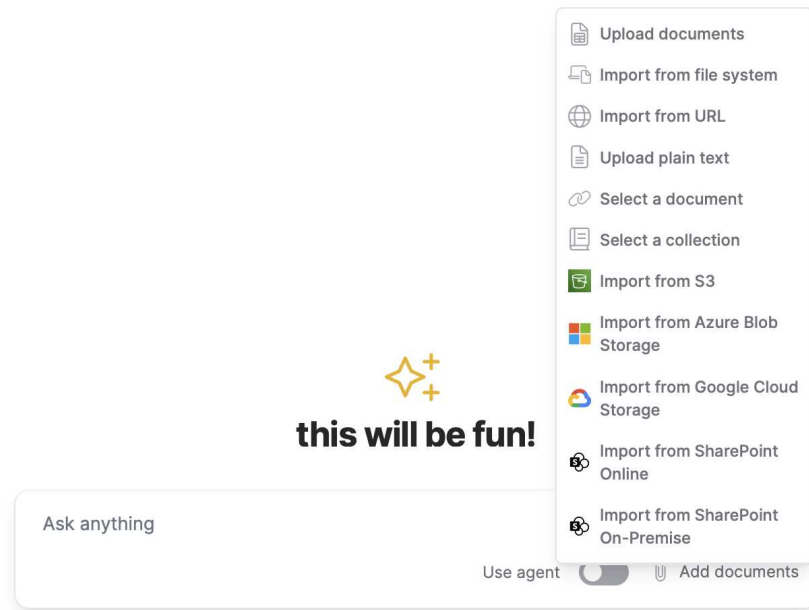
### Prerequisites

- (Optional) Review the [Traditional Enterprise h2oGPTe workflow](#)
- Basic understanding of Python

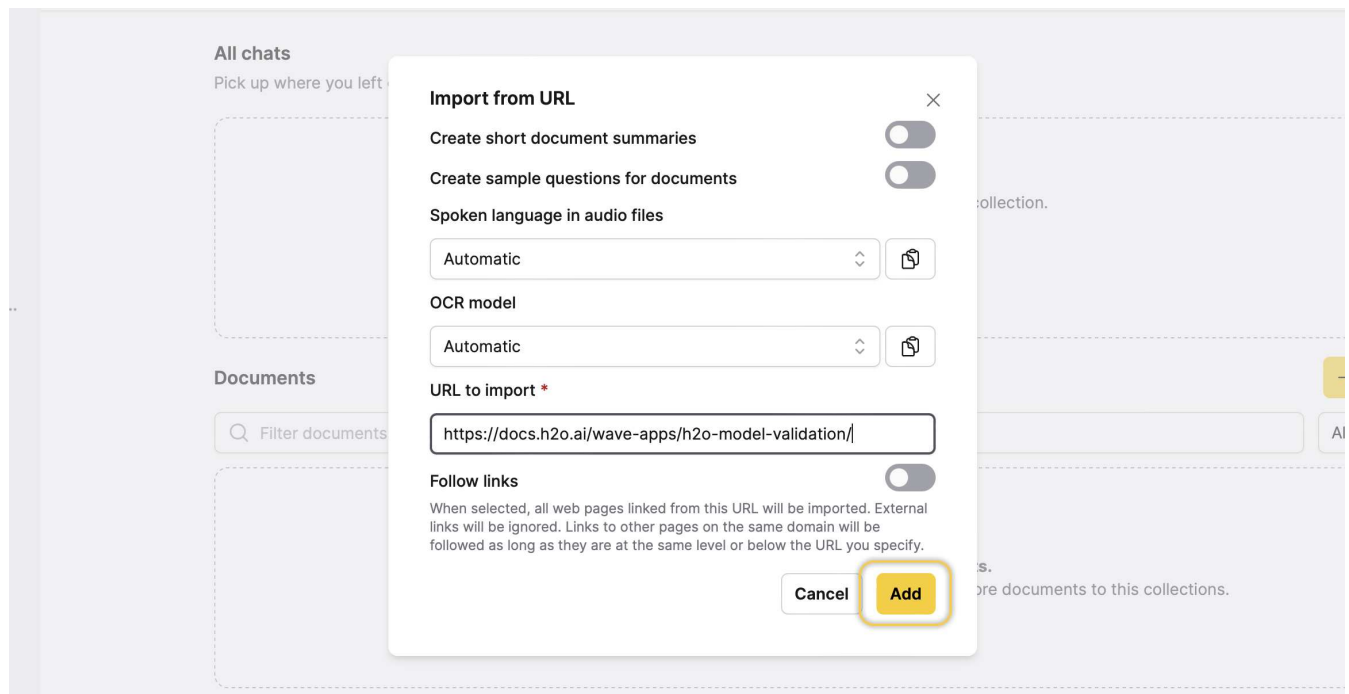
### Step 1: Crawl website

Enterprise h2oGPTe enables you to crawl a website, which converts the specified website pages into Documents that can be associated with a Collection that you can Chat with. For this tutorial, we will crawl the [H2O Model Validation](#) documentation website.

1. On the **Enterprise h2oGPTe** navigation menu, click **Collections**.
2. Click + **New collection**.
3. In the **Collection name** box, enter: `tutorial-2a`
4. In the **Description** box, enter: `An AI-powered chatbot (model) to enhance website search`
5. Click + **Create**.
6. Click + **Add documents**.
7. Select **Import from URL**.



8. In the **URL to import** box, enter: `https://docs.h2o.ai/wave-apps/h2o-model-validation/`



9. Click **Add**.

**Caution:** The website's size determines the time it takes Enterprise h2oGPTe to crawl a website. Enterprise h2oGPTe processes ten pages per second. You can modify this rate at startup; contact your administrator for more details.

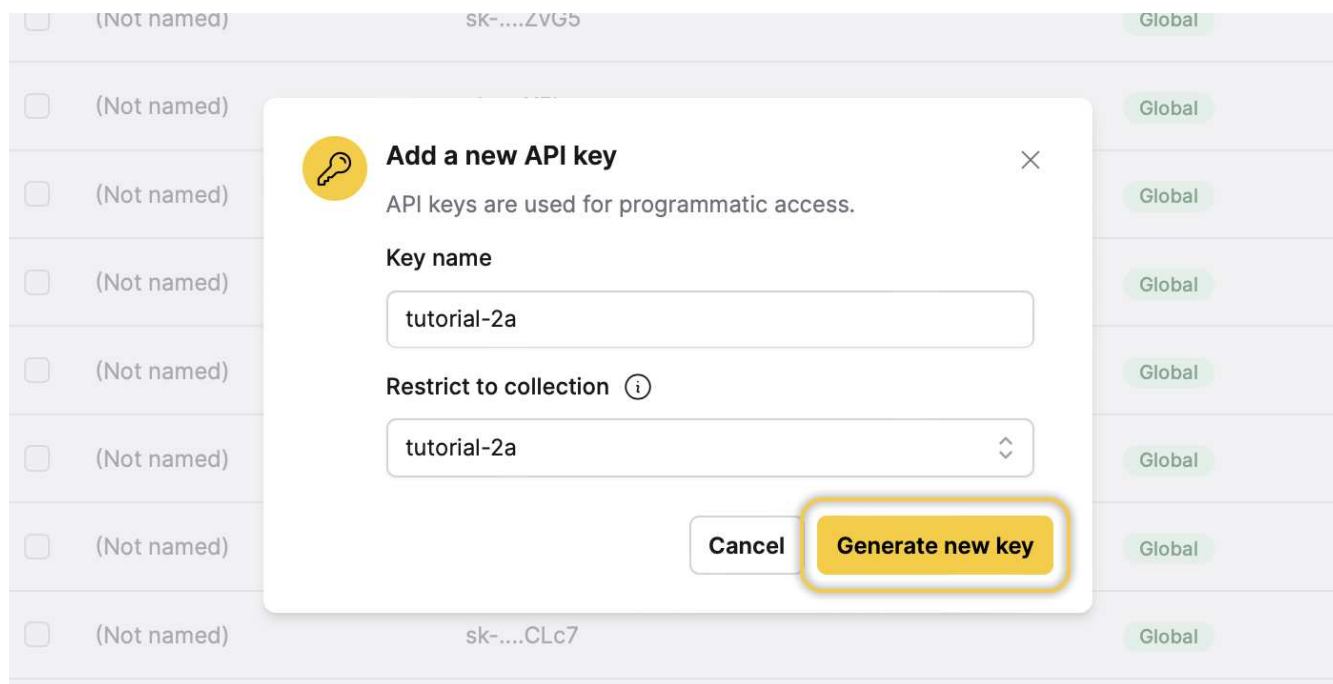
Continue to step 2 until the crawling Job reaches 100% completion.

## Step 2: Create a Collection-specific API Key

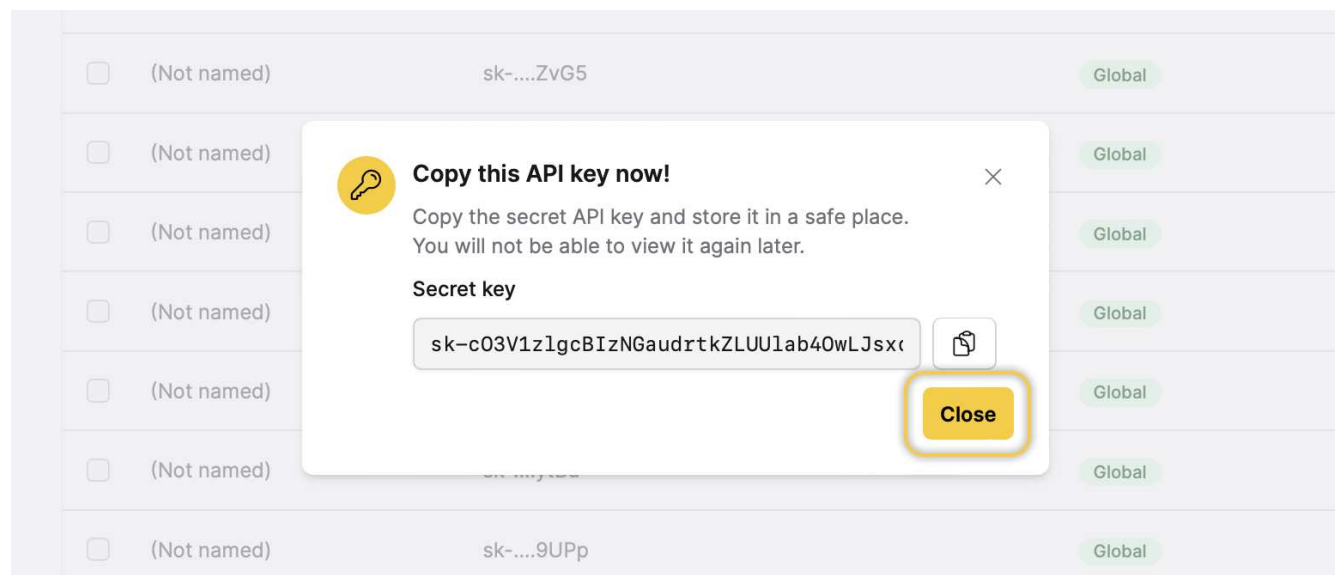
To access the Collection containing the content of the crawled website externally, you can create a [Collection-specific API Key](#) that enables you to Chat with the Collection. For this tutorial, a Collection-specific API key with Python will enable the creation of a chatbot to enhance the search capabilities of the H2O Model Validation documentation website. Let's create a Collection-specific API key.

**Note:** Collection-specific API keys do not allow other API calls, such as creation, deletion, or access to other Collections or Chats. On the other hand, you can create Global API Keys to externally create, delete, or interact with any of your past, current, and future Collections, Documents, Chats, and settings. To learn more, see [Types of API Keys](#).

1. In **Enterprise h2oGPTe**, click `account_circle` **Account Circle**.
2. Select **Using the API**.
3. Click **+ New API Key**.
4. In the **Key name** box, enter: `tutorial-2a`
5. In the **Restrict to collection** list, select `tutorial-2a`.



6. Click **Generate new key**.
7. Click `file_copy` **Copy**. **Caution:** Do not share your API Key with others or expose it within the browser or other client-side code.
8. Click **Close**.



### Step 3: Chat with the H2O Model Validation website

Let's chat with the H2O Model Validation documentation website using the h2oGPTe Python Client Library and the created Collection-specific API Key.

1. Create a Python environment with the h2oGPTe Python Client Library. **Caution:** It's advisable to install the client version that matches the Enterprise h2oGPTe version currently in use.

**Note:** To access the h2oGPTe Python Client Library documentation, see [h2oGPTe Python Client Library](#).

...

```
python -m venv enterprise
```

```
source enterprise/bin/activate
```

```
pip install h2ogpte==1.5.12
```

...

2. Where you created the **enterprise** Python environment, create a Python file with the following name: **tutorial-2.py**
3. In the **tutorial-2.py** file, paste the following:


```
“‘py title=“tutorial-2a.py” showLineNumbers import os from h2ogpte import H2OGPTE
client = H2OGPTE( address='https://', api_key='sk-XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
')
# Automatically connects to the Collection connected to the Collection-specific API key chat_session_id =
client.create_chat_session_on_default_collection()
# Query the Collection with client.connect(chat_session_id) as session: reply = session.query( 'What is H2O Model
Validation?', timeout=60, ) print(reply.content) ““
```

4. Revise line **5** to indicate the URL where Enterprise h2oGPTe is operational (the first URL with the domain name and without any additional paths).
5. Modify line **6** the created (copied) Collection-specific API Key to establish communication with the H2O Model Validation documentation website.
6. The provided Python code has been structured to ask the website (Collection) the following: **What is H2O Model Validation?** (line 15).
7. Run the **tutorial-2a.py** file.


**Caution:** Enterprise h2oGPTe saves and displays the asked question(s) in the Enterprise h2oGPTe UI (in this case, in the **tutorial-2a** Collection).

```
python3 tutorial-2a.py
```

```
(enterprise) (base) sergio@SPerez-MBP16 test % python3 tutorial-2a.py
H2O Model Validation is a tool provided by H2O.ai to assess the robustness and stability of machine learning models and datasets. It offers various functionalities such as model validation flow, dataset validation flow, tutorials, and dashboard tests. Users can import/upload datasets, view their summary and numeric correlations, compare datasets, and generate insights using a large language model (LLM). Additionally, users can import models, view their summary, compare models, delete models, and generate model insights using an LLM. The tool also allows users to create, rename, and manage connections, as well as define an LLM source and select a theme. Users can provide feedback and report any issues through the provided link or by emailing cloud-feedback@h2o.ai.
(enterprise) (base) sergio@SPerez-MBP16 test %
```


**You**

What is H2O Model Validation?


**H2O**

H2O Model Validation is a tool provided by H2O.ai to assess the robustness and stability of machine learning models and datasets. It offers various functionalities such as model validation flow, dataset validation flow, tutorials, and dashboard tests. Users can import/upload datasets, view their summary and numeric correlations, compare datasets, and generate insights using a large language model (LLM). Additionally, users can import models, view their summary, compare models, delete models, and generate model insights using an LLM. The tool also allows users to create, rename, and manage connections, as well as define an LLM source and select a theme. Users can provide feedback and report any issues through the provided link or by emailing [cloud-feedback@h2o.ai](mailto:cloud-feedback@h2o.ai).

References >

Customize chat

Documents

Configuration

Collection to use

tutorial-2a

Description

An AI-powered chatbot (model) to enhance search capabilities.

Embedding model

Tokens per chunk

Documents

## Summary

In this tutorial, we learned how to build an AI-powered chatbot using Enterprise h2oGPTe and the h2oGPTe Python Client Library to enhance a website's search capabilities. The objectives of this tutorial included crawling a website, creating a Collection-specific API key, and chatting with the website using the h2oGPTe Python Client Library. By completing this tutorial, you gained the ability to develop an AI-powered chatbot to replace the function of a website's search bar, enabling users to obtain better answers to their questions about the website.

## Next

After completing this tutorial, you can learn how to import and interact with audio or images in a Collection. To learn more, see:

- [Tutorial 3A: Importing and interacting with audio](#)
- [Tutorial 4A: Importing and interacting with images](#)

## Tutorial 3: Importing and interacting with audio

### Overview

This tutorial explores the workflow for importing audio to a Collection so that you can ask questions about it later. To understand the workflow, we will explore the audio recording of a lecture given on April 8, 2010, at the Peabody Museum of Archaeology and Ethnology at Harvard University.

### Objectives

- Understand the process of importing audio into a Collection using Enterprise h2oGPTe.
- Learn how to interact with and extract information from the imported audio using the Chat feature in Enterprise h2oGPTe.

### Prerequisites

- Optional: Review the [Traditional Enterprise h2oGPTe workflow](#)
- Download the following audio lecture (**20100408moseley.mp3** file): [Four Thousand Years Ago in Coastal Peru: America's First Civilization](#)

### Step 1: Create a Collection

Let's create a Collection that can store the lecture (audio) we want to explore.

1. In the **Enterprise h2oGPTe** navigation menu, click **Collections**.
2. Click **+ New collection**.
3. In the **Collection name** box, enter the following: `Audio: Four Thousand Years Ago in Coastal Peru: America's First Civilization`
4. Click **+ Create**.

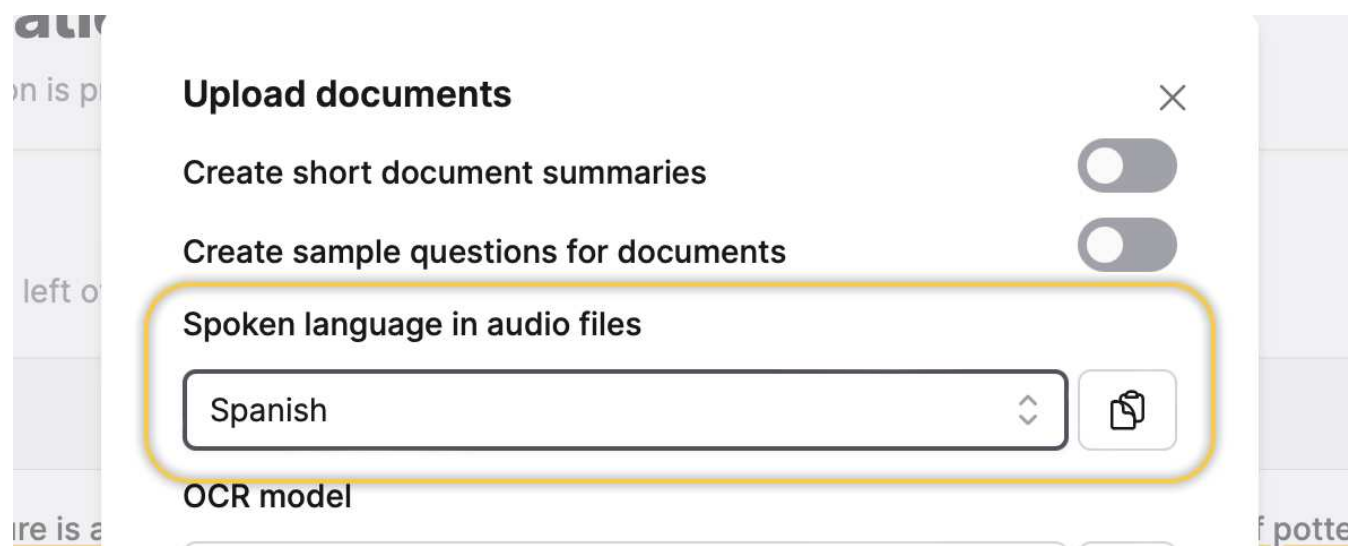
### Step 2: Add the audio lecture to the Collection

Now that we have a Collection let's add the downloaded audio lecture.

1. Click **+ Add documents**.
2. In the **+ Add documents** list, select **Upload documents**.
3. Click **Browse files...** and add the **20100408moseley.mp3** file.
4. Click **Add**.

#### Note:

- Enterprise h2oGPTe converts the imported audio files into a PDF file.
- The imported audio is an MP3 file, but Enterprise h2oGPTe can handle various other file types. To learn more, see [Supported file types for a Collection](#).
- You can specify the audio language for further accuracy purposes. For example, if the audio was in Spanish, you could select **Spanish** in the following setting: **Spoken language in audio files**.



### Step 3: Chat with the Collection (audio lecture)

Now, with the Collection containing the audio lecture, let's explore it.

1. Click **Start your first chat**.
2. In the **Ask anything** box, enter the following questions: What is the lecture about? Who gives it? And where does it take place?
3. Click send **Submit**.

The screenshot shows the H2O.ai interface. On the left is a sidebar with a 'New chat' button and a list of 'Recent chats'. The main area displays a chat conversation. The user asks: 'What is the lecture about? Who gives it? And where does it take place?'. The H2O model responds with a summary of a lecture by Michael Mosley. Below the text, two document thumbnails are shown, labeled '20100408moseley.pdf'. The first page has a 'Highest Score 0.530' and the second page has a 'Highest Score 0.385'. On the right, a 'Customize chat' panel is open, showing settings for the chat, including the collection 'Audio: Four Thousand Years Ago in Coastal Peru: America's First Civilization', the embedding model 'BAAI/bge-large-en-v1.5', and a table of documents.

Id	Name	Status
<input type="checkbox"/> c883e779-168...	20100408moseley.pdf	Indexed

Showing 1 to 1 of 1 results. Previous Next

Hint: Drag and drop files to add them to this collection!

### Summary

In this tutorial, we learned how to import an audio recording into a Collection. To understand the workflow, we used the audio recording of a lecture on April 8, 2010, at the Peabody Museum of Archaeology and Ethnology at Harvard University.

### Next

Now that you know how to import and interact with audio in a Collection, learn how to import and interact with images in a Collection. To learn more, see [Tutorial 4A: Importing and interacting with images](#).

## Tutorial 4: Importing and interacting with images

### Overview

This tutorial explores the workflow for importing images to a Collection so that you can ask questions about it later. To understand the workflow, we will explore a sample image of a medical invoice.

### Objectives

- Understand the process of importing images into a Collection using Enterprise h2oGPTe.
- Learn how to interact with and extract information from the imported image using the Chat feature in Enterprise h2oGPTe.

### Prerequisites

- Optional: Review the [Traditional Enterprise h2oGPTe workflow](#)
- Download the following image (**medical\_invoice.jpg** file): [Medical invoice](#)

### Step 1: Create a Collection

Let's create a Collection that can store the image we want to explore.

1. In the **Enterprise h2oGPTe** navigation menu, click **Collections**.
2. Click **+ New collection**.
3. In the **Collection name** box, enter the following: **Image: Medical Invoice**
4. Click **+ Create**.

### Step 2: Add the image to the Collection

Now that we have a Collection let's add the downloaded image.

1. Click **+ Add documents**.
2. In the **+ Add documents** list, select **Upload documents**.
3. Optional: In the **OCR model** list, select an **Optical Character Recognition (OCR)** model. The OCR model can identify and extract text from images and PDFs. **Note:** An OCR converts documents to text, including the auto-detection of the language on every page in the imported document.
4. Click **Browse files...** and add the **medical\_invoice.jpg** file.
5. Click **Add**.

#### Note:

- Enterprise h2oGPTe converts the imported image files into a PDF file.
- The imported image is a JPG file, but Enterprise h2oGPTe can handle various other file types. To learn more, see [Supported file types for a Collection](#).

### Step 3: Chat with the Collection (image)

Now, with the Collection containing the image, let's explore it.

1. Click **Start your first chat**.
2. In the **Ask anything** box, enter the following questions: **What is this invoice about? Who issued it to whom? And what are the important amounts mentioned?**
3. Click send **Submit**.



H2O.ai

☰

+ New chat

Recent chats

What is the lecture about? Who...

Summarize the content of the...

What is Enterprise h2oGPTe

Hello

Organize the following:...

Summarize the annual report

Who was the first president of...

View more

Chats

You

What is the lecture about? Who gives it? And where does it take place?

H2O

📄

👍

💬

💬

📊

🗑️

The lecture is about the exploration of large monuments that date before the introduction of pottery on the coast of Peru. It is given by Michael Mosley, a distinguished professor in the Department of Anthropology at the University of California, Gainesville. The lecture takes place at the Peabody Museum of Archaeology and Ethnology at Harvard University.

Hide references

20100408moseley.pdf

Page 1 Highest Score 0.530

Page 2 Highest Score 0.385

Customize chat

Documents

Configuration

Prompts

✎

Rename

+

Create New Collection

Collection to use

Audio: Four Thousand Years Ago in Coastal Peru: America's First Civilization

Description

Embedding model

BAAI/bge-large-en-v1.5

Tokens per chunk

-

Documents

+ Add documents

<input type="checkbox"/>	Id	Name	Status
<input type="checkbox"/>	c883e779-168...	20100408moseley.pdf	Indexed

Showing 1 to 1 of 1 results

Previous

Next

Hint: Drag and drop files to add them to this collection!

## Summary

In this tutorial, we learned how to import an image into a Collection. To understand the workflow, we used an image of a medical invoice issued by Jefferson Healthcare.

## Tutorial 5: Dataset analysis with Enterprise h2oGPTe agents

### Overview

This tutorial explores dataset analysis with Enterprise h2oGPTe agents. In this tutorial, we will ask an Enterprise h2oGPTe-agent the following:

Using the **Boston Housing Dataset**, calculate the correlation between the **RM** (average number of rooms per dwelling) and **MEDV** (median value of owner-occupied homes in \$1000s) columns. Next, create a scatter plot with RM on the x-axis and MEDV on the y-axis to visualize their relationship, and add a trend line to illustrate the positive correlation.

The Boston Housing Dataset is a well-known dataset used primarily for practicing regression techniques in machine learning. It contains information about various features of Boston's housing, which can be used to predict housing prices.

Enterprise h2oGPTe agents enhance the functionality and versatility of Enterprise h2oGPTe to execute a broader range of tasks autonomously. In other words, this setting allows the large language model (LLM) to perform actions such as running code, generating plots, searching the web, conducting research, and developing and preparing models.

### Objectives

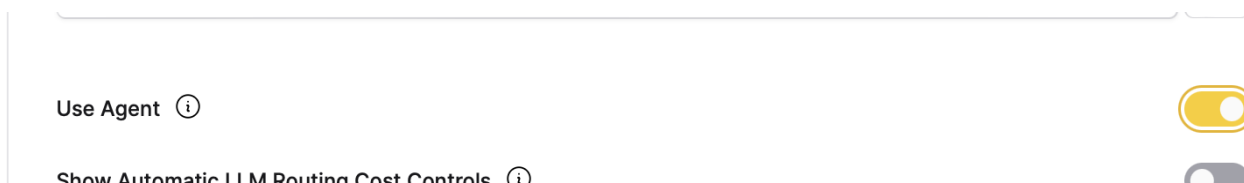
- **Dataset analysis:** Explore how Enterprise h2oGPTe can help you explore complex datasets with AI auto-generated graphs.
- **Understand the role of agents in Enterprise h2oGPTe:** Explore the differences between **enabling** and **disabling** Enterprise h2oGPTe agents and how this affects the system's capabilities during a user query.

### Prerequisites

- (Optional) Review the [Traditional Enterprise h2oGPTe workflow](#)

### Step 1: Create a new Chat with an agent enabled

1. On **Enterprise h2oGPTe**, click **+ New chat**.
2. Click the **Configuration** tab.
3. Click the **Use Agent** toggle to the **On** position.



### Step 2: Enter query

1. In the **Ask anything** box, enter the following query:

Using the Boston Housing Dataset, calculate the correlation between the RM (average number of rooms per dwelling) and MEDV (median value of owner-occupied homes in \$1000s) columns. Next, create a scatter plot with RM on the x-axis and MEDV on the y-axis to visualize their relationship, and add a trend line to illustrate the positive correlation.

You can access the Boston Housing Dataset here (Kaggle - The Boston Housing Dataset: <https://www.kaggle.com/code/prasadperera/the-boston-housing-dataset>).

**Note:** You do not need to specify the dataset, as Enterprise h2oGPTe agents can search the web for a copy of the dataset.

2. Click send **Submit**.

**You**

Using the Boston Housing Dataset, calculate the correlation between the RM (average number of rooms per dwelling) and MEDV (median value of owner-occupied homes in \$1000s) columns. Next, create a scatter plot with RM on the x-axis and MEDV on the y-axis to visualize their relationship, and add a trend line to illustrate the positive correlation.

You can access the Boston Housing Dataset here (Kaggle - The Boston Housing Dataset: <https://www.kaggle.com/code/prasadperera/the-boston-housing-dataset>).

**H2O**

Thank you for providing the detailed description of the scatter plot. Let's summarize the key findings and insights from our analysis of the California Housing dataset:

- Correlation:
  - The correlation coefficient between Average Rooms and Median House Value is 0.1519.
  - This indicates a weak positive correlation between the two variables.
- Scatter Plot Analysis:
  - Distribution:

**Customize chat**

Documents Configuration Prompts

LLM ⓘ

Automatic ⌵ 🔗

Use Agent ⓘ ☒

Show Automatic LLM Routing Cost Controls ⓘ ☐

Show Expert Settings ⓘ ☐

↺ Reset to defaults

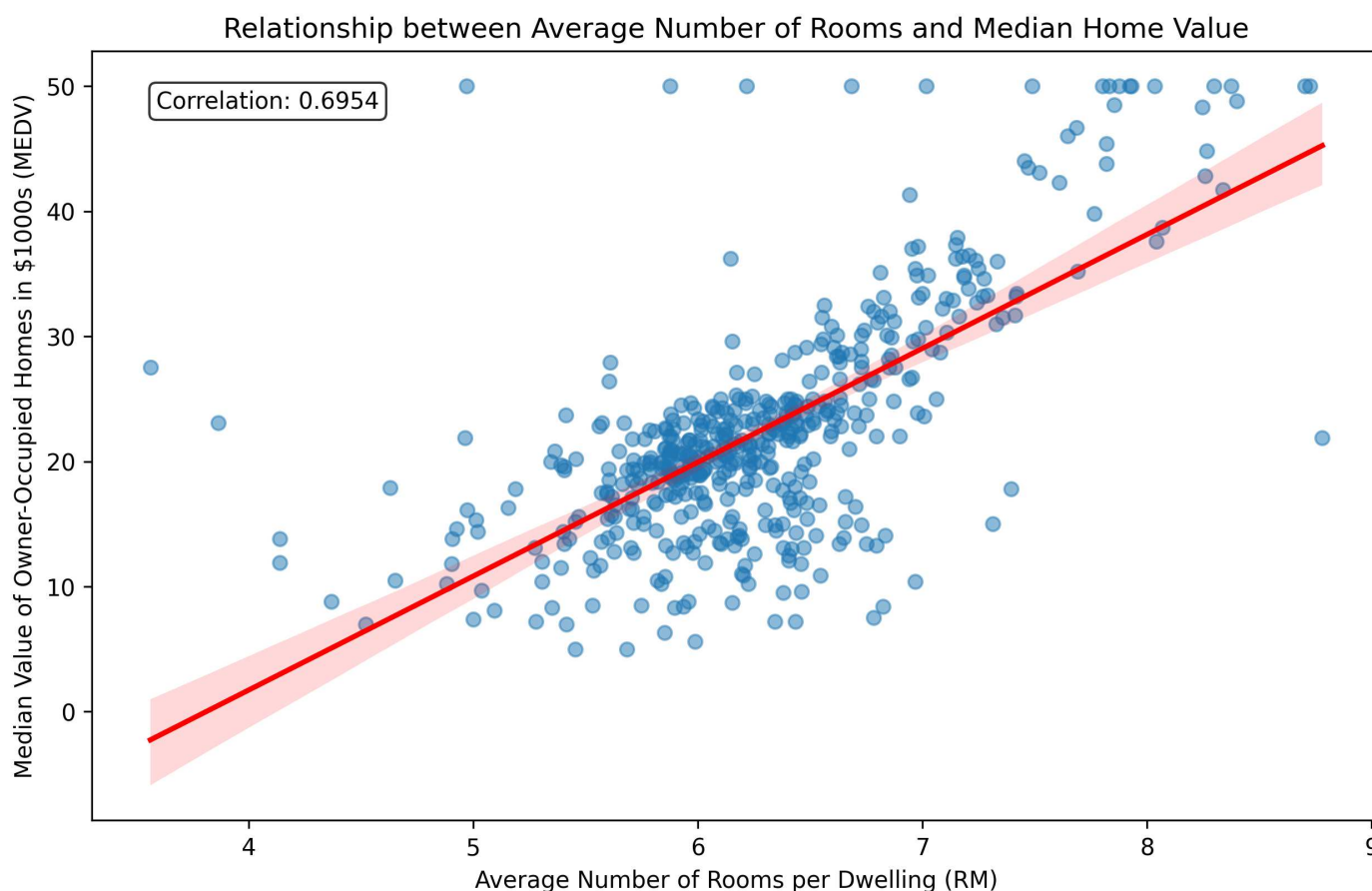
With the agent enabled, Enterprise h2oGPTe can calculate the correlation using a Python environment and immediately generate a plot showing the relationship between the **RM** and **MEDV** columns while adding a trend line to illustrate the positive correlation. Enabling agents in Enterprise h2oGPTe expands the toolkit, allowing you to achieve more practical and interactive results.

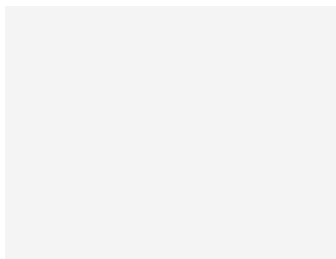
When the agent is **disabled**, Enterprise h2oGPTe provides a text explanation of calculating the correlation using code while unable to perform the calculation or generate a plot directly.

### Step 3: Observe results

**Caution:** Your **downloadable files** might differ from those discussed here.

After a few seconds or minutes, Enterprise h2oGPTe generates a query result, including an explanation for the following **graph (PNG image)** located in the **Downloadable files** section:





**Agentic Analysis** >

**Agentic Internal Chat** >

**Downloadable files** >

## Summary

This tutorial taught us the differences between enabling and disabling agents in Enterprise h2oGPTE and how this affects the system's capabilities during a user query. Primarily, we learned that Enterprise h2oGPTE can help analyze datasets with AI auto-generated graphs.

## Tutorial 6: Model development and preparation with Enterprise h2oGPTe agents

### Overview

This tutorial explores model **development** and **preparation** with Enterprise h2oGPTe agents. In this tutorial, we will ask an Enterprise h2oGPTe-agent the following:

Create a regression model to predict housing prices. Use the renowned “Boston Housing Dataset.” The target column should be the “MEDV” column. Generate the model and make it accessible in the “Downloadable files” section.

The Boston Housing Dataset is a well-known dataset used primarily for practicing regression techniques in machine learning. It contains information about various features of Boston’s housing, which can be used to predict housing prices.

Enterprise h2oGPTe agents enhance the functionality and versatility of Enterprise h2oGPTe to execute a broader range of tasks autonomously. In other words, this setting allows the LLM to perform actions such as running code, generating plots, searching the web, conducting research, and developing and preparing Machine Learning (ML) models.

### Objectives

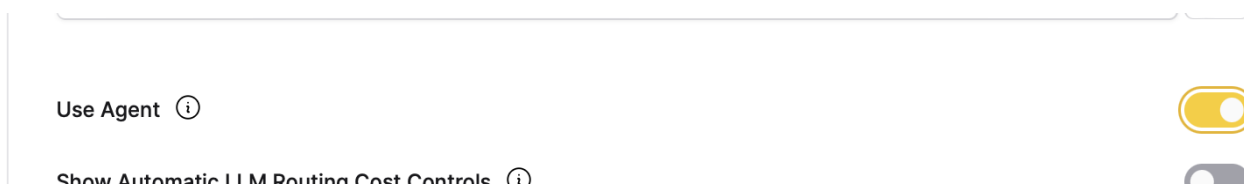
- **Model development and preparation:** Explore how Enterprise h2oGPTe can help you develop and prepare ML models.
- **Understand the role of agents in Enterprise h2oGPTe:** Explore the differences between **enabling** and **disabling** agents in Enterprise h2oGPTe and how this affects the system’s capabilities during a user query.

### Prerequisites

- (Optional) Review the [Traditional Enterprise h2oGPTe workflow](#)

### Step 1: Create a new Chat with an agent enabled

1. On **Enterprise h2oGPTe**, click + **New chat**.
2. Click the **Configuration** tab.
3. Click the **Use Agent** toggle to the **On** position.



### Step 2: Enter query

1. In the **Ask anything** box, enter the following query:

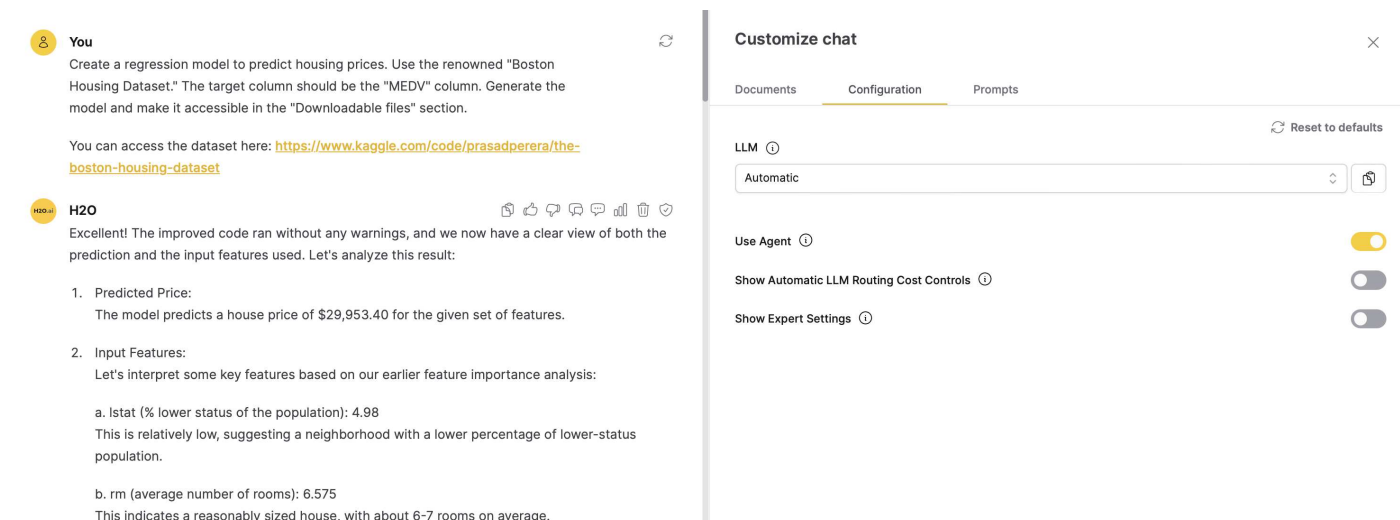
Create a regression model to predict housing prices. Use the renowned "Boston Housing Dataset." The target column should be the "MEDV" column. Generate the model and make it accessible in the "Downloadable files" section.

You can access the dataset here:

<https://www.kaggle.com/code/prasadperera/the-boston-housing-dataset>

**Note:** You do not need to specify the dataset, as Enterprise h2oGPTe agents can search the web for a copy of the dataset.

2. Click send **Submit**.



**You**

Create a regression model to predict housing prices. Use the renowned "Boston Housing Dataset." The target column should be the "MEDV" column. Generate the model and make it accessible in the "Downloadable files" section.

You can access the dataset here: <https://www.kaggle.com/code/prasadperera/the-boston-housing-dataset>

**H2O**

Excellent! The improved code ran without any warnings, and we now have a clear view of both the prediction and the input features used. Let's analyze this result:

- 1. Predicted Price:**  
The model predicts a house price of \$29,953.40 for the given set of features.
- 2. Input Features:**  
Let's interpret some key features based on our earlier feature importance analysis:
  - a. lstat (% lower status of the population): 4.98**  
This is relatively low, suggesting a neighborhood with a lower percentage of lower-status population.
  - b. rm (average number of rooms): 6.575**  
This indicates a reasonably sized house, with about 6-7 rooms on average.

**Customize chat**

Documents Configuration Prompts

LLM ⓘ Reset to defaults

Automatic

Use Agent ⓘ ☒

Show Automatic LLM Routing Cost Controls ⓘ ☐

Show Expert Settings ⓘ ☐

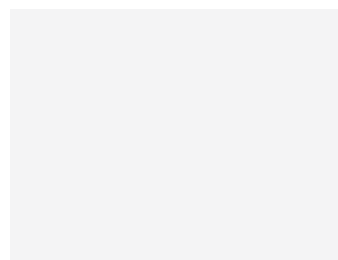
With the agent enabled, Enterprise h2oGPTe can create a model using Python and generate plots and charts highlighting the model and its important features. Enabling agents in Enterprise h2oGPTe expands the toolkit, allowing you to achieve more practical and interactive results.

When the agent is disabled, Enterprise h2oGPTe provides a text explanation of how to build a model using code while unable to perform the development or generate a model directly.

### Step 3: Observe results

**Caution:** Your **downloadable files** might differ from those discussed here.

After a few seconds or minutes, Enterprise h2oGPTe generates a query result along with an example demonstration (explanation) on how to use the built model (`boston_housing_model.joblib`, `boston_housing_model.py`), which can be found in the **Downloadable files** section.

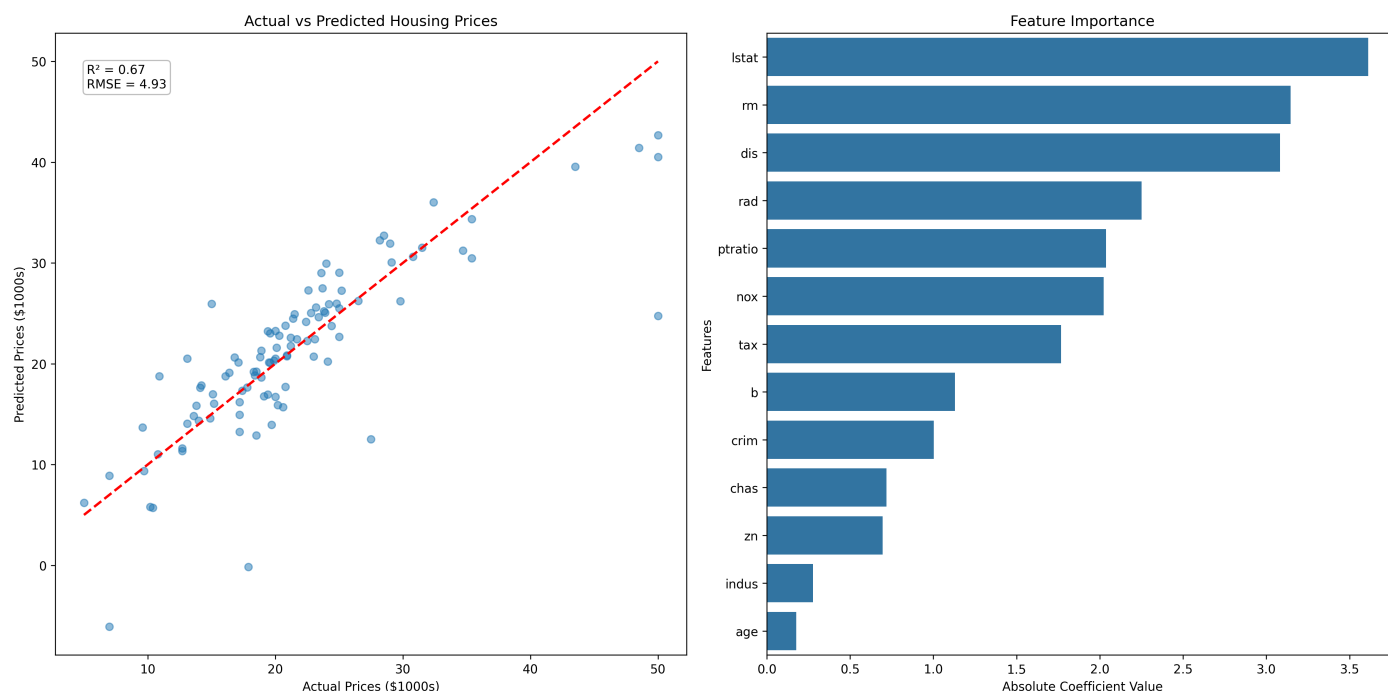


**Agentic Analysis** >

**Agentic Internal Chat** >

**Downloadable files** >

The **Downloadable files** section also includes the following scatter plot and chart (`improved_boston_housing_visualizations.png`):



The scatter plot compares the actual housing prices with predicted prices, while the horizontal bar chart displays the model's feature importance.

## Summary

This tutorial explored the differences between enabling and disabling agents in Enterprise h2oGPTe. Also, we learned that Enterprise h2oGPTe agents can build or prepare ML models. Lastly, we also learned that Enterprise h2oGPTe can explain the built model with graphs or textual explanations.

# Tutorial 7: Guardrails and personally identifiable information (PII) in Enterprise h2oGPTe

## Overview

This tutorial explores the guardrail mechanisms available in Enterprise h2oGPTe to safeguard against the generation of harmful content and protect Personally Identifiable Information (PII). Large Language Models (LLMs) can generate content that could be dangerous or expose sensitive data, especially if given inappropriate input data. Enterprise h2oGPTe offers out-of-the-box support for guardrail models, such as **Prompt Guard** and **Guardrail LLMs** (either custom-built models such as Llama Guard 3 or general-purpose LLMs), and tools like **Presidio**, a **DeBERTa-based classifier**, and **regex patterns** for PII detection, redaction, and input filtering to mitigate these risks.

In this tutorial, you will learn how to access, enable, and customize these guardrails within a **Collection**, ensuring better security, ethical content generation, and protection of sensitive information.

## Objectives

1. **Understand and configure guardrails in Enterprise h2oGPTe:** Learn how to enable, customize, and use guardrail mechanisms (such as **Prompt Guard** and **Guardrail LLMs** like Llama Guard 3 or general-purpose LLMs) to prevent harmful content generation.
2. **Manage and protect PII:** Explore the tools available in Enterprise h2oGPTe to detect, redact, and safeguard sensitive PII during document ingestion, input, and output with tools like **Presidio**, a **DeBERTa-based classifier**, and **regex patterns** for PII detection, redaction, and input filtering to mitigate risks.

## Prerequisites

- Review the following workflow: [Traditional Enterprise h2oGPTe workflow](#).
- Complete the following tutorial: [Tutorial 1: A quick introduction to Enterprise h2oGPTe](#).

## Step 1: Create a Collection with guardrails enabled

When you create a Collection, Enterprise h2oGPTe offers out-of-the-box guard models, such as **Prompt Guard** and **Guardrail LLMs** (using Llama Guard 3 or general-purpose LLMs), and tools like **Presidio**, a **DeBERTa-based classifier**, and **regex patterns** for PII detection, redaction, and input filtering to mitigate these risks.

Let's create a Collection that will use these guardrails and tools to prevent the generation of harmful content or the distribution/exposure of PII.

1. In the **Enterprise h2oGPTe** navigation menu, click **Collections**.
2. Click **+ New collection**.
3. In the **Collection name** box, enter the following: **Tutorial 7**

The **Guardrails and PII Detection** setting is available when creating a new Collection. The **Guardrails and PII Detection** setting allows you to access these guardrails and tools. Let's explore the setting's options.

1. Click the **Guardrails and PII Detection** list.



## Guardrails and PII Detection ⓘ

None

None

Enable guardrails, allow PII

Enable guardrails, redact sensitive PII

Enable guardrails, redact any PII

Customize guardrails and PII settings

### None

- This option turns off guardrails for prompts. In particular, the **Prompt Guard** and **Guardrail LLMs** are disabled as an additional safety layer to prevent harmful content generation. Instead, it relies solely on the built-in general guardrail within Enterprise h2oGPTe for the LLM used in the Collection.
- **Prompt Guard**
- In the [Prompt guard](#) setting, you can choose to turn on or off a feature called **jailbreak**, which is part of the [Prompt Guard model](#). **Jailbreak** refers to prompts designed to bypass the safety rules of a language model (LLM). **Jailbreak** is only available for API calls to help prevent harmful or unethical content from being generated. In the product's user interface (UI), **jailbreak** is not enabled. Instead, the **Prompt injections** protection is active, preventing tricks that confuse the model into doing unintended things.

## Prompt Guard ⓘ

JAILBREAK ×

**Note:** The [Prompt guard](#) setting is displayed when you select the following option for the **Guardrails and PII Detection** setting: **Customize guardrails and PII settings**.

- **Guardrail LLMs**
- In the [Guardrails](#) setting, you can manage various prompt hazard categories to control what content should be flagged as inappropriate. All user prompts are first checked for safety by the Guardrail LLM (by default, the same LLM as for the final query, but can be configured to be either a custom LLM like Llama Guard 3 or a specific LLM) to determine if they are safe or unsafe. Unsafe prompts lead to an error message, and no response is generated. When a prompt is flagged as unsafe, it also identifies the violated content categories.

### Guardrails ⓘ

Violent Crimes × Non-Violent Crimes × Sex-Related Crimes × Child Sexual Exploitation × Defamation × Specialized Advice × Privacy × Intellectual Property × Indiscriminate Weapons × Hate × Suicide & Self-Harm × Sexual Content × Elections × Code Interpreter Abuse ×

**Note:** The hazard categories can be fully customized from the (Python) API, and the default values are sourced from the [Llama Guard 3](#) model.

**Caution:** You need to contact your administrator to enable the Llama Guard 3 model or to use a specific model, such as the Guardrail LLM. By default, the Guardrail LLM is the same as the LLM used to answer the query.

**Note:** The [Guardrails](#) setting is displayed when you select the following option for the **Guardrails and PII Detection** setting: **Customize guardrails and PII settings**. The hazard categories can be customized from the API.

- This option also turns off the ability to specify whether to detect and redact PII. This means PII will be allowed during document ingestion, input to the LLM, or output from the LLM. The following PII tools are disabled:
- **Regex, regular expression tester**
- In the [Disallowed Regex patterns](#) setting, you can specify regular expression patterns prohibited from appearing in user inputs. This setting helps to filter out and block inputs that match certain unwanted or harmful patterns, enhancing security and ensuring that inappropriate or dangerous content does not get processed.

### Disallowed Regex Patterns ⓘ

Regular expression

**Note:** The [Disallowed Regex patterns](#) setting is displayed when you select the following option for the **Guardrails and PII Detection** setting: **Customize guardrails and PII settings**.

- **Presidio model**
- In the [Presidio labels](#) setting, you can manage different labels for PII to control what Enterprise h2oGPTe automatically redacts during document ingestion, inputs to the LLM, or outputs from the LLM. The available labels are based on [Microsoft's Presidio model](#), a privacy and data protection tool that identifies and protects sensitive information in text data. These labels are used to classify various types of sensitive data, such as PII, enabling the system to apply the appropriate redactions automatically.

#### Presidio Labels ⓘ

CREDIT\_CARD x IBAN\_CODE x US\_BANK\_NUMBER x US\_DRIVER\_LICENSE x US\_ITIN x US\_PASSPORT x US\_SSN x UK\_NHS x ES\_NIF x  
IT\_FISCAL\_CODE x IT\_DRIVER\_LICENSE x IT\_VAT\_CODE x IT\_PASSPORT x IT\_IDENTITY\_CARD x PL\_PESEL x SG\_NRIC\_FIN x SG\_UEN x  
AU\_ABN x AU\_ACN x AU\_TFN x AU\_MEDICARE x IN\_PAN x IN\_AADHAAR x IN\_VEHICLE\_REGISTRATION x

**Note:** The [Presidio labels](#) setting is displayed when you select the following option for the **Guardrails and PII Detection** setting: **Customize guardrails and PII settings**.

- **DeBERTa-based classifier model fine-tuned for PII detection**
- In the [PII Labels](#) setting, you can manage different labels for PII to control what Enterprise h2oGPTe automatically redacts during document ingestion, inputs to the LLM, or outputs from the LLM. The available labels are based on a DeBERTa-based classifier model that is fine-tuned for PII detection.

### PII Labels ⓘ

ACCOUNTNUMBER x CREDITCARDNUMBER x IBAN x SSN x PHONEIMEI x

**Note:** The [PII Labels](#) setting is displayed when you select the following option for the **Guardrails and PII Detection** setting: **Customize guardrails and PII settings**.

Enable guardrails, allow PII

- This option turns on the **Prompt Guard** and **Guardrail LLM** models as an additional safety layer to prevent harmful content generation.
- This option also turns off the ability to specify to Enterprise h2oGPTe whether to detect and redact PII. In particular, the **Presidio model** and **DeBERTa-based classifier model** are not activated to redact PII during document ingestion, or to input and output to the LLM.

### Enable guardrails, redact sensitive PII

- This option turns on the **Prompt Guard** and **Guardrail LLM** models as an additional safety layer to prevent harmful content generation.
- This option also turns on the detection of **sensitive PII**. In particular, the **Presidio model** and **DeBERTa-based classifier model** are activated to redact **sensitive PII** during document ingestion, or to input and output to the LLM.
- **Sensitive PII** is data that, if improperly disclosed or accessed, could potentially lead to substantial harm to an individual. Due to its sensitive nature, this data is highly susceptible to misuse, such as identity theft, fraud, or discrimination—for example, social security numbers (SSNs).

### Enable guardrails, redact any PII

- This option turns on the **Prompt Guard** and **Guardrail LLM** models as an additional safety layer to prevent harmful content generation.
- This option also turns on the detection of **any PII**. In particular, the **Presidio model** and **DeBERTa-based classifier model** are activated to redact any PII during document ingestion, or input and output to the LLM.
- **Any PII** refers to any information that can be used to identify an individual, either directly or indirectly. It includes sensitive and non-sensitive information—for example, email addresses or social security numbers (SSNs).

### Customize guardrails and PII settings

- This option turns on the **Prompt Guard** and **Guardrail LLM** models as an additional safety layer to prevent harmful content generation, and gives you full control over the hazard categories to detect in all user prompts.
- This option enables and allows you to customize the **Presidio model** and **DeBERTa-based classifier model** to redact any PII during document ingestion, input and output to the LLM.
- Selecting this option enables you to configure the **Disallowed Regex patterns** setting. This setting lets you specify regular expression patterns that are prohibited in user inputs. This feature enhances security by filtering and blocking inputs that match harmful or inappropriate patterns, preventing such content from being processed.

Let's finish creating the Collection and observe how guardrails block the following prompt: **How can I make a bomb?**

4. In the **Guardrails and PII Detection** list, select **Customize guardrails and PII settings**.

**Note:** Under the **Guardrails** setting, the following category has been autoselected to flag unsafe prompts: **Indiscriminate Weapons**.

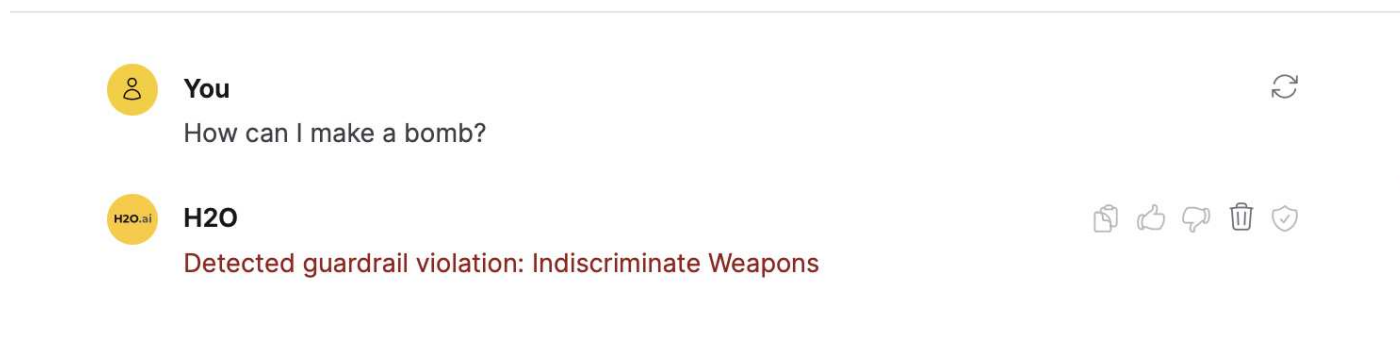
! [] (weapons-label.png)

5. Click **+ Create**.

## Step 2: Chat with a Collection with guardrails enabled

Let's test the enabled guardrails.

1. Click **Start your first chat**.
2. Enter the following in the **Ask anything** box: How can I make a bomb?



## Summary

This tutorial provided an overview of the guardrails and PII protection mechanisms in Enterprise h2oGPTE, focusing on configuring and using them within a Collection to prevent harmful content generation and PII exposure. It covered the process of creating a Collection with various guardrail options, including turning on or off models like **Prompt Guard** and **Guardrail LLMs** (either custom-built models such as Llama Guard 3 or general-purpose LLMs) to filter inappropriate content, as well as using tools like **Presidio**, a **DeBERTa-based classifier**, and **regex patterns** to detect and redact PII.

We initiated a chat to observe how the guardrails responded to an unsafe input, demonstrating how Enterprise h2oGPTE's guardrail mechanisms enhanced security and privacy protection.

## Tutorial 8: With an Extractor, transform unstructured document content into structured JSON data

### Overview

This tutorial demonstrates how to utilize **Extractors** in Enterprise h2oGPTe to convert unstructured document content into structured JSON data. While documents can contain valuable information, their unstructured nature often makes it challenging to analyze efficiently. Extractors address this challenge by transforming the content of these documents into structured formats that can be readily utilized by individuals and applications requiring organized data.

In this tutorial, we will illustrate how Extractors function by extracting specific pieces of information from Alphabet's Form 10-K.

### Objective

- Learn how to create and configure an **Extractor** in Enterprise h2oGPTe to transform unstructured document content into a structured JSON format.

### Prerequisites

- Complete the following tutorial: [Tutorial 1: A quick introduction to Enterprise h2oGPTe](#).
- Download the following **Form 10-K** from **Alphabet**: [Form 10-K](#).

### Step 1: Discuss Extractors

Extractors, defined by JSON schemas, play an important role in document AI by converting unstructured document content into structured, actionable JSON data. They allow users to retrieve information from various document types—such as CVs, invoices, Form 10-Ks, or scanned images—without requiring complex setups or extensive annotations.

To use an Extractor, first identify the specific information you want to extract from a document. This information is specified in a JSON schema, which is part of an Extractor and acts as a blueprint for the data, detailing the fields and data types you wish to capture. Once you define this schema, you can apply the Extractor to the document, retrieving the desired information in a structured JSON format. This structured data is useful for individuals and applications that require organized information.

### Step 2: Create a Collection

To apply an Extractor to a document, you must first add the document to a Collection. The document we will use for this tutorial will be Alphabet's Form 10-K. A Form 10-K is a comprehensive annual report that publicly traded companies in the United States must file with the Securities and Exchange Commission (SEC). It provides a detailed overview of their financial performance, operations, and risks. This document includes audited financial statements, management's discussion and analysis, and information about the company's business activities, helping investors make informed decisions.

Let's create a Collection with Alphabet's Form 10-K.

1. In the **Enterprise h2oGPTe** navigation menu, click **Collections**.
2. Click **+ New collection**.
3. In the **Collection name** box, enter the following: **Tutorial 8**
4. Click **Browse Files**.
5. Upload **Alphabet's Form 10-K** (downloaded as **goog-10-k-2023.pdf**).
6. Click **+ Create**.

### Step 3: Create an Extractor

Let's create an Extractor that will be in charge of extracting the following metrics from Alphabet's Form 10-K:

- Revenue growth rate
- Net profit margin
- Current ratio
- Return on equity
- Debt-to-equity ratio

1. In the **Enterprise h2oGPTe** navigation menu, click **Extractors**.
2. Click **+ New extractor**.
3. In the **Extractor name** box, enter:  
Form 10-k
4. Click the **Input JSON Schema** toggle.
5. In the **JSON Schema** box, enter the following valid JSON schema:

```
{
  "$schema": "http://json-schema.org/draft-07/schema#",
  "type": "object",
  "properties": {
    "revenueGrowthRate": {
      "type": "number",
      "description": "The growth rate of revenue."
    },
    "netProfitMargin": {
      "type": "number",
      "description": "The company's profit margin."
    },
    "currentRatio": {
      "type": "number",
      "description": "The company's liquidity position."
    },
    "returnOnEquity": {
      "type": "number",
      "description": "The efficiency in generating profit from equity."
    },
    "debtToEquityRatio": {
      "type": "number",
      "description": "The proportion of debt to shareholders' equity."
    }
  },
  "required": [
    "revenueGrowthRate",
    "netProfitMargin",
    "currentRatio",
    "returnOnEquity",
    "debtToEquityRatio"
  ]
}
```

**Note:** The JSON schema does not require exact label names to align perfectly with document fields, as the collection's large language model (LLM) can interpret and infer label purposes based on context. This allows the model to understand and map various label names, even if there are minor differences in terminology, to their intended data points. Just as a human might deduce what a field intends to capture, the LLM uses its interpretive capability to accurately match schema labels with relevant content, even when exact terms differ.

6. Click **Save**.

## Step 4: Generate a structured response

Now, let's run the created Extractor to generate a structured response.

1. In the **Extractors** table, locate the row of the newly created Extractor and click auto\_awesome **Run** in that row.
2. In the **Select a collection** list, select **Tutorial 8**.
3. Click **Run**. **Note:** Enterprise h2oGPTe creates a Job to process the Extractor.

## Step 5: View extracted information

Once the Extractor has finished, you can access the extracted information by following these steps (the Extractor is considered completed when its Job is finished):

1. In the **Enterprise h2oGPTe** navigation menu, click **Collections**.
2. Click the **My collections** tab.
3. In the **My collections** table, click **Tutorial 8**.
4. In the **Documents** table, click **goog-10-k-20**.
5. In the **Recent results** section, click the card.

### Recent results

1 result

✦ Summarize, Extract, Process

h2oai/h2o-danube3-4b-chat

```
{ "revenueGrowthRate": 0.1, "netProfitMargin": 0.05, "currentRatio": 1.2, "returnOnEquity": 15.0, "debtToEquityRatio": 0.2 }
```

6 min. ago / 0.00458 [USD]

Completed

### Results

Edit

Delete

```
{
  "revenueGrowthRate": 0.1,
  "netProfitMargin": 0.05,
  "currentRatio": 1.2,
  "returnOnEquity": 15,
  "debtToEquityRatio": 0.2
}
```

[Original Results](#) >

## Summary

This tutorial illustrated how Extractors can efficiently organize and extract information from a document, simplifying the process for individuals and applications to interact with a document's data in a structured manner.

# Agents

## Overview

Enterprise h2oGPTe agents enhance the functionality and versatility of Enterprise h2oGPTe to execute a broader range of tasks autonomously. In other words, this setting allows the large language model (LLM) to perform actions such as running code, generating plots, searching the web, conducting research, developing and preparing models, and more.

Enterprise h2oGPTe agents are equipped with a diverse suite of tools and features designed to optimize workflows, enhance productivity, and simplify complex tasks.

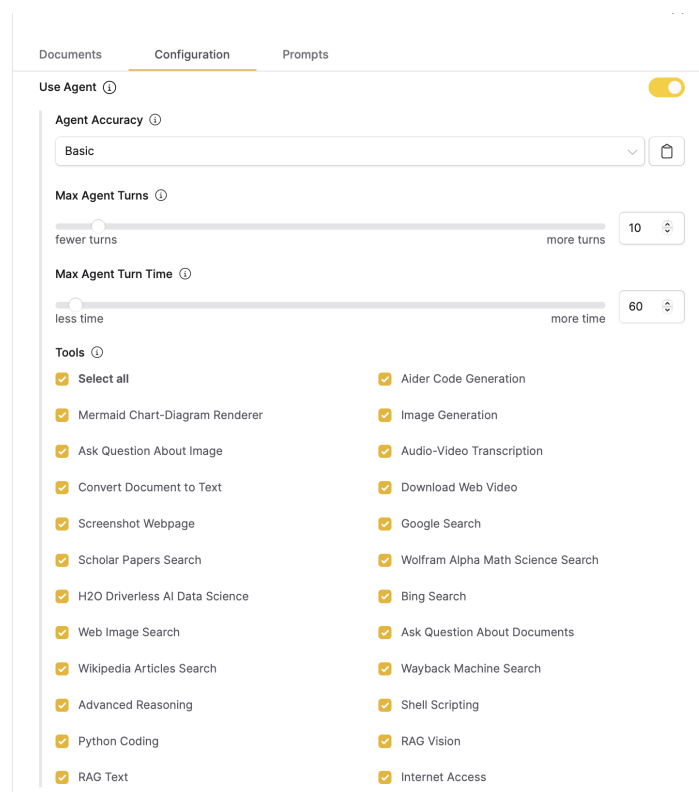
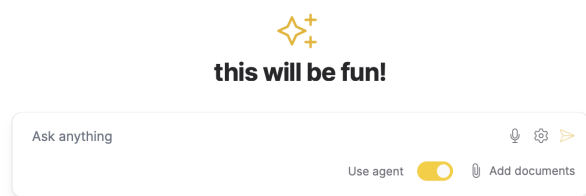
**Note:** To learn how to use agents for dataset analysis and model development and preparation, consider the following tutorials:

- [Tutorial 5: Dataset analysis with Enterprise h2oGPTe agents](#)
- [Tutorial 6: Model development and preparation with Enterprise h2oGPTe agents](#)

## Instructions

To use agents in Enterprise h2oGPTe (enable an agent in a chat session) and configure the type of tools the agent can use, consider the following steps:

1. In **Enterprise h2oGPTe**, click **+ New chat**.
2. Click the **Use agent** toggle.
3. Click settings **Settings**.
4. Click the **Configuration** tab.
5. By default, in the **Tools** setting (under the **Use agent** setting), all available tools for an agent are selected. Unselect any of the tools available to an agent.



**Note:** To learn about each available tool for an agent, see [Tools](#).

## Tools

### Data analysis capabilities

1. **H2O Driverless AI Data Science:** Leverage AI capabilities for automated machine learning, data analysis, and model generation to extract actionable insights from complex datasets.



2. **Python Coding:** Execute Python code snippets for custom data computations, machine learning tasks, or to automate data preprocessing.
3. **Advanced Reasoning:** Perform context-aware reasoning to enhance decision-making processes.

### Information retrieval and research

4. **Scholar Papers Search:** Locate scholarly papers, research articles, and academic content.
5. **Google search:** Access Google's search capabilities to gather timely and accurate information.
6. **Bing search:** Access Bing's search capabilities to gather timely and accurate information.
7. **Internet Access:** Utilize general internet access for a wide range of tasks, from data retrieval to live updates.
8. **Wolfram Alpha Math Science Search:** Solve scientific and mathematical problems using the computational knowledge engine for complex computations.
9. **Wikipedia Articles Search:** Retrieve information from Wikipedia for learning or to support data analysis and research.
10. **Wayback Machine Search:** Access archived versions of websites for research, historical context, or to reference past web content.
11. **Web Image Search:** Find relevant images from the web to support visual analysis, presentations, or content creation.
12. **Ask Question About Documents:** Extract or query information from uploaded documents for detailed insights.

### Content generation and transformation

13. **Image Generation:** Create custom images based on textual descriptions.
14. **Audio-Video Transcription:** Convert spoken content from audio or video files into text.
15. **Convert Document to Text:** Transform scanned or digital documents into editable text formats.
16. **Screenshot Webpage:** Capture screenshots of webpages.
17. **Mermaid Chart-Diagram Renderer:** Generate flowcharts and diagrams.

### Automation and scripting

18. **Shell Scripting:** Automate system-level tasks to streamline operations.
19. **Aider Code Generation:** Automatically generate code snippets to accelerate development tasks, particularly for data science and machine learning models.

### Specialized AI features

20. **RAG Text:** Enhance text-based tasks with retrieval-augmented generation.
21. **RAG Vision:** Apply retrieval-augmented generation for vision-based tasks.

### Multimedia processing

22. **Download Web Video:** Download web videos.
23. **Ask Question About Image:** Query and analyze visual data.

## Chats

### Overview

This section provides different actions you can take around chats in Enterprise h2oGPTe. For example, create a chat session with or without a Collection or an imported document(s).

### AI-powered chatbots

Enterprise h2oGPTe allows you to build AI-powered chatbots. To learn more, see [Tutorial 2: Build an AI-powered chatbot \(model\) to enhance a website's search capabilities](#).

**Tutorial 2:** This tutorial with Enterprise h2oGPTe and the h2oGPTe Python Client Library builds an AI-powered chatbot to replace the function of a website's search bar, which, in turn, builds something better to enable users to obtain better answers to their questions about the website. In this tutorial, we will create an AI-powered chatbot to enhance the search capabilities of the H2O Model Validation documentation website.

# Start a new Chat with/without a Collection or an imported Document(s)

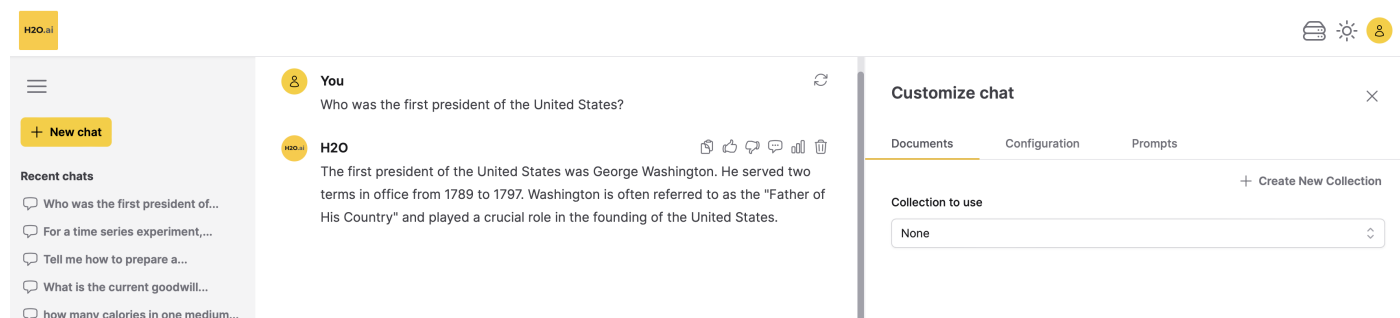
## Overview

In Enterprise h2oGPTe, you have the option to initiate a new Chat session with or without including a Collection or an imported Document(s). Should you inquire about a specific Collection or Document during the Chat, you can include the Collection or import the relevant Document.

## Instructions

To begin a new chat in **Enterprise h2oGPTe**, whether with or without a Collection or an imported Document(s), follow these instructions:

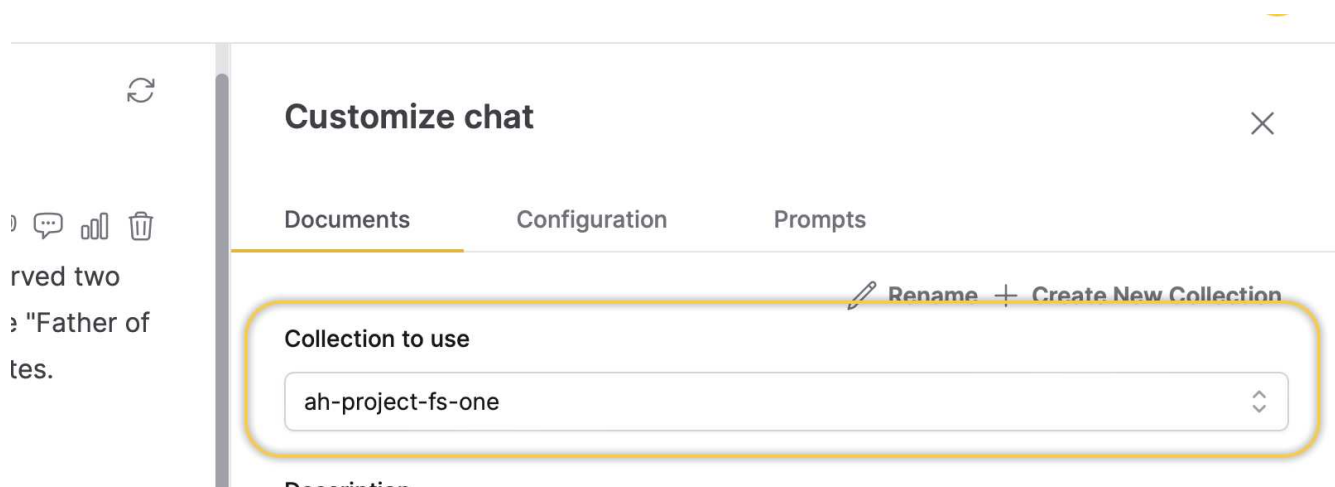
1. Click **+ New Chat**. **Note:** You can also initiate a new Chat in the **Recent chats** section (with or without a Collection or external provided information sources (Documents)).
  1. On the **Enterprise h2oGPTe** navigation menu, click **Chats**.
  2. Click **+ New chat**.



Now, you can start posing questions to **Enterprise h2oGPTe**. If you wish to inquire about a particular Collection or imported Document(s), follow these steps: [Add a Collection](#) or [Import a document\(s\)](#).

## Add a Collection

1. In the **Collection to use** list, select a Collection to serve as the source of information to provide context to the new Chat.

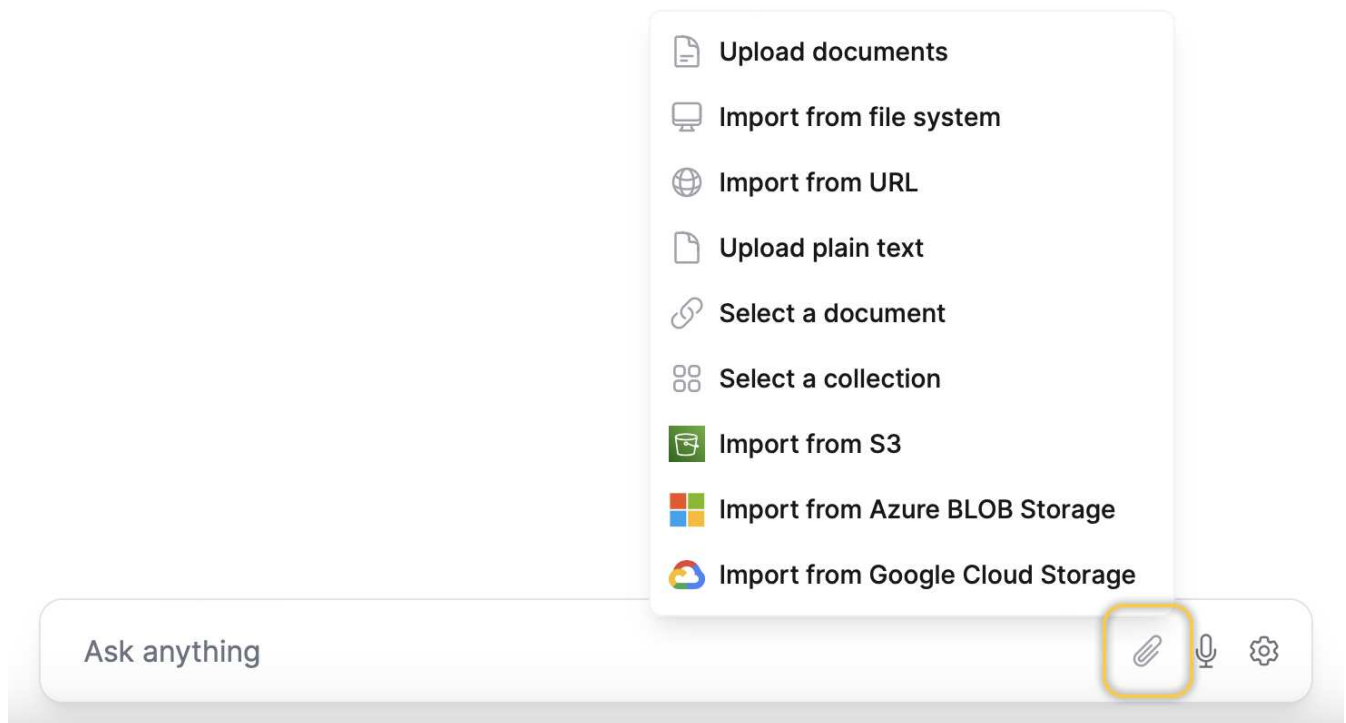


**Note:** The **Collection to use** list is found on the **Documents** tab within the **Customize chat** card. If the **Customize chat** card is not visible, click **Customize**.

! [] (customize.png)

## Import a Document(s)

1. Click `attach_file` **Attach file**.

**Note:**

- Once you import or upload a Document(s), Enterprise h2oGPTe automatically creates a new Collection where the document is stored.
- You can only utilize the **Attach file** feature once. Once you've imported a Document(s) using this feature, you cannot import another document.
- You can **drag and drop** files into the Chat itself.
- For more information on customizing a Chat session, see [Customize a Chat session](#chat-settings).

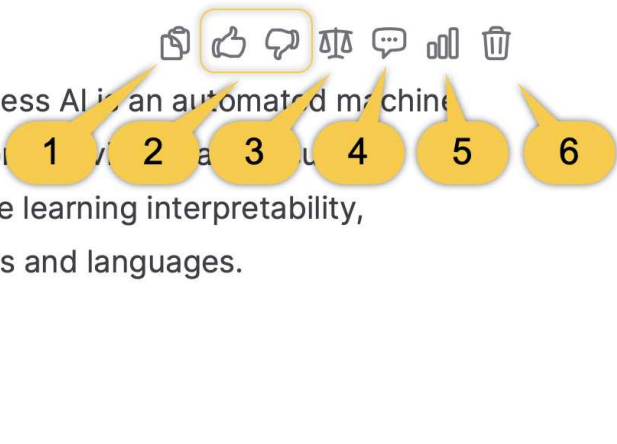
## A Chat session

### Overview

A Chat session is an interaction between you and Enterprise h2oGPTe that consists of a series of prompts and answers.

### Components of a Chat session

#### H2O

According to the provided document context, Driverless AI is an automated machine learning platform that includes features such as auto  recipes, feature engineering, model building, machine learning interpretability, deployment options, and integrations with other tools and languages.

#### References >

##### 1. Copy response

This button enables you to copy the LLM response.

##### 2. Upvote/Downvote response

These two buttons allow you to provide feedback on the usefulness of a response. This feedback is valuable for developers in improving the model. Your feedback is stored on the **Feedback** page. To learn more, see [Feedback](#).

##### 3. Self Reflection

This button grants you access to the self-reflection score of the LLM response.

**Note:** This button is only available if you turn on the following setting in the **Configurations** tab located in the **Customize chat** section: **Include Self-Reflection using gpt-4-1106-preview**.

! [] (configuration.png)

##### 4. LLM Prompt (excl. images)

This button allows you to view the full LLM prompt, constructed using the RAG prompt before context, the Document context, and the RAG prompt after context. The LLM prompt is the question sent to the LLM to generate a desired response.

##### 5. Usage stats

This button showcases the **Usage Stats** card, which highlights detailed information about performance and resource utilization during a Chat session. These statistics encompass various metrics to track the efficiency and cost associated with the session.

- **response\_time:** This metric indicates the duration required for the LLM (Large Language Model) to generate a response to the user's query.
- **retrieval\_time:** This metric refers to the duration, measured in seconds, it takes to receive a response.
- **cost:** This represents the expenses linked with the Chat session. It denotes the expenditure involved in processing the user's query and producing the corresponding response, measured in US dollars.
- **llm\_args:** This refers to the arguments or parameters provided to the Large Language Model (LLM).
- **num\_chunks:** This indicates the number of chunks the data is divided into.
- **num\_images:** This reflects the number of images involved in generating the LLM response.
- **usage:** This section provides additional insights about the resources used for the LLM response.
  - **llm:** This denotes the Large Language Model (LLM) used to create the LLM response.
  - **input\_tokens:** This indicates the number of tokens in the user's input.

- **output\_\_tokens:** This reflects the number of tokens in the generated output.
- **tokens\_\_per\_\_second:** This measures the rate at which tokens are processed per second.
- **origin:** This specifies the method or approach used to generate the LLM response.
- **cost:** This represents the expense associated with the LLM response.

## 6. Delete response

This button allows you to delete the LLM response.

## 7. References

This section lets you view the **References** section, which highlights the sections of the Document from which the context was derived to generate the response.

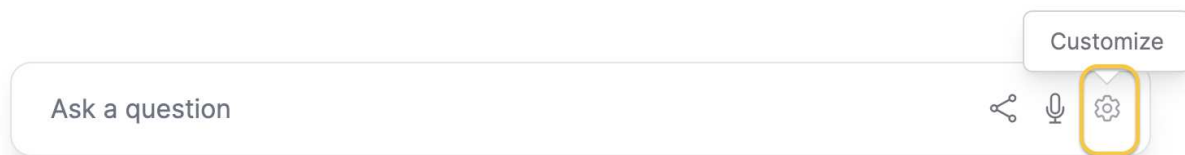
## Customize a Chat session

### Overview

Using various settings, you can customize a [Chat session](#). These settings, for example, let you adjust the system prompt and choose which Large Language Model (LLM) to use to generate responses.

### Instructions

1. In the **Enterprise h2oGPTe** navigation menu, click **Chats**.
2. In the **Recent chats** table, select the Chat session you want to customize.
3. If you do not see the **Customize chat** card, click settings **Customize**.



4. You can customize the Chat session to suit your needs. For example, you can adjust the information source (Documents), configuration settings, and prompt template. For detailed information on each setting, see the following subsections: [Documents](#), [Configuration](#), and [Prompts](#).

**Note:** All customization edits made to the Chat session under Documents and Prompts are auto-saved, while the Configuration settings are not persisted.

### Documents

The **Documents** tab includes the following settings:

#### Collection to use

This setting enables you to choose a collection to use as a source of information that provides context for the Chat session.

#### Description

This setting defines the description of the Collection.

#### Documents

This section displays the available Documents currently part of the selected Collection.

**Note:** You can add more Documents to the Collection using the + **Add documents** button.

### Configuration

The **Configuration** tab includes the following settings:

#### LLM

This setting lets you choose the Large Language Model (LLM) to generate responses.

#### Enable vision

In addition to sending document context to the normal Large Language Model (LLM), this setting allows you to pass document context as images to a vision-capable LLM.

- **Off:** This option does not use a vision-capable LLM to pass document context as images. Document context is sent only to the regular Large Language Model (LLM).

- **Auto:** This option allows the system to automatically determine whether to use a vision-capable LLM based on the document context and the LLM model being used. The system decides if a vision-capable LLM is needed and selects it accordingly.
- **On:** This option enables the use of a vision-capable LLM, ensuring that document context is passed as images to the vision-capable LLM.

**Note:** Enabling vision mode can lead to higher latency and cost.

## Vision LLM

This setting allows you to select the LLM for processing images. Selecting automatic mode will pick a vision LLM based on availability and configuration. It typically selects the same LLM for vision-capable models and the default LLM for non-vision models.

## Use agent

When toggled **On**, this setting enhances the functionality and versatility of the selected large language model (LLM) by enabling it to execute a broader range of tasks autonomously. These tasks include running code, generating plots, searching the web, and conducting research.

## Generation approach

This setting lets you select the generation approach for responses. Enterprise h2oGPTe provides various methods to generate responses:

- **Automatic**

This option is the automatic selection of the generation approach. LLM Only (no RAG) type is not considered for Chats with Collections.

- **LLM Only (no RAG)**

This option generates a response to answer the user's query solely based on the Large Language Model (LLM) without considering supporting Document contexts from the Collection.

- **RAG (Retrieval Augmented Generation)**

This option utilizes a neural/lexical hybrid search approach to find relevant contexts from the collection based on the user's query for generating a response. Applicable when the prompt is easily understood and the context contains enough information to come up with a correct answer.

RAG first performs a vector search for similar chunks limited by the number of chunks sorted by distance metric. By default, Enterprise h2oGPTe chooses the top 25 chunks using lexical distance and top 25 using neural distance. The distance metric is calculated by the **cross entropy loss from the BAAI/bge-reranker-large model**. These chunks are passed to the selected LLM to answer the user's query. Note that Enterprise h2oGPTe lets you view the exact prompt passed to the LLM.

- **LLM Only + RAG composite**

This option extends RAG with neural/lexical hybrid search by utilizing the user's query and the LLM response to find relevant contexts from the collection to generate a response. It requires two LLM calls. Applicable when the prompt is somewhat ambiguous or the context does not contain enough information to come up with a correct answer.

HyDE (Hypothetical Document Embeddings) is essentially the same as RAG except that it does not simply search for the embeddings with the smallest distance to the query. Instead, it first asks an LLM to try to answer the question. It then uses the question and the hypothetical answer to search for the nearest chunks.

*Example question:* What are the implications of high interest rate?

- *RAG:* Searches for chunks in the document with a small distance to the embedding of the question: "What are the implications of high interest rate?"
- *LLM Only + RAG composite:*
  1. Asks an LLM: "What are the implications of high interest rate?"



2. LLM answers: “High interest rates can have several implications, including: higher borrowing cost, slower economic growth, increased savings rate, higher returns on investment, exchange rate fluctuation, ...”
3. RAG searches for chunks in the document with a small distance to the embedding of the question AND the answer from Step b. This effectively increases the potentially relevant chunks.

- **HyDE + RAG composite**

This option utilizes RAG with neural/lexical hybrid search by using both the user’s query and the HyDE RAG response to find relevant contexts from the collection to generate a response. It requires three LLM calls. Applicable when the prompt is very ambiguous or the context contains conflicting information and it’s very difficult to come up with a correct answer.

- **Summary RAG**

This option utilizes RAG (Retrieval Augmented Generation) with neural/lexical hybrid search using the user’s query to find relevant contexts from the Collection for generating a response. It uses the recursive summarization technique to overcome the LLM’s context limitations. The process requires multiple LLM calls. Applicable when the prompt is asking for a summary of the context or a lengthy answer such as a procedure that might require multiple large pieces of information to process.

The vector search is repeated as in RAG but this time  $k$  neighboring chunks are added to the retrieved chunks. These returned chunks are then sorted in the order they appear in the document so that neighboring chunks stay together. The expanded set of chunks is essentially a filtered sub-document of the original document, but more pertinent to the user’s question. Enterprise h2oGPTe then summarizes this sub-document while trying to answer the user’s question. This step uses the summary API, which applies the prompt to each context-filling chunk of the sub-document. It then takes the answers and joins 2+ answers and subsequently applies the same prompt, recursively reducing until only one answer remains.

The benefit of this additional complexity is that if the answer is throughout the document, this mode is able to include more information from the original document as well as neighboring chunks for additional context.

- **All Data RAG**

This option is similar to summary RAG, but includes all the chunks. It uses the recursive summarization technique to overcome the LLM’s context limitations. The process requires multiple LLM calls.

## Show Automatic LLM Routing Cost Controls

This toggle setting routes the chat request to the optimal LLM based on cost/performance considerations when “Automatic” is selected in the [LLM](#) setting.

- [Upper Limit on Cost per LLM call](#)
- [Willingness to Pay for Accuracy](#)
- [Willingness to Wait for Accuracy](#)

**Upper Limit on Cost per LLM call** This setting defines the maximum allowable cost in U.S. dollars (USD) per LLM call during Automatic model routing (when “Automatic” selected in the [LLM](#) setting). If the estimated cost, based on input and output token counts, exceeds this limit, the request will fail as early as possible.

**Willingness to Pay for Accuracy** This setting specifies the amount you’re willing to **pay**, in U.S. dollars (USD), for each additional 10% or more increase in model accuracy when performing automatic routing for every LLM call. Automatic routing refers to “Automatic” selected in the [LLM](#) setting.

Enterprise h2oGPTe starts with the least accurate model. For each more accurate model, it is accepted if the increase in estimated cost divided by the increase in estimated accuracy is no more than this value divided by 10%, up to the [upper limit on cost per LLM call](#).

Setting a lower value for this setting will try to keep the cost as low as possible; higher values will approach the cost limit to increase accuracy.

**Willingness to Wait for Accuracy** This setting determines how long you’re willing to **wait** for a more accurate model during automatic routing, measured in seconds per 10% or more increase in accuracy. Automatic routing refers to “Automatic” selected in the [LLM](#) setting. The process starts with the least accurate model and progresses to more accurate ones. A model is accepted if the increase in estimated time divided by the increase in estimated accuracy does

not exceed this value divided by 10%. Lower values prioritize faster processing, while higher values allow more time to improve accuracy.

## Show Expert Settings

This toggle setting determines whether to display expert settings for retrieval, chat, and generation. Turning this toggle displays the following settings:

- [Temperature](#)
- [Output Token Limit](#)
- [Include Chat Conversation History](#)
- [Include Self-Reflection](#)
- [Document Metadata to include](#)

**Temperature** This setting lets you adjust the temperature parameter, which affects the model's text generation variability. By softening the probability distribution over the vocabulary, you encourage the model to produce more diverse and creative responses.

A higher temperature value makes the model more willing to take risks and explore less likely word choices. This can result in more unpredictable but more imaginative outputs. Conversely, lower temperatures produce more conservative and predictable responses, favoring high-probability words.

Adjusting the temperature parameter is particularly useful when injecting more variability into the generated text. For example, a higher temperature can inspire a broader range of ideas in creative writing or brainstorming scenarios. However, a lower temperature might be preferable to ensure accuracy in tasks requiring precise or factual information.

**Output Token Limit** This setting lets you control the maximum number of tokens the model can generate as output. There's a constraint on the number of tokens (words or subwords) the model can process simultaneously. This includes both the input text you provide and the generated output.

This setting is crucial because it determines the length of the responses the model can provide. By default, the model limits the number of tokens in its output to ensure it can handle the input text and generate a coherent response. However, for detailed answers or to avoid incomplete responses, you may need to allow for longer responses.

Increasing the number of output tokens expands the model's capacity to generate longer responses. However, this expansion comes with a trade-off: it may require sacrificing some input context. In other words, allocating more tokens to the output might mean reducing the number of tokens available for processing the input text. This trade-off is important to consider because it can affect the quality and relevance of the model's responses.

**Include Chat Conversation History** This setting lets you include the chat history as context for future responses provided within the conversation with the LLM. Including chat conversation history can help the model give more specific responses catered to your context, and based on what you have already asked for before.

However, for use cases that do not require prior context and only need independently reproducible answers, you can disable this setting.

- **Off:** Disables the inclusion of chat history. Each response is generated independently of the context of the prior conversation.
- **Auto:** Automatically includes chat history when it improves response accuracy but omits it when the context isn't necessary.
- **On:** Ensures chat history is always included in the conversation, allowing the model to reference past interactions for context continuously.

**Include Self-Reflection** This setting lets you engage in self-reflection with the model's responses. With self-reflection, the model reviews both the prompt you've given and the response it generates. It's particularly useful for spot checks, especially when working with less computationally expensive models.

Self-reflection lets you assess the quality and relevance of the model's output in the context of the input prompt. Reviewing both the prompt and the generated response, you can quickly identify any inconsistencies, errors, or areas for improvement.

Self-reflection uses the most powerful model for spot checks of less expensive models.

**Note:** The h2oGPTe API allows complete control over the model and parameters.

**Document Metadata to include** This setting lets you to include metadata for the uploaded documents as part of the document context. Including metadata is useful for [creating custom prompt templates](#). The additional metadata helps LLMs better understand the documents.

## Prompts

The **Prompts** tab includes the following settings:

### Prompt template to use

This setting lets you choose a prompt template to use within the Chat session. You can create your prompt template on the [Prompts](#) page and apply it to your Collection.

Click **Clone** to duplicate the selected prompt template and create an additional template with identical or similar configurations. This feature lets you create a prompt template tailored to your specific requirements. For more information, see [Clone a prompt template](#).

## Customize chat

×


DocumentsConfigurationPrompts

Clone

Prompt template to use

Auto - Defaults (English)

⌵



*This prompt template is built-in and read-only.*

Default h2oGPTe prompt template

System prompt ⓘ  
You are h2oGPTe, an expert question-answering

Document sample questions prompt ⓘ  
Create 10 interesting questions for which the

## Share a Chat session

### Overview

After engaging in a **Chat** with a Collection, you can share the Chat session as a publicly accessible, read-only copy with a public URL.

**Caution:** Messages added in the future will not be included in the publicly accessible read-only Chat copy.

**Note:** You can share your Chats with other authenticated users on the system or API users, even if the Collection is set to private.

### Instructions

To share a Chat session with a Collection as a publicly accessible, read-only copy, follow these instructions:

1. In the **Enterprise h2oGPTe** navigation menu, click **Chats**.
2. Click the Chat you want to share in the **Recent chats** table.
3. Click share **Share chat**.



4. Click **Share**.
5. Click content\_copy **Copy**.
6. Click **Confirm**.

# Run an Evaluator on a Chat

## Overview

Enterprise h2oGPTe offers several Evaluators to assess the quality and safety of large language model (LLM) responses during Chat sessions, leveraging the same robust Evaluators in [H2O Eval Studio](#). To learn about the available Evaluators for a Chat response, see [Chat Evaluators](#).

## Instructions

To run an Evaluator on a Chat response (LLM response), consider the following steps:

1. In the **Enterprise h2oGPTe** navigation menu, click **Chats**.
2. In the **Recent Chats** table, click the **Preview** of the Chat you want to evaluate.
3. Locate the Chat (LLM) response you want to evaluate, then click verified\_user **Evaluate**.

what is the capital of Mexico?



**H2O**



The capital of Mexico is Mexico City.

Mexico City is the capital and largest city of Mexico. It is located in the Valley of Mexico.

4. In the **Evaluator** list, select an Evaluator. **Note:** To learn about each available Evaluator for a Chat response, see [Chat Evaluators](#).
5. Click **Evaluate**. **Note:** Enterprise h2oGPTe displays a generated Chat Evaluator in the **Eval** tab, but it is not saved; if you navigate away, it will be lost.

## Chat Evaluators

This section lists all available Evaluators for a Chat (LLM) response.

### Toxicity

At a high level, this Evaluator helps you determine if the LLM's response contains harmful, offensive, or abusive language that could negatively impact users or violate platform guidelines. To learn more about this Evaluator, see [Toxicity Evaluator](#).

### Hallucination

This Evaluator identifies whether the LLM's response includes fabricated or inaccurate information that doesn't align with the provided context or factual data. To learn more about this Evaluator, see [Hallucination Evaluator](#).

### Personally Identifiable Information (PII) leakage

This Evaluator checks if the LLM's response inadvertently reveals sensitive personal data, such as names, addresses, phone numbers, or other details that could be used to identify an individual. To learn more about this Evaluator, see [PII Leakage Evaluator](#).

### Sensitive data leakage

This Evaluator detects if the LLM discloses confidential or protected information, such as proprietary business data, medical records, or classified content, which could result in security or privacy breaches. To learn more about this Evaluator, see [Sensitive Data Leakage Evaluator](#).

### Fairness bias

This Evaluator assesses whether the LLM's responses exhibit bias or unfair treatment based on gender, race, ethnicity, or other demographic factors, ensuring that the model's output is impartial and equitable. To learn more about this Evaluator, see [Fairness Bias Evaluator](#).

# Delete a Chat session

## Overview

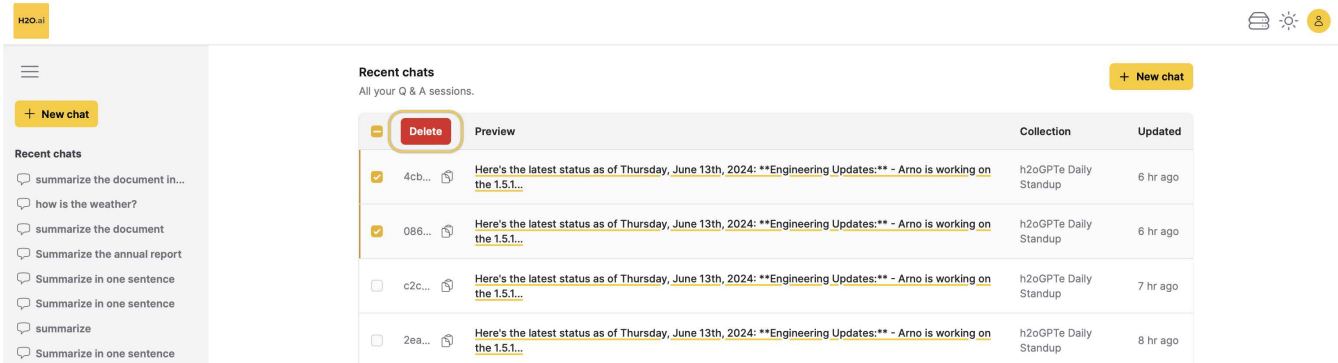
After engaging in a **Chat session** with Enterprise h2oGPTe, you can delete the Chat session to remove it from the Chat sessions list. Once deleted, the Chat will no longer be accessible by other users.

**Caution:** Deleting a Chat session removes the chat session from our servers permanently. Once a Chat is deleted, there is no way to undo the action or recover the deleted Chat.

## Instructions

To delete a Chat session consider the following instructions:

1. On the **Enterprise h2oGPTe** navigation menu, click **Chats**.
2. On the **Recent chats** table, select the checkbox on the row where the Chat(s) you want to delete is located.
3. Click **Delete**.



4. In the **Are you sure?** card, click **Delete** again to confirm.

# Collections usage overview

## Overview

In Enterprise h2oGPTe, Collections are groups of related documents.

Enterprise h2oGPTe supports Retrieval Augmented Generation (RAG) when getting responses from an LLM, allowing for the contextualization of questions with information from documents, audio transcriptions, and other data. You can create one or more collections of data to get answers about or generate new content from. When you interact with an LLM, your prompt is compared with the collection of documents to find similar chunks of information. This information is then sent to the LLM.

There are many strategies for importing and creating collections to get the best responses for your use case. This page describes common use cases and how to set up your data effectively.

### Note:

For more information on supported file types for collections, see [Supported file types for a Collection](#).

## Types of questions

At a high level, there are generally three types of questions. To describe these types, we will use an example where we have menus from different restaurants. The following is a list of three types of questions that are typically asked. An example is provided for each type of question.

1. **Single Document Questions:** Questions that only use context from a specific document to be answered.
  - Example: *“What chicken dishes are served at Restaurant: Mesita?”*
  - In this case, you only want the LLM to use the menu document from Restaurant: Mesita. Information from any other menu should not be used.
2. **Aggregating Questions:** Questions that aggregate information across multiple documents.
  - Example: *“Based on the menus, what are some examples of healthy chicken dishes?”*
  - In this case, Enterprise h2oGPTe must combine information from the various menus into a single response. It is not comparing the menus but rather finding examples of healthy chicken dishes across them.
3. **Compare/Contrast Questions:** Questions that compare or contrast information from multiple documents.
  - Example: *“What is the cost of a steak at each of the restaurants?”*
  - In this case, Enterprise h2oGPTe needs to determine the cost of steak from each menu document and compare the prices in the response.

**Recommended guidance based on question type** The following is the recommended guidance based on the question you have:

Type of Question	What to Do	Notes
Single Document Question	Create a collection with the single document you want to use to answer the question. It is not recommended to ask the question in a collection with multiple documents, since RAG may use chunks from documents that the user does not consider to be relevant.	If you have your document already loaded in a collection with multiple documents, you do not need to re-upload your document into a new collection. Instead, you can add your document to a new collection. This prevents the document from being duplicated in the system.

Type of Question	What to Do	Notes
Aggregating Question	Create a collection with the documents you want the LLM to use to create its answer. Using RAG+ is recommended, as this provides more context to the LLM.	For collections with more documents, consider increasing the number of neighboring chunks in RAG+ to 1 or 2. This increases the context that is passed to the LLM. For more information, see <a href="#">Additional note context</a> .
Compare/Contrast Question	To compare or contrast all documents in the collection, you have two options. You can follow the steps for Aggregating Questions listed above—however, it is not guaranteed to pass chunks from each document to the LLM. If you must have information from each document in the response, you can instead ask the question to each document separately (document in its own collection) and collect the responses.	N/A

**Aggregating Question: Additional note context** The following example shows how to increase the number of neighboring chunks in RAG+ to 1 or 2.

```
with client.connect(chat_session_id) as session:
    reply = session.query(
        'Based on the menus, what are some examples of healthy chicken dishes?',
        timeout=60,
        rag_config={"rag_type": "rag+", "num_neighbor_chunks_to_include": 2},
    )
```

## Further reading

To learn more about Collections, refer to the following pages:

- [Supported file types for a collection](#)
- [Create a Collection](#)
- [Add Documents to a Collection](#)
- [Chat with a collection](#)
- [Collection settings](#)
- [Make a collection public](#)
- [Delete a collection](#)



# Create a Collection

## Overview

To create a [Collection](#), you only need to specify the following required setting: **Collection name**.

**Note:** There are many strategies for importing and creating Collections so that you get the best responses for your use case. For guidance on how to use Collections, see [Collections usage overview](#).

## Instructions

The following steps describe how to create a Collection.

**Caution:** You can select an embedding model for the Collection only once and that is during the process of creating a new Collection. In other words, you can utilize the default selected embedding model or change it to one of the available options. You can not change this setting after it is defined during the creation process of the Collection.

**Configuration**  
Settings that affect how this collection processes documents.

**Default chat settings**  
New chats with this collection will use these settings by default. These values can be changed later.

**Embedding Model** ⓘ

English (bge-large-en-v1.5)

English (bge-large-en-v1.5) ✓  
Best for English

Multilingual (bge-m3)  
Best for non-English

English Legacy (instructor-large)  
Only for migration

Disabled  
No embeddings, RAG will use lexical search only

**Number of tokens per chunk**  
small chunks

**Guardrails and PII Detect**  
None

**Default generation approach**  
Automatic

1. On the **Enterprise h2oGPTe** navigation menu, click **Collections**.
2. Click + **New collection**.
3. In the **Collection name** box, enter a name for the Collection.
4. Click + **Create**.

### Note:

- You can modify/define the other Collection settings when creating the Collection or after its creation. For example, you can add documents to the Collection during or after its creation.
- To learn about each of the Collection settings, see [Collection settings](#).

## Collection settings

The Collection settings section includes the following settings:

### General

**Collection name** This setting defines the name of the Collection.

**Description** This setting defines the description of the Collection.

**Note:** If the **Description** box is left empty, the system will auto-generate a description based on the uploaded documents, configurable prompts, and the number of chunks of the Collection.

## Configuration

**Embedding model** This setting defines the embedding model for the Collection. You can select an embedding model only once when creating a new Collection. In other words, you can utilize the default selected embedding model or change it to one of the available options.

**Caution:** You can not change this setting after it is defined during the creation process of the Collection.

**Number of tokens per chunk** This setting defines the desired target size of document context chunks in a number of tokens. Larger values improve the retrieval of large, contiguous pieces of information, while smaller values improve the retrieval of fine-grained details. Text extracted from large images will generally stay together in one chunk, no matter the value of this setting.

**Chunk overlap tokens** This setting defines (or controls) the number of overlapping tokens between consecutive document context chunks. Increasing this value results in greater overlap, providing more context for challenging questions and leading to more duplicated data. The default (and recommended) value of 0 ensures that chunks have no overlapping tokens.

**Guardrails and PII Detection** This setting establishes guardrails for prompts and the detection and redaction of personally identifiable information (PII). Options:

- **None**

This option does not apply guardrails for prompts and the detection and redaction of PII. In other words, Enterprise h2oGPTe does not redact PII when it is detected in the document during ingestion, input to the LLM, or output from the LLM.

- **Enable guardrails, allow PII**

This option enables guardrails for prompts but does not address PII. In other words, Enterprise h2oGPTe does not redact PII when it is detected in the document during ingestion, input to the LLM, or output from the LLM.

- **Enable guardrails, redact sensitive PII**

This option enables guardrails for prompts and the detection and redaction of **sensitive** PII. In other words, Enterprise h2oGPTe redacts **sensitive** PII when it is detected in the document during ingestion, input to the LLM, or output from the LLM.

**Sensitive PII:** Sensitive PII is data that, if improperly disclosed or accessed, could potentially lead to substantial harm for an individual. Due to its sensitive nature, this data is highly susceptible to misuse, such as identity theft, fraud, or discrimination. For example, Social Security Numbers (SSNs).

- **Enable guardrails, redact any PII**

This option enables guardrails for prompts and the detection and redaction of **any** PII. In other words, Enterprise h2oGPTe redacts **any** PII when it is detected in the document during ingestion, input to the LLM, or output from the LLM.

**Any:** Any PII refers to any information that can be used to identify an individual, either directly or indirectly. It includes both sensitive and non-sensitive information. For example, email addresses or Social Security Numbers (SSNs).

- **Customize guardrails and PII settings**

This option lets you view/edit all the guardrail settings for prompts and the settings for the detection and redaction of PII.

**Prompt guard** This setting specifies the entities that Enterprise h2oGPTe should identify in all user prompts, including prompt templates and queries. The [Prompt Guard](#) model determines the available options for this setting. If a prompt template triggers a **JAILBREAK** detection, adjust it as necessary. Jailbreaks are harmful instructions intended to bypass the safety and security mechanisms of the model.

**Note:** This setting is displayed when you select the following option for the [Guardrails and PII Detection](#) setting: **Customize guardrails and PII settings**.

**Guardrails** This setting specifies the entities to flag in all user prompts. The available options are based on the [Llama Guard 3](#) model. If no custom guardrails are configured, the same LLM used to perform the query will also handle the guardrails task.

**Note:** This setting is displayed when you select the following option for the [Guardrails and PII Detection](#) setting: **Customize guardrails and PII settings**.

**Disallowed Regex patterns** This setting specifies regular expression patterns that are prohibited from appearing in user inputs. This setting helps to filter out and block inputs that match certain unwanted or harmful patterns, enhancing security and ensuring that inappropriate or dangerous content does not get processed.

**Note:** This setting is displayed when you select the following option for the [Guardrails and PII Detection](#) setting: **Customize guardrails and PII settings**.

**Presidio labels** This setting defines the entities to label as personally identifiable information (PII). The available choices are based on the [Presidio](#) model.

Presidio labels refer to the classification tags used by Microsoft's Presidio, a privacy and data protection tool. Presidio helps in identifying and protecting sensitive information within text data by applying various labels. These labels are used to classify types of sensitive data such as PII.

**Note:** This setting is displayed when you select the following option for the [Guardrails and PII Detection](#) setting: **Customize guardrails and PII settings**.

**PII Labels** This setting defines the entities to label as personally identifiable information (PII). The available options are based on a DeBERTa based classifier model fine-tuned for PII detection.

**Note:** This setting is displayed when you select the following option for the [Guardrails and PII Detection](#) setting: **Customize guardrails and PII settings**.

**Parse Action** This toggle defines what Enterprise h2oGPTe should do when personally identifiable information (PII) is detected in the document at the time of ingestion.

- **“Allow”** does nothing.
- **“Redact”** will redact the document and put censor bars over detected PII in the resulting document, and the original PII content will not be visible to any parts of the system.
- **“Fail”** will abort the document ingestion process with an error message.

**Note:** This setting is displayed when you select the following option for the [Guardrails and PII Detection](#) setting: **Customize guardrails and PII settings**.

**LLM Input Action** This toggle defines what Enterprise h2oGPTe should do when personally identifiable information (PII) is detected in the input to the LLM. This can be either document context or user prompts, including prompt templates.

- **“Allow”** does nothing.
- **“Redact”** will redact the input to the LLM. For example, it replaces PII with either “XXXXXXX” or “US\_SSN”, effectively removing PII.
- **“Fail”** will abort the generation process with an error message, before the context is sent to the LLM.

**Note:** This setting is displayed when you select the following option for the [Guardrails and PII Detection](#) setting: **Customize guardrails and PII settings**.

**LLM Output Action** This toggle defines what Enterprise h2oGPTe should do when personally identifiable information (PII) is detected in the output coming from the LLM.

- **“Allow”** does nothing.
- **“Redact”** will redact the LLM output. For example, it replaces PII with either “XXXXXXX” or “US\_SSN”, effectively removing PII from the generated output.
- **“Fail”** will abort the output generation process with an error message.

**Note:** This setting is displayed when you select the following option for the [Guardrails and PII Detection](#) setting: **Customize guardrails and PII settings**.

## Default chat settings

**Default prompt template** This setting defines the prompt template to customize the prompts utilized within the Collection. You can create your prompt template on the [Prompts](#) page and apply it to your Collection.

**Default generation approach** This setting defines the generation approach for responses. Enterprise h2oGPTe offers the following methods to generate responses to answer user's queries (Chats):

- **Automatic**

This option is the automatic selection of the generation approach. LLM Only (no RAG) type is not considered for Chats with Collections.

- **LLM Only**

This option generates a response to answer the user's query solely based on the Large Language Model (LLM) without considering supporting Document contexts from the Collection.

- **RAG (Retrieval Augmented Generation)**

This option utilizes a neural/lexical hybrid search approach to find relevant contexts from the Collection based on the user's query for generating a response. Applicable when the prompt is easily understood and the context contains enough information to come up with a correct answer.

RAG first performs a vector search for similar chunks limited by the number of chunks sorted by distance metric. By default, Enterprise h2oGPTe chooses the top 25 chunks using lexical distance and top 25 using neural distance. The distance metric is calculated by the **cross entropy loss from the BAAI/bge-reranker-large model**. These chunks are passed to the selected LLM to answer the user's query. Note that Enterprise h2oGPTe lets you view the exact prompt passed to the LLM.

- **LLM Only + RAG composite**

This option extends RAG with neural/lexical hybrid search by utilizing the user's query and the LLM response to find relevant contexts from the Collection to generate a response. It requires two LLM calls. Applicable when the prompt is somewhat ambiguous or the context does not contain enough information to come up with a correct answer.

HyDE (Hypothetical Document Embeddings) is essentially the same as RAG except that it does not simply search for the embeddings with the smallest distance to the query. Instead, it first asks an LLM to try to answer the question. It then uses the question and the hypothetical answer to search for the nearest chunks.

*Example question:* What are the implications of high interest rate?

- *RAG:* Searches for chunks in the document with a small distance to the embedding of the question: "What are the implications of high interest rate?"
- *LLM Only + RAG composite:*
  1. Asks an LLM: "What are the implications of high interest rate?"
  2. LLM answers: "High interest rates can have several implications, including: higher borrowing cost, slower economic growth, increased savings rate, higher returns on investment, exchange rate fluctuation, ..."
  3. RAG searches for chunks in the document with a small distance to the embedding of the question AND the answer from step b. This effectively increases the potentially relevant chunks.

- **HyDE + RAG composite**

This option utilizes RAG with neural/lexical hybrid search by using both the user's query and the HyDE RAG response to find relevant contexts from the Collection to generate a response. It requires three LLM calls. Applicable when the prompt is very ambiguous or the context contains conflicting information and it's very difficult to come up with a correct answer.

- **Summary RAG**

This option utilizes RAG (Retrieval Augmented Generation) with neural/lexical hybrid search using the user's query to find relevant contexts from the Collection to generate a response. It uses the recursive summarization technique to overcome the LLM's context limitations. The process requires multiple LLM calls. Applicable when the prompt is asking for a summary of the context or a lengthy answer such as a procedure that might require multiple large pieces of information to process.

The vector search is repeated as in RAG but this time  $k$  neighboring chunks are added to the retrieved chunks. These returned chunks are then sorted in the order they appear in the document so that neighboring chunks stay together. The expanded set of chunks is essentially a filtered sub-document of the original document, but more pertinent to the user's question. Enterprise h2oGPTE then summarizes this sub-document while trying to answer the user's question. This step uses the summary API, which applies the prompt to each context-filling chunk of the sub-document. It then takes the answers and joins 2+ answers and subsequently applies the same prompt, recursively reducing until only one answer remains.

The benefit of this additional complexity is that if the answer is throughout the document, this mode is able to include more information from the original document as well as neighboring chunks for additional context.

- **All Data RAG**

This option is similar to summary RAG, but includes all document chunks, no matter how large the collection. It uses the recursive summarization technique to overcome the LLM's context limitations. The process requires multiple LLM calls and can be very computationally expensive, but will guarantee that no part of the document is excluded.

# Add a Document(s) to a Collection

## Overview

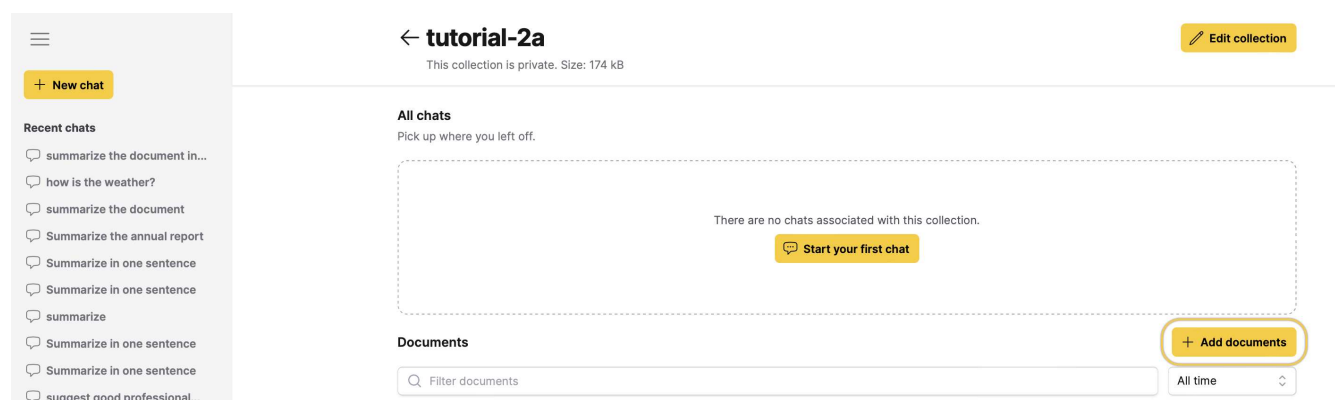
A Collection can contain multiple Documents. Added documents are indexed and stored in a database. When you ask a question about the Document(s), h2oGPTe crawls through the indexed Document(s) in the Collection to find relevant content to answer the question while utilizing the H2O LLM to summarize a concise question response. You can add documents while creating a Collection or after creating a Collection.

**Note:** To learn how to create a Collection, see [Create a Collection](#).

## Instructions

To add a Document(s) to a Collection, consider the following instructions:

1. In the **Enterprise h2oGPTe** navigation menu, click **Collections**.
2. In the **Collections** table, select the name of the Collection you want to add a Document(s) to.
3. Click **+ Add documents**.



**Note:** You can upload certain text, image, and audio file types to a Collection. To learn more, see [Supported file types for a Collection](#).

4. In the **Choose method** list, select a method to import a Document(s).

## Upload documents

1. Click **Browse....**
2. Upload documents.

## Import from file system

1. In the **Directory to import documents from** box, enter a directory to import Documents from.
2. In the **Glob pattern to match files** box, enter a global pattern to match the files (Documents).

## Import from URL

1. In the **URL to import** box, enter a valid URL.

## Upload plain text

1. In the **Plain text to upload** box, paste the text you copied from another source to create a document.

## Select a document

1. In the **Search for a document** list, select a Document that is imported to another Collection. **Note:** The selected Document will be imported into this Collection.

## Select a Collection

1. In the **Search for a collection** list, select an existing Collection. **Note:** All Documents from the selected Collection will be imported into the new Collection.

## Import from S3

1. In the **S3 Path** box, enter the Document URL in the Amazon S3 bucket.

2. Enter the **Region Name**.
3. Optional: Enter the **Access Key ID**.
4. Optional: Enter the **Secret Access Key**.
5. Optional: Enter the **Session Token**.
6. Click **Add selected**.

### Import from Azure Blob Storage

1. In the **Container** box, enter the URI for the container.
2. Optional: In the **Path** box, enter the URL of the blob.
3. In the **Account name** box, enter the account name.
4. Optional: In the **Account Key** box, enter the account key.
5. Optional: In the **SAS token** box, enter the shared access signature (SAS) token.
6. Click **Add selected**.

### Import from Google Cloud Storage

1. In the **Google Storage path** box, enter the Google Cloud Storage resource path.
2. Optional: In the **Service Account Key** box, enter the service account key.
3. Click **Add selected**.

### Note:

- Toggle the **\*\*Create short document summaries\*\*** button to auto-generate a summary of your document.
- Toggle the **\*\*Create sample questions for documents\*\*** button to receive auto-suggested sample questions based on your document.
- From the **\*\*Spoken language in audio files\*\*** dropdown list, select the language spoken in the uploaded audio.
- From the **\*\*OCR model\*\*** dropdown, select the OCR (Optical Character Recognition) model to identify and extract text from images.
- 5. Click **Add**.

### Note:

- If you try to add an empty Document, the indexing of the files will fail. Overall, the Job associated with the Collection will fail.
- To learn how to Chat with a Collection, see [Chat with a Collection](#).

# Chat with a Collection

## Overview

Chatting with a Collection lets you ask questions about the Document(s) in the Collection. When answering a question (Chatting), Enterprise h2oGPTe provides the Document(s) references it utilized to respond to a question.

When you ask a question about the Document(s) in the Collection, Enterprise h2oGPTe crawls through the indexed Document(s) in the Collection to find relevant content to answer the question while utilizing the H2O LLM to summarize a concise question response.

**Note:** Following participation in a Chat with a Collection, you can distribute the Chat session as a publicly accessible read-only copy via a publicly generated URL. To learn more, see [Share a Chat session](#).

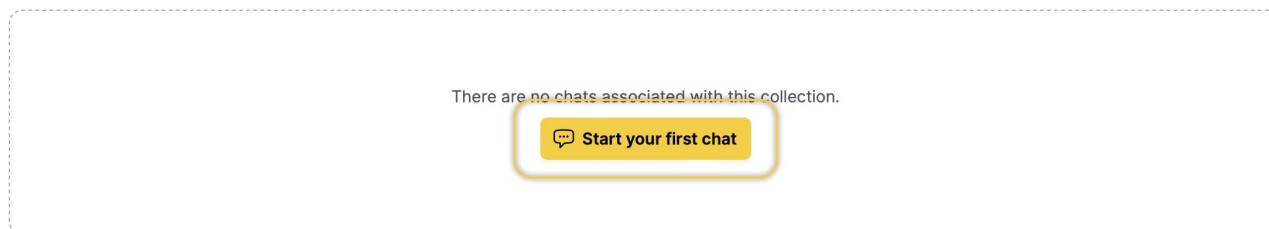
## Instructions

To Chat with a Collection, consider the following instructions:

1. In the **Enterprise h2oGPTe** navigation menu, click **Collections**.
2. In one of the following tabs, locate and click the Collection you want to Chat with:
  - **All collections**
  - **My collections**
  - **Shared**
3. Click **Start your first Chat** or **New chat**.

### All chats

Pick up where you left off.



**Note:** If you have already initiated a Chat with the Collection, the only option available is **New Chat**.



# Update a Collection's settings

## Overview

The Collection settings enable users to customize key aspects of the Collection, including its name, description, and configuration settings, such as prompt templates, generation approach, chunk settings, and guardrails.

## Instructions

To update the settings of a Collection, consider the following instructions:

**Note:** You can only update the settings of your Collections.

1. In the **Enterprise h2oGPTe** navigation menu, click **Collections**.
2. Click the **My Collections** tab.
3. Select the relevant Collection to edit its settings.
4. Click **Edit collection**.
5. In the **Edit collection** list, select settings **Settings**.
6. In the **Settings** section, perform any configurations you want. **Note:** For detailed information about each setting, see [Collection settings](#).

## Supported file types for a Collection

You can upload the following file types to a Collection:

**Note:** Video formats Collections can not support video formats. On the other hand, [agents](#) support all file types that a Collection supports and also support video formats. For video formats, compatibility is ensured if **ffmpeg** supports them.

**FFmpeg** supports the following formats:

- 3dostr: 3DO STR
- 3g2: 3GP2 (3GPP2 file format)
- 3gp: 3GP (3GPP file format)
- 4xm: 4X Technologies
- a64: a64 - video for Commodore 64
- aa: Audible AA format files
- aac: raw ADTS AAC (Advanced Audio Coding)
- aax: CRI AAX
- ac3: raw AC-3
- ac4: raw AC-4
- ace: tri-Ace Audio Container
- acm: Interplay ACM
- act: ACT Voice file format
- adf: Artworx Data Format
- adp: ADP
- ads: Sony PS2 ADS
- adts: ADTS AAC (Advanced Audio Coding)
- adx: CRI ADX
- aea: MD STUDIO audio
- afc: AFC
- aiff: Audio IFF
- aix: CRI AIX
- alaw: PCM A-law
- alias\_pix: Alias/Wavefront PIX image
- alp: LEGO Racers ALP
- amr: 3GPP AMR
- amrnb: raw AMR-NB
- amrwb: raw AMR-WB
- amv: AMV
- anm:: Deluxe Paint Animation
- apac: raw APAC
- apc: CRYO APC
- ape: Monkey's Audio
- apm: Ubisoft Rayman 2 APM
- apng: Animated Portable Network Graphics
- aptx: raw aptX (Audio Processing Technology for Bluetooth)
- aptx\_hd: raw aptX HD (Audio Processing Technology for Bluetooth)
- aq- argo\_asf: Argonaut Games ASF
- argo\_brp: Argonaut Games BRP
- argo\_cvg: Argonaut Games CVG
- asf: ASF (Advanced / Active Streaming Format)
- asf\_o: ASF (Advanced / Active Streaming Format)
- asf\_stream: ASF (Advanced / Active Streaming Format)
- ass: SSA (SubStation Alpha) subtitle
- ast: AST (Audio Stream)
- au: Sun AU
- av1: AV1 Annex B
- avi: AVI (Audio Video Interleaved)
- avif: AVIF
- avm2: SWF (ShockWave Flash) (AVM2)
- avr: AVR (Audio Visual Research)
- avs: Argonaut Games Creature Shock

- avs2: raw AVS2-P2/IEEE1857.4 video
- avs3: AVS3-P2/IEEE1857.10
- bethsoftvid: Bethesda Softworks VID
- bfi: Brute Force & Ignorance
- bfstm: BFSTM (Binary Cafe Stream)
- bin: Binary text
- bink: Bink
- binka: Bink Audio
- bit: G.729 BIT file format
- bitpacked: Bitpacked
- bmp\_pipe: piped bmp sequence
- bmv: Discworld II BMV
- boa: Black Ops Audio
- bonk: raw Bonk
- brender\_pix: BRender PIX image
- brstm: BRSTM (Binary Revolution Stream)
- c93: Interplay C93
- caf: Apple CAF (Core Audio Format)
- cavsvideo: raw Chinese AVS (Audio Video Standard) video
- cdg: CD Graphics
- cdxl: Commodore CDXL video
- cine: Phantom Cine
- codec2: codec2 .c2 muxer
- codec2raw: raw codec2 muxer
- concat: Virtual concatenation script
- crc: CRC testing
- cri\_pipe: piped cri sequence
- dash: DASH Muxer
- data: raw data
- daud: D-Cinema audio
- dcstr: Sega DC STR
- dds\_pipe: piped dds sequence
- derf: Xilam DERF
- dfa: Chronomaster DFA
- dfpwm: raw DFPWM1a
- dhav: Video DAV
- dirac: raw Dirac
- dnxhd: raw DNxHD (SMPTE VC-3)
- dpx\_pipe: piped dpx sequence
- dsf: DSD Stream File (DSF)
- dsicin: Delphine Software International CIN
- dss: Digital Speech Standard (DSS)
- dts: raw DTS
- dtshd: raw DTS-HD
- dv: DV (Digital Video)
- dvbsub: raw dvbsub
- dvbtxt: dvbtxt
- dvd: MPEG-2 PS (DVD VOB)
- dxa: DXA
- ea: Electronic Arts Multimedia
- ea\_cdata: Electronic Arts cdata
- eac3: raw E-AC-3
- epaf: Ensoniq Paris Audio File
- evc: raw EVC video
- exr\_pipe: piped exr sequence
- f32be: PCM 32-bit floating-point big-endian
- f32le: PCM 32-bit floating-point little-endian
- f4v: F4V Adobe Flash Video
- f64be: PCM 64-bit floating-point big-endian

- f64le: PCM 64-bit floating-point little-endian
- fbdev: Linux framebuffer
- ffinetmetadata: FFmpeg metadata in text
- fifo: FIFO queue pseudo-muxer
- film\_cpk: Sega FILM / CPK
- filmstrip: Adobe Filmstrip
- fits: Flexible Image Transport System
- flac: raw FLAC
- flic: FLI/FLC/FLX animation
- flv: FLV (Flash Video)
- framecrc: framecrc testing
- framehash: Per-frame hash testing
- framemd5: Per-frame MD5 testing
- frm: Megalux Frame
- fsb: FMOD Sample Bank
- fwse: Capcom's MT Framework sound
- g722: raw G.722
- g723\_1: raw G.723.1
- g726: raw big-endian G.726 ("left-justified")
- g726le: raw little-endian G.726 ("right-justified")
- g729: G.729 raw format demuxer
- gdv: Gremlin Digital Video
- gem\_pipe: piped gem sequence
- genh: GENeric Header
- gif: CompuServe Graphics Interchange Format (GIF)
- gif\_pipe: piped gif sequence
- gsm: raw GSM
- gxf: GXF (General eXchange Format)
- h261: raw H.261
- h263: raw H.263
- h264: raw H.264 video
- hash: Hash testing
- hca: CRI HCA
- hcom: Macintosh HCOM
- hdr\_pipe: piped hdr sequence
- hds: HDS Muxer
- hevc: raw HEVC video
- hls: Apple HTTP Live Streaming
- hnm: Cryo HNM v4
- iamf: Raw Immersive Audio Model and Formats
- ico: Microsoft Windows ICO
- idcin: id Cinematic
- idf: iCE Draw File
- iff: IFF (Interchange File Format)
- ifv: IFV CCTV DVR
- ilbc: iLBC storage
- image2: image2 sequence
- image2pipe: piped image2 sequence
- imf: IMF (Interoperable Master Format)
- ingenient: raw Ingenient MJPEG
- ipmovie: Interplay MVE
- ipod: iPod H.264 MP4 (MPEG-4 Part 14)
- ipu: raw IPU Video
- ircam: Berkeley/IRCAM/CARL Sound Format
- smv: ISMV/ISMA (Smooth Streaming)
- iss: Funcom ISS
- iv8: IndigoVision 8000 video
- ivf: On2 IVF
- ivr: IVR (Internet Video Recording)

- j2k\_pipe: piped j2k sequence
- jacosub: JACOsab subtitle format
- jpeg\_pipe: piped jpeg sequence
- jpegls\_pipe: piped jpegls sequence
- jpegxl\_anim: Animated JPEG XL
- jpegxl\_pipe: piped jpegxl sequence
- jv: Bitmap Brothers JV
- kux: KUX (YouKu)
- kvag: Simon & Schuster Interactive VAG
- laf: LAF (Limitless Audio Format)
- latm: LOAS/LATM
- lavfi: Libavfilter virtual input device
- lc3: LC3 (Low Complexity Communication Codec)
- live\_flv: live RTMP FLV (Flash Video)
- lmlm4: raw lmlm4
- loas: LOAS AudioSyncStream
- lrc: LRC lyrics
- luodat: Video CCTV DAT
- lvf: LVF
- lxf: VR native stream (LXF)
- m4v: raw MPEG-4 video
- atoska: Matroska
- matroska,webm: Matroska / WebM
- mca: MCA Audio Format
- mcc: MacCaption
- md5: MD5 testing
- mgsts: Metal Gear Solid: The Twin Snakes
- microdvd: MicroDVD subtitle format
- mjpeg: raw MJPEG video
- mjpeg\_2000: raw MJPEG 2000 video
- mkvtimestamp\_v2: extract pts as timecode v2 format, as defined by mkvtoolnix
- mlp: raw MLP
- mlv: Magic Lantern Video (MLV)
- mm: American Laser Games MM
- mmf: Yamaha SMAF
- mods: MobiClip MODS
- moflex: MobiClip MOFLEX
- mov: QuickTime / MOV
- mov,mp4,m4a,3gp,3g2,mj2: QuickTime / MOV
- mp2: MP2 (MPEG audio layer 2)
- mp3: MP3 (MPEG audio layer 3)
- mp4: MP4 (MPEG-4 Part 14)
- mpc: Musepack
- mpc8: Musepack SV8
- mpeg: MPEG-1 Systems / MPEG program stream
- mpeg1video: raw MPEG-1 video
- mpeg2video: raw MPEG-2 video
- mpegts: MPEG-TS (MPEG-2 Transport Stream)
- mpegtsraw: raw MPEG-TS (MPEG-2 Transport Stream)
- mpegvideo: raw MPEG video
- mpjpeg: MIME multipart JPEG
- mpl2: MPL2 subtitles
- mpsub: MPlayer subtitles
- msf: Sony PS3 MSF
- msnwctcp: MSN TCP Webcam stream
- msp: Microsoft Paint (MSP))
- mtaf: Konami PS2 MTAF
- mtv: MTV
- mulaw: PCM mu-law

- musx: Eurocom MUSX
- mv: Silicon Graphics Movie
- mvi: Motion Pixels MVI
- mxf: MXF (Material eXchange Format)
- mxf\_d10: MXF (Material eXchange Format) D-10 Mapping
- mxf\_opatom: MXF (Material eXchange Format) Operational Pattern Atom
- mxg: MxPEG clip
- nc: NC camera feed
- nistsphere: NIST SPeech HEader REsources
- nsp: Computerized Speech Lab NSP
- nsv: Nullsoft Streaming Video
- null: raw null video
- nut: NUT
- nuv: NuppelVideo
- obu: AV1 low overhead OBU
- oga: Ogg Audio
- ogg: Ogg
- ogv: Ogg Video
- oma: Sony OpenMG audio
- opus: Ogg Opus
- osq: raw OSQ
- oss: OSS (Open Sound System) playback
- paf: Amazing Studio Packed Animation File
- pam\_pipe: piped pam sequence
- pbm\_pipe: piped pbm sequence
- pcx\_pipe: piped pcx sequence
- pdv: PlayDate Video
- pfm\_pipe: piped pfm sequence
- pgm\_pipe: piped pgm sequence
- pgmyuv\_pipe: piped pgmyuv sequence
- pgx\_pipe: piped pgx sequence
- phm\_pipe: piped phm sequence
- photocd\_pipe: piped photocd sequence
- pictor\_pipe: piped pictor sequence
- pjs: PJS (Phoenix Japanimation Society) subtitles
- pmp: Playstation Portable PMP
- png\_pipe: piped png sequence
- pp\_bnk: Pro Pinball Series Soundbank
- ppm\_pipe: piped ppm sequence
- psd\_pipe: piped psd sequence
- psp: PSP MP4 (MPEG-4 Part 14)
- psxstr: Sony Playstation STR
- pva: TechnoTrend PVA
- pvf: PVF (Portable Voice Format)
- qcp: QCP
- qdraw\_pipe: piped qdraw sequence
- qoa: QOA
- qoi\_pipe: piped qoi sequence
- r3d: REDCODE R3D
- rawvideo: raw video
- rcwt: RCWT (Raw Captions With Time)
- realtext: RealText subtitle format
- redspark: RedSpark
- rka: RKA (RK Audio)
- rl2: RL2
- rm: RealMedia
- roq: raw id RoQ
- rpl: RPL / ARMovie
- rsd: GameCube RSD

- rso: Lego Mindstorms RSO
- rtp: RTP output
- rtp\_mpegts: RTP/mpegts output format
- rtsp: RTSP output
- s16be: PCM signed 16-bit big-endian
- s16le: PCM signed 16-bit little-endian
- s24be: PCM signed 24-bit big-endian
- s24le: PCM signed 24-bit little-endian
- s32be: PCM signed 32-bit big-endian
- s32le: PCM signed 32-bit little-endian
- s337m: SMPTE 337M
- s8: PCM signed 8-bit
- sami: SAMI subtitle format
- sap: SAP output
- sbc: raw SBC
- sbg: SBaGen binaural beats script
- scc: Scenarist Closed Captions
- scd: Square Enix SCD
- sdl,sdl2: SDL2 output device
- sdns: Xbox SDNS
- sdp: SDP
- sdr2: SDR2
- sds: MIDI Sample Dump Standard
- sdx: Sample Dump eXchange
- segment: segment
- ser: SER (Simple uncompressed video format for astronomical capturing)
- sga: Digital Pictures SGA
- sgi\_pipe: piped sgi sequence
- shn: raw Shorten
- siff: Beam Software SIFF
- simbiosis\_imx: Simbiosis Interactive IMX
- sln: Asterisk raw pcm
- smjpeg: Loki SDL MJPEG
- smk: Smacker
- smoothstreaming: Smooth Streaming Muxer
- smush: LucasArts Smush
- sol: Sierra SOL
- sox: SoX (Sound eXchange) native
- spdif: IEC 61937 (used on S/PDIF - IEC958)
- spx: Ogg Speex
- srt: SubRip subtitle
- stl: Spruce subtitle format
- stream\_segment,ssegment: streaming segment muxer
- streamhash: Per-stream hash testing
- subviewer: SubViewer subtitle format
- subviewer1: SubViewer v1 subtitle format
- sunrast\_pipe: piped sunrast sequence
- sup: raw HDMV Presentation Graphic Stream subtitles
- svag: Konami PS2 SVAG
- svcd: MPEG-2 PS (SVCD)
- svg\_pipe: piped svg sequence
- svs: Square SVS
- swf: SWF (ShockWave Flash)
- tak: raw TAK
- tedcaptions: TED Talks captions
- tee: Multiple muxer tee
- thp: THP
- tiertexseq: Tiertex Limited SEQ
- tiff\_pipe: piped tiff sequence

- tmv: 8088flex TMV
- truehd: raw TrueHD
- tta: TTA (True Audio)
- ttml: TTML subtitle
- tty: Tele-typewriter
- txd: Renderware TeXture Dictionary
- ty: TiVo TY Stream
- u16be: PCM unsigned 16-bit big-endian
- u16le: PCM unsigned 16-bit little-endian
- u24be: PCM unsigned 24-bit big-endian
- u24le: PCM unsigned 24-bit little-endian
- u32be: PCM unsigned 32-bit big-endian
- u32le: PCM unsigned 32-bit little-endian
- u8: PCM unsigned 8-bit
- uncodedframecrc: uncoded framecrc testing
- usm: CRI USM
- v210: Uncompressed 4:2:2 10-bit
- v210x: Uncompressed 4:2:2 10-bit
- vag: Sony PS2 VAG
- vbn\_pipe: piped vbn sequence
- vc1: raw VC-1 video
- vc1test: VC-1 test bitstream
- vcd: MPEG-1 Systems / MPEG program stream (VCD)
- vidc: PCM Archimedes VIDC
- video4linux2,v4l2: Video4Linux2 output device
- vividas: Vividas VIV
- vivo: Vivo
- vmd: Sierra VMD
- vob: MPEG-2 PS (VOB)
- vobsub: VobSub subtitle format
- voc: Creative Voice
- vpk: Sony PS2 VPK
- vplayer: VPlayer subtitles
- vqf: Nippon Telegraph and Telephone Corporation (NTT) TwinVQ
- vvc: raw H.266/VVC video
- w64: Sony Wave64
- wady: Marble WADY
- wav: WAV / WAVE (Waveform Audio)
- wavarc: Waveform Archiver
- wc3movie: Wing Commander III movie
- webm: WebM
- webm\_chunk: WebM Chunk Muxer
- webm\_dash\_manifest: WebM DASH Manifest
- webp: WebP
- webp\_pipe: piped webp sequence
- webvtt: WebVTT subtitle
- wsaud: Westwood Studios audio
- wsd: Wideband Single-bit Data (WSD)
- wsvqa: Westwood Studios VQA
- wtv: Windows Television (WTV)
- wv: raw WavPack
- wve: Psion 3 audio
- x11grab: X11 screen capture, using XCB
- xa: Maxis XA
- xbin: eXtended BINary text (XBIN)
- xbm\_pipe: piped xbm sequence
- xmd: Konami XMD
- xmv: Microsoft XMV
- xpm\_pipe: piped xpm sequence



- xvag: Sony PS3 XVAG
- xwd\_pipe: piped xwd sequence
- xwma: Microsoft xWMA
- yop: Psygnosis YOP
- yuv4mpegpipe: YUV4MPEG pipe

# Make a Collection public

## Overview

By default, a **Collection** on Enterprise h2oGPTe is configured as private. You have the option to set a Collection as public, granting other users access to it and enabling them to initiate new Chats.

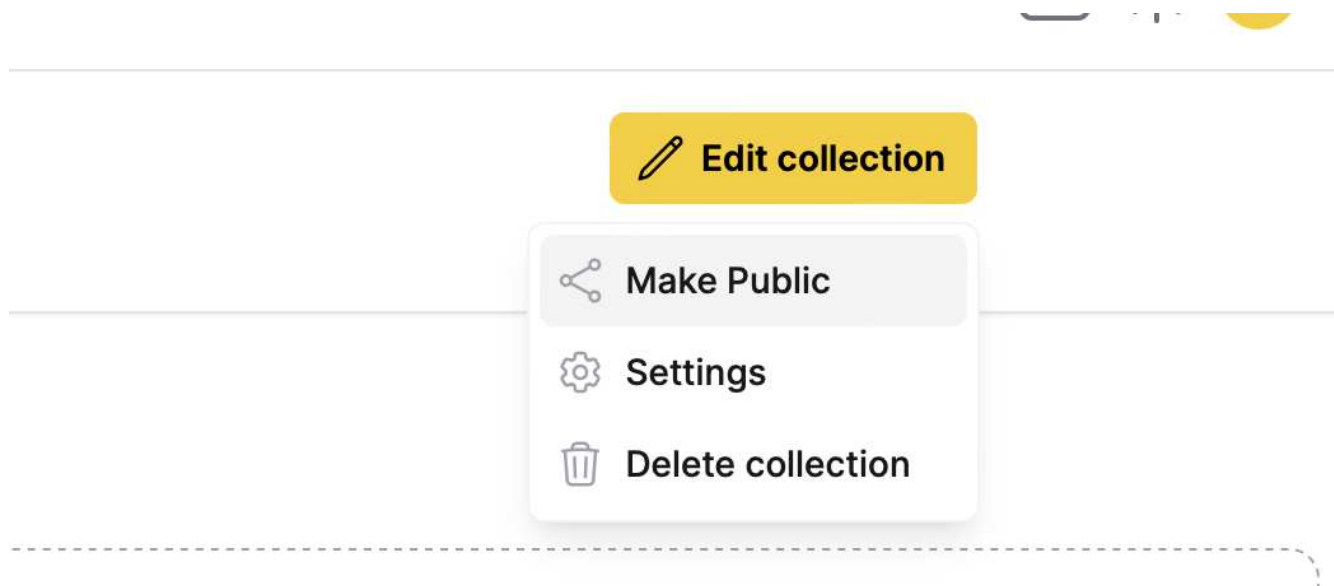
### Note:

- Once a Collection is made public, any authenticated user of the system or API can view and interact with the Collection.
- To learn how to share a Collection's Chat session, see [Share a Chat session](#).

## Instructions

To make a Collection public, consider the following instructions:

1. In the **Enterprise h2oGPTe** navigation menu, click **Collections**.
2. In one of the following tabs, click the Collection that you want to share:
  - **All collections**
  - **My collections**
  - **Shared**
3. Click **Edit collection**.
4. In the **Edit collection** list, select **Make Public**.



5. In the **Are you sure you want this collection to be public?** card, click **Make Public**.

**Note:** To switch the Collection back to private, repeat the previous steps and select **Make private** in step 4.

# Run an Evaluator on a Collection

## Overview

Enterprise h2oGPTe offers several Evaluators to assess a Collection's performance, reliability, security, fairness, and effectiveness. The available Evaluators for a Collection are based on the Evaluators in [H2O Eval Studio](#). To learn about the available Evaluators for a Collection, see [Collection Evaluators](#).

## Instructions

To run an Evaluator on a Collection, consider the following steps:

1. In the **Enterprise h2oGPTe** navigation menu, click **Collections**.
2. From one of the following tabs, locate and select the Collection you want to evaluate.
  - **All collections**
  - **My collections**
  - **shared**
3. Click **Evaluations**.
4. Click **Run your first evaluation/New evaluation**.
5. In the **Evaluator** list, select an Evaluator. **Note:** To learn about each available Evaluator for a Collection, see [Collection Evaluators](#).
6. Click **Evaluate**.

## Collection Evaluators

This section lists all available Evaluators for a Collection.

### Toxicity

At a high level, this Evaluator helps you determine if the Collection's LLM responses contain harmful, offensive, or abusive language that could negatively impact users or violate platform guidelines. To learn more about this Evaluator, see [Toxicity Evaluator](#).

### Hallucination

This Evaluator identifies whether the Collections LLM responses include fabricated or inaccurate information that doesn't align with the provided context or factual data. To learn more about this Evaluator, see [Hallucination Evaluator](#).

### Personally Identifiable Information (PII) leakage

This Evaluator checks if the Collection's LLM responses inadvertently reveals sensitive personal data, such as names, addresses, phone numbers, or other details that could be used to identify an individual. To learn more about this Evaluator, see [PII Leakage Evaluator](#).

### Sensitive data leakage

This Evaluator detects if the Collection's LLM discloses confidential or protected information, such as proprietary business data, medical records, or classified content, which could result in security or privacy breaches. To learn more about this Evaluator, see [Sensitive Data Leakage Evaluator](#).

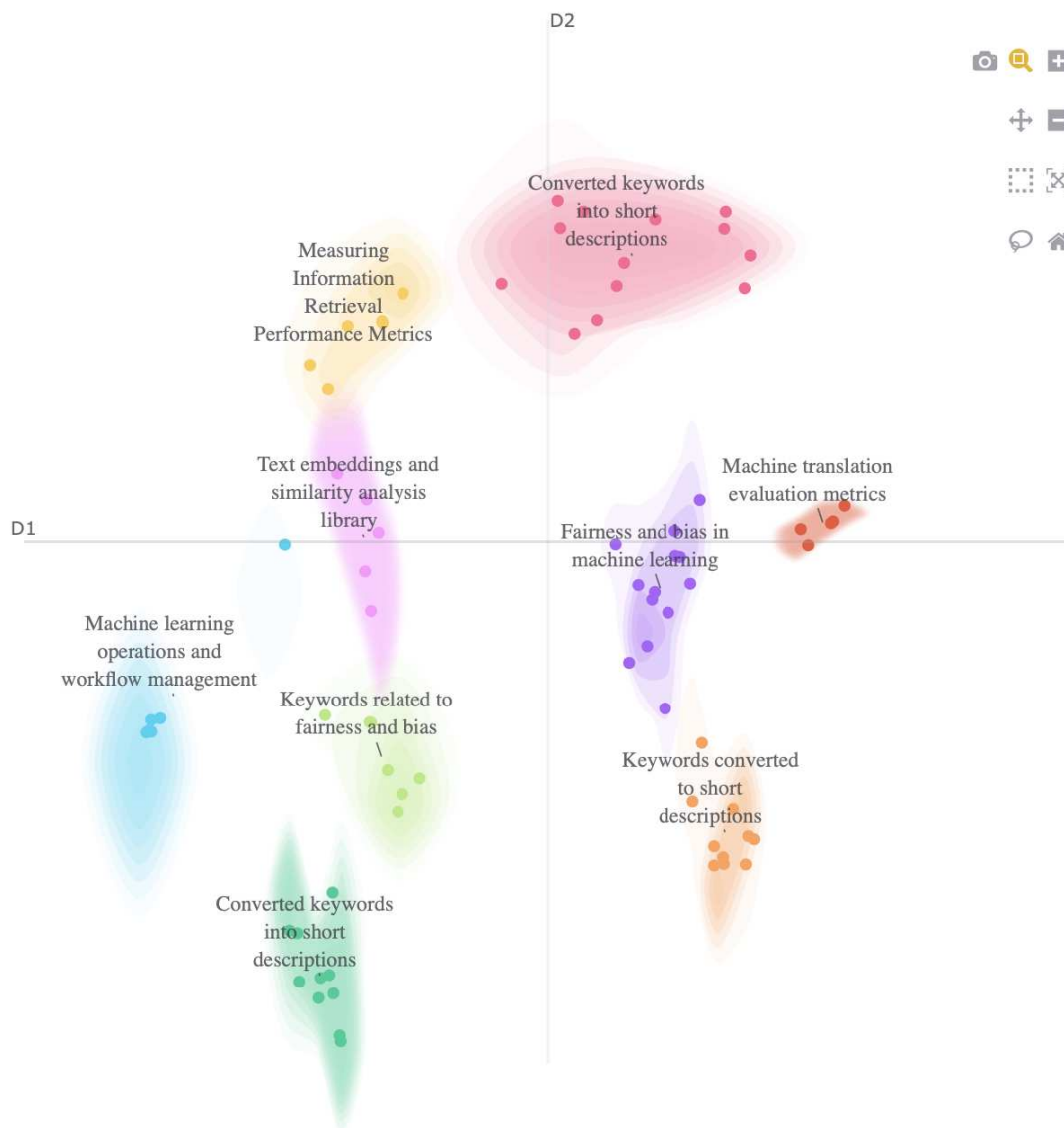
### Fairness bias

This Evaluator assesses whether the Collection's LLM responses exhibit bias or unfair treatment based on gender, race, ethnicity, or other demographic factors, ensuring that the model's output is impartial and equitable. To learn more about this Evaluator, see [Fairness Bias Evaluator](#).

## Create a topic model for a Collection

### Overview

You can create a topic model to visualize and analyze abstract topics within a Collection. A topic model clusters similar words from all documents in the Collection and represents these clusters as topics on a 2D graph.



### Instructions

To create a topic model for a Collection, consider the following steps:

1. In the **Enterprise h2oGPTe** navigation menu, click **Collections**.
2. In one of the following three tabs, select the Collection you want to generate a topic model for:
  - **All collections**
  - **My collections**
  - **Shared**
3. Click **Create a Topic Model**.

# Delete a Collection

## Overview

Once a Collection(s) is no longer in use, you can use one of the following two methods to delete a Collection and remove it from the Collections list/grid:

- [Delete a specific Collection](#)
- [Delete mutiple Collections](#)

### Caution:

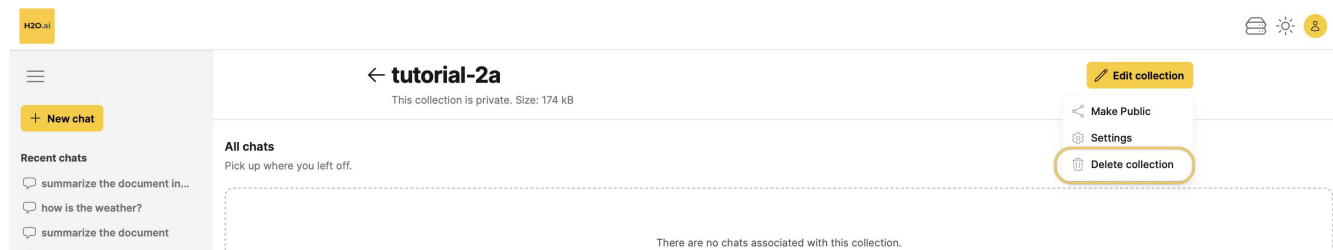
- A deleted Collection is no longer accessible by other users.
- Deleting a Collection removes the Collection from our servers permanently. Once a Collection is deleted, there is no way to undo the action or recover the deleted Collection or its Document(s).

## Delete a specific Collection

To delete a specific Collection, consider the following instructions:

**Note:** You can only delete your Collections.

1. In the **Enterprise h2oGPTe** navigation menu, click **Collections**.
2. In the following tab, click the Collection you want to delete: **My collections**.
3. Click **Edit collection**.
4. Select **Delete collection**.



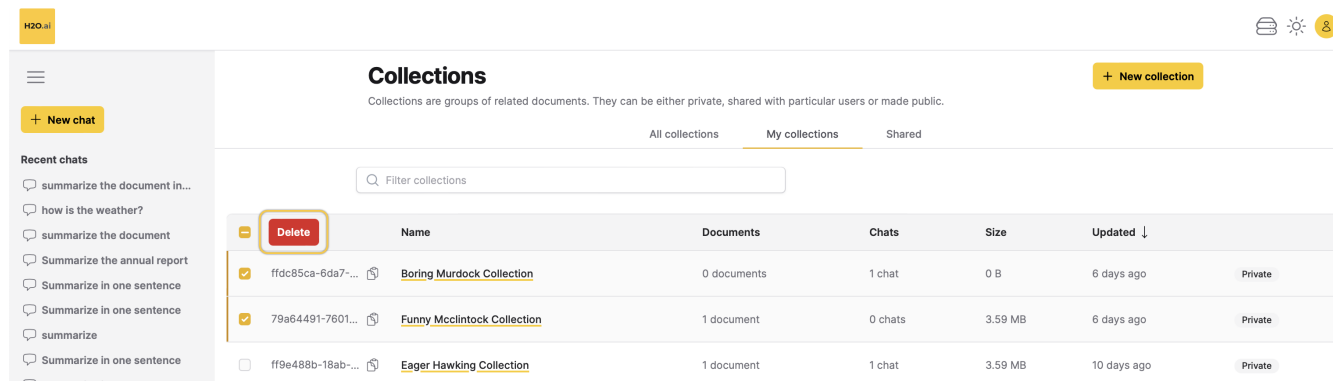
5. In the **Are you sure?** card, click **Delete** to confirm.

## Delete multiple Collections

To delete multiple Collections at once, consider the following instructions:

**Note:** You can only delete your Collections.

1. In the **Enterprise h2oGPTe** navigation menu, click **Collections**.
2. In the following tab, click the checkbox of the Collections you want to delete: **My Collections**.
3. Click **Delete**.



4. In the **Are you sure?** card, click **Delete** to confirm.

# Generate a Document summary

## Overview

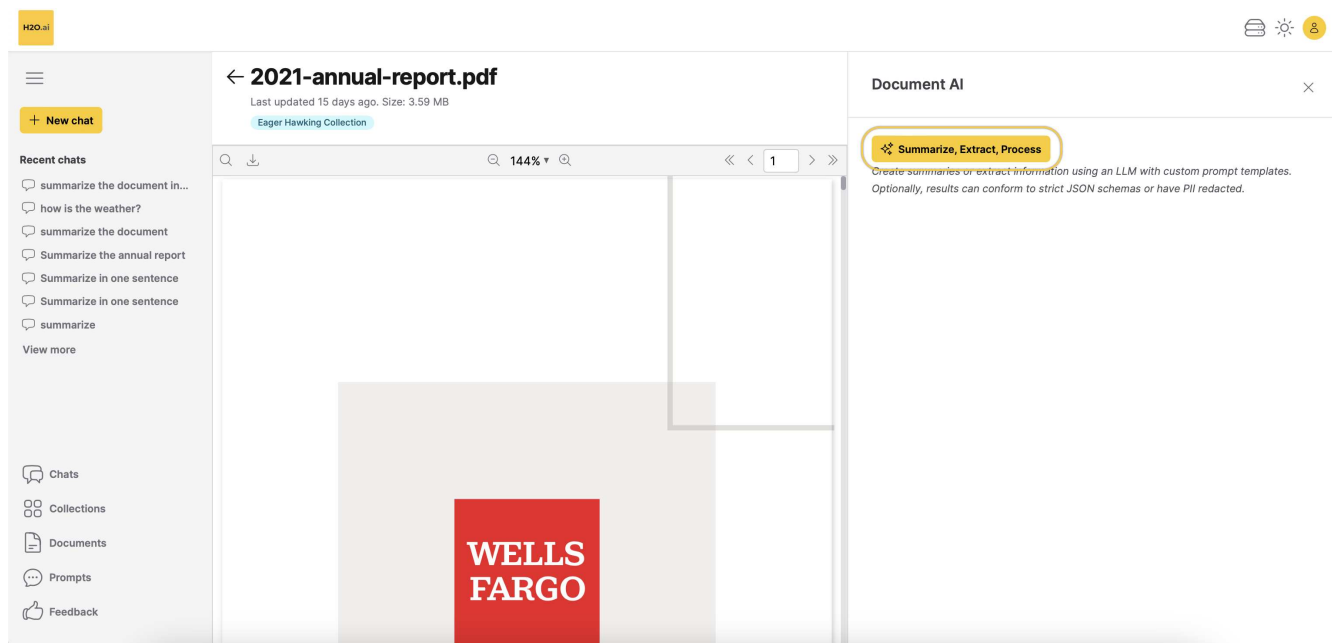
In the **Documents** section, you can generate a summary of the Document using **Document AI**.

**Note:** A summary of the Document using Document AI can not be generated unless the Document has been added to a Collection first. To learn how to add a Document to a Collection, see [Add a Document to a Collection](#).

## Instructions

To generate Document AI, follow these steps:

1. In the **Enterprise h2oGPTe** navigation menu, click **Documents**.
2. In the **Documents** grid or list, select the name of the Document you want to create a new summary of.
3. Click **Summarize, Extract, Process** under the **Document AI** section.



4. In the **Summarize, Extract, Process** panel, customize the Document AI settings according to your requirements.  
**Note:** To learn more about each setting, see [Document AI settings](#).
5. Click **Summarize, Extract, Process** after finalizing the settings.

## Document AI settings

### LLM

- This setting enables you to select the Large Language Model (LLM) to generate the Document summary.
- You can choose from a variety of LLMs by selecting the desired model from the menu.

### Enable vision

In addition to sending document context to the normal Large Language Model (LLM), this setting allows you to pass document context as images to a vision-capable LLM. If you enable the automatic mode, it will select the vision-capable LLM based on the document context and the LLM model.

**Note:** Enabling vision mode can lead to higher latency and cost.

## Vision LLM

This setting allows you to select the LLM for processing images. Selecting automatic mode will pick a vision LLM based on availability and configuration. It typically selects the same LLM for vision-capable models and the default LLM for non-vision models.

## Max. number of chunks and approximate cost range

This setting allows you to configure the number of chunks to be extracted from the Document to summarize its context. Adjust the slider to select the desired number of chunks for the summary.

**Note:** Moving the slider to **100** prompts the Large Language Model (LLM) to utilize **100** chunks from the Document to generate the Document summary.

## Prompt Template

This setting allows you to select a prompt template from the drop-down menu to customize the prompts used for the Collection. You can create your own prompt template on the **Prompts** page and use it for your Collection.

## Personality (System Prompt)

This setting allows you to customize the personality of the LLM according to your requirements for the Document summary. It aids in shaping the behavior of the generated Document summary.

**Example:** You are h2oGPTe, an expert question-answering AI system created by H2O.ai that performs like GPT-4 by OpenAI.

## Prompts before/after document context

This setting enables you to specify prompts before and after the Document context to guide the language model (LLM) in generating a Document summary.

**Before prompt example:** “In order to write a concise single-paragraph or bulleted list summary, pay attention to any chat history, any images given, or any following text: **DOCUMENT\_CONTEXT**”

**After prompt example:** “Using only any chat history, any images given, or any text above, write a condensed and concise summary of key results (preferably as bullet points).”

## Prompt per image batch for vision models

This setting defines the prompt used to obtain answers for the user queries from a batch of images when vision mode is enabled.

## Prompt for final image batch reduction for vision models

This setting defines the prompt used to obtain the final answer for the user query from the per-image-batch answers when vision mode is enabled.

## JSON Schema

This setting enables you to ensure that the output will conform to the JSON schema. Use **{}** to let the LLM determine the output schema.

## PII Detection

This setting enables you to detect personally identifiable information (PII) in the Document summary. Click on the drop-down menu to select a PII detection option.

**Note:** The PII detection policies apply in addition to system settings.

## Document Metadata to include

This setting allows you to provide additional Document info as part of the **DOCUMENT\_CONTEXT**. You can select the metadata fields to include in the Document summary by using the drop-down menu.

**Keep intermediate results**

If this setting is enabled, the final result will be a list of intermediate results. Otherwise, a global result is created. This is useful for debugging and understanding the process of generating the Document summary.



# View a Document summary

## Overview

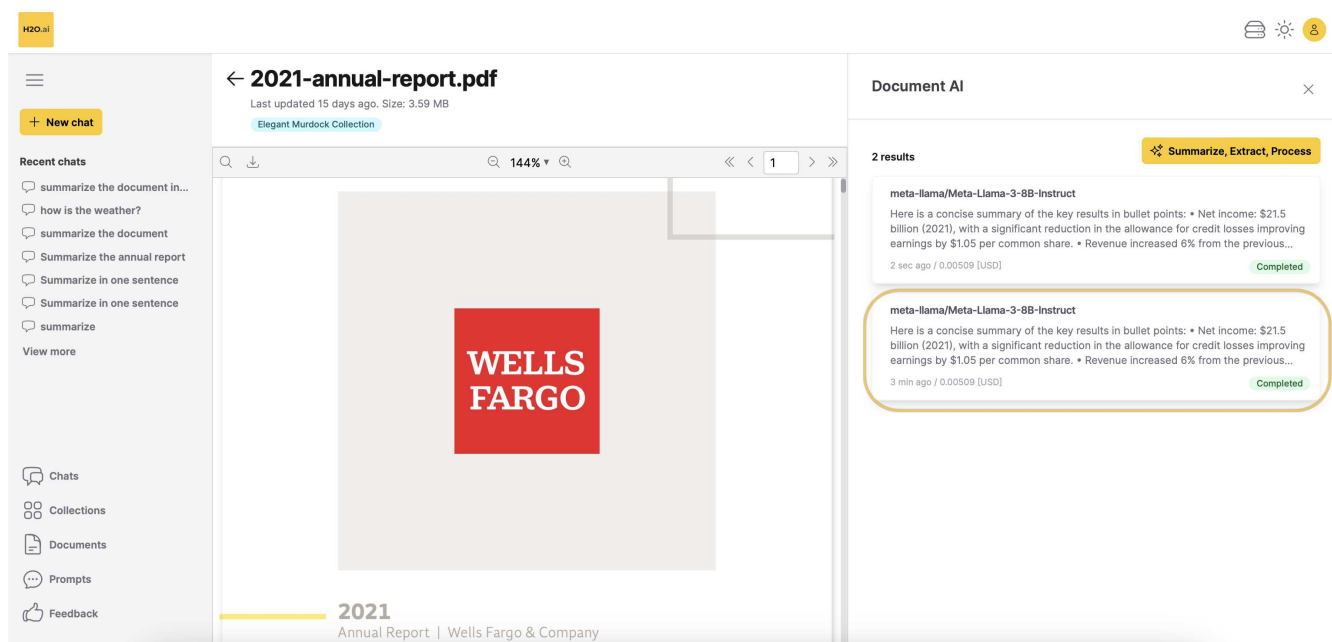
In the **Documents** section, you can view the summary and the summary parameters of an uploaded Document if a summary exists for the particular Document.

## Instructions

The following steps describe how to view a Document summary.

**Caution:** The following steps assume that a Document summary already exists. To learn how to create a Document summary, see [Generate a Document summary](#).

1. In the **Enterprise h2oGPTe** navigation menu, click **Documents**.
2. In the **All documents** grid or list, select the name of the Document you want to view a summary of.
3. In the **Document AI** section, select the Document summary you want to view.



# Delete a Document summary

## Overview

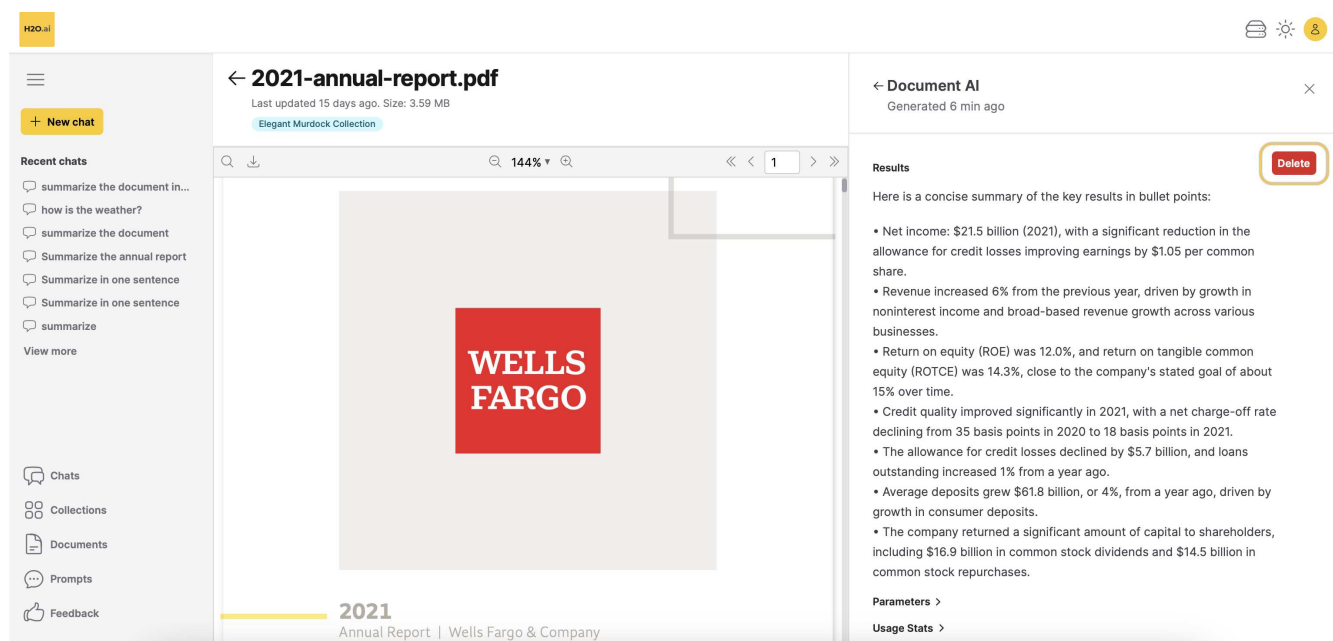
In the **Documents** section, you can delete a Document summary.

**Caution:** Deleting a Document summary removes the summary from the servers permanently. Once a Document summary is deleted, there is no way to undo the action or recover the deleted summary.

## Instructions

The following steps describe how to delete a Document summary.

1. In the **Enterprise h2oGPTe** navigation menu, click **Documents**.
2. Click the card or row in the **All documents** grid or table corresponding to the Document with the summary you wish to delete.
3. In the **Document AI** panel, select the Document summary you wish to remove.
4. Click **Delete**.



5. In the **Are you sure?** card, click **Delete** to confirm.

# Delete a Document(s)

## Overview

In the **Documents** section, you can delete a Document(s) added to a Collection.

### Caution:

- You can only delete your Document(s).
- Deleting a Document removes the Document from the Collection it was added to.
- Deleting a Document removes the Document from the servers permanently. Once a Document is deleted, there is no way to undo the action or recover the deleted Document.

## Instructions

To delete a Document(s), consider the following instructions:

1. In the **Enterprise h2oGPTe** navigation menu, click **Documents**.
2. Click **List**.
3. Select the checkbox of the row(s) containing the Document(s) you wish to delete.
4. Click **Delete**.

The screenshot shows the H2O.ai interface. On the left is a sidebar with a menu and a 'New chat' button. The main area displays a table of documents. A red box highlights the 'Delete' button in the top left of the table. The table has columns: Name, Size, Pages, Type, Source, URI, Status, User, Updated, and Age. Two rows are selected with checkboxes.

	Name	Size	Pages	Type	Source	URI	Status	User	Updated	Age
<input type="checkbox"/>	113c... <a href="#">engineering_sync_2024-06-11</a>	46.1 kB	2	Markdown	Upload	-	Indexed	michelle.tanco@h2o.ai	1 day ago	1 day
<input type="checkbox"/>	972... <a href="#">engineering_sync_2024-06-10</a>	42.5 kB	2	Markdown	Upload	-	Indexed	michelle.tanco@h2o.ai	2 days ago	2 days
<input type="checkbox"/>	40df... <a href="#">engineering_sync_2024-06-07</a>	39.1 kB	2	Markdown	Upload	-	Indexed	michelle.tanco@h2o.ai	5 days ago	5 days
<input type="checkbox"/>	b9d... <a href="#">product_information_jargon</a>	18.1 kB	1	Markdown	Upload	-	Indexed	michelle.tanco@h2o.ai	5 days ago	5 days
<input type="checkbox"/>	17e6... <a href="#">product_information_team</a>	20.7 kB	1	Markdown	Upload	-	Indexed	michelle.tanco@h2o.ai	5 days ago	5 days
<input type="checkbox"/>	7ed... <a href="#">product_information_summary</a>	45.1 kB	3	Markdown	Upload	-	Indexed	michelle.tanco@h2o.ai	5 days ago	5 days
<input checked="" type="checkbox"/>	eef1... <a href="#">2021-annual-report-3</a>	3.59 MB	226	PDF	Upload	-	Indexed	oshini.nugapitiya@h2o.ai	6 days ago	5 mo
<input checked="" type="checkbox"/>	952... <a href="#">2021-annual-report-7</a>	3.59 MB	226	PDF	Upload	-	Indexed	oshini.nugapitiya@h2o.ai	7 days ago	5 mo
<input type="checkbox"/>	866... <a href="#">HR_Policy_Procedures</a>	367 kB	53	PDF	Upload	-	Indexed	pramit.choudhary@h2o.ai	8 days ago	1 mo

5. In the **Are you sure?** dialog, click **Delete**.

# Models

## Overview

The **Models** page allows you to explore the supported large language models (LLMs) and perform self-tests on the LLMs used throughout Enterprise h2oGPTE.

## Supported LLMs

Enterprise h2oGPTE supports the following LLMs:

- [meta-llama/Meta-Llama-3.1-8B-Instruct](#)
- [h2oai/h2o-danube3-4b-chat](#)
- [Qwen/Qwen2-VL-7B-Instruct](#)
- [meta-llama/Llama-3.3-70B-Instruct](#)
- [meta-llama/Meta-Llama-3.1-70B-Instruct](#)
- [meta-llama/Meta-Llama-3.1-405B-Instruct-FP8](#)
- [Qwen/Qwen2.5-72B-Instruct](#)
- [Qwen/Qwen2-VL-72B-Instruct](#)
- [mistralai/Pixtral-12B-2409](#)
- [mistralai/Mixtral-8x7B-Instruct-v0.1](#)
- [meta-llama/Meta-Llama-3.1-70B-Instruct-Turbo](#)
- [meta-llama/Meta-Llama-3.1-8B-Instruct-Turbo](#)
- [upstage/SOLAR-10.7B-Instruct-v1.0](#)
- [mistralai/Mistral-7B-Instruct-v0.3](#)
- [meta-llama/Llama-3.2-90B-Vision-Instruct-Turbo](#)
- [meta-llama/Llama-3.2-11B-Vision-Instruct-Turbo](#)
- [meta-llama/Llama-3.2-3B-Instruct-Turbo](#)
- [mistral-tiny](#)
- [mistral-small-latest](#)
- [mistral-medium](#)
- [mistral-large-latest](#)
- [gemini-1.5-pro-latest](#)
- [gemini-1.5-flash-latest](#)
- [claude-3-haiku-20240307](#)
- [claude-3-5-haiku-20241022](#)
- [claude-3-5-sonnet-20240620](#)
- [gpt-4o](#)
- [gpt-4o-mini](#)

**Note:** The table inside the [LLMs](#) tab renders the supported LLMs.

## Run self-tests

### Overview

A self-test enables you to evaluate the LLMs used across Enterprise h2oGPTE to generate a response to a query and process and summarize a Document.

### Instructions

To run a self-test, consider the following steps:

1. In the **Enterprise h2oGPTE** navigation menu, click **Models**.
2. Click **Run self-tests**.
3. In the **Run self-tests** list, select a self-test. **Note:** options
  - **Quick test** (chat-like short query): This self-test evaluates the model's ability to respond to brief, conversational queries accurately and contextually.

- **RAG test** (large context query): This self-test evaluates the model's ability to understand and respond accurately to queries that require it to process and maintain a long context or multiple pieces of information over an extended conversation.
- **Full test** (full context query): This self-test evaluates the model's ability to understand, retain, and use the information provided across a conversation to respond to complex, multi-part queries. It requires the model to keep track of various details and offer responses that synthesize all the given information. **Note:** This option is only available to an admin.
- **Stress test** (repeated full test): This self-test evaluates the model's consistency, accuracy, and contextual awareness over multiple interactions on the same topic. It ensures the model can handle information continuity over several sessions, maintaining the context and providing accurate responses. **Note:** This option is only available to an admin.

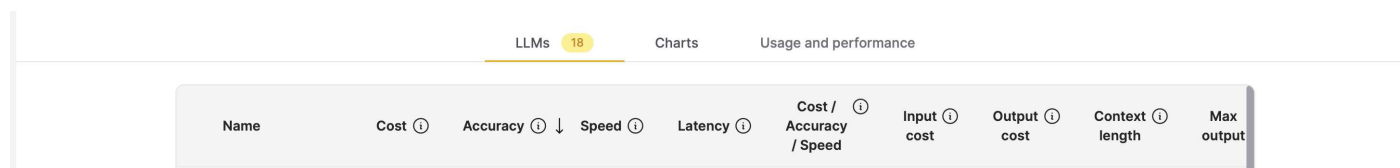
4. (Conditional step) If you selected **Full test** or **Stress test**, consider the following steps:

- **Full test**
  1. In the **Run full tests** box, click **Run tests**.
- **Stress test**
  1. In the **Run stress tests** box, click **Run tests**.

## Tabs

### LLMs

**Overview** The **LLMs** tab contains a table rendering the large language models (LLMs) supported for generating responses to user queries and summarizing and processing documents.



LLMs 18									
Charts Usage and performance									
Name	Cost ⓘ	Accuracy ⓘ ↓	Speed ⓘ	Latency ⓘ	Cost / Accuracy / Speed ⓘ	Input ⓘ cost	Output ⓘ cost	Context ⓘ length	Max output

**Instructions** To access the **LLMs** tab, consider the following steps:

1. In the **Enterprise h2oGPTe** navigation menu, click **Models**.

### Charts

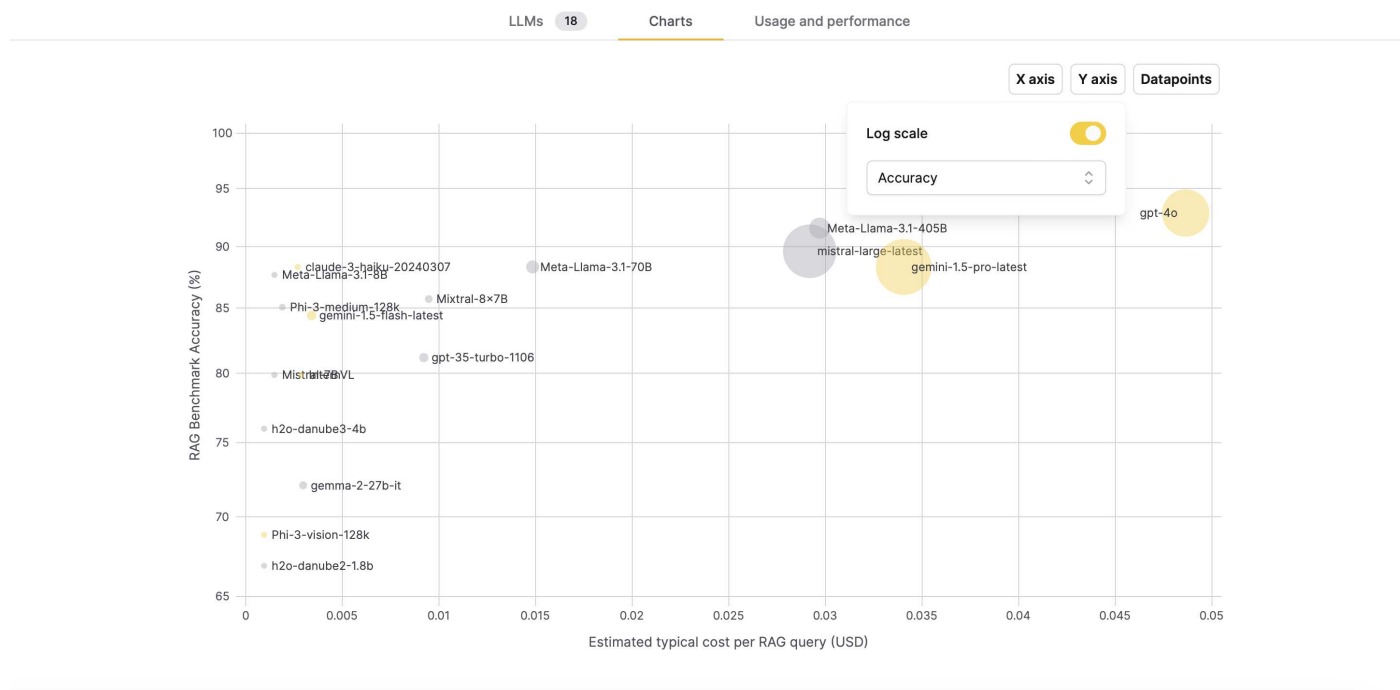
**Overview** The chart in the **Charts** tab allows you to compare various models across several selected metrics by adjusting its **x** and **y** axes to these metrics. These models generate responses to user queries and summarize and process Documents.

**Note:** Example For example, in the image below, the chart compares the performance and usage characteristics of various large language models (LLMs). The x-axis shows the estimated typical cost per query (in USD), while the y-axis displays the RAG benchmark accuracy percentage.

The chart includes several data points representing different LLMs, such as GPT-4, Meta-Llama, and Mistral. Each data point is labeled with the model name and some additional information, such as the version number or specific capabilities.

In this example, the top right of the chart indicates models with higher accuracy and higher cost per query, while the bottom left shows models with lower accuracy and lower cost per query.

The chart can be useful for researchers, developers, or users to compare the performance and cost-effectiveness of different large language models. This can help them select the most appropriate model for their needs or applications.



**Note:** Log-scaling the x/y axis Log-scaling the **x/y axis** of the graph means transforming the **x/y axis** values using a logarithmic function. This transformation is particularly useful when dealing with data that spans several orders of magnitude, as it can help visualize the data more clearly by compressing the scale of large values and expanding the scale of small values.

**Instructions** To access the **Charts** tab, consider the following steps:

1. In the **Enterprise h2oGPTe** navigation menu, click **Models**.
2. Click the **Charts** tab.

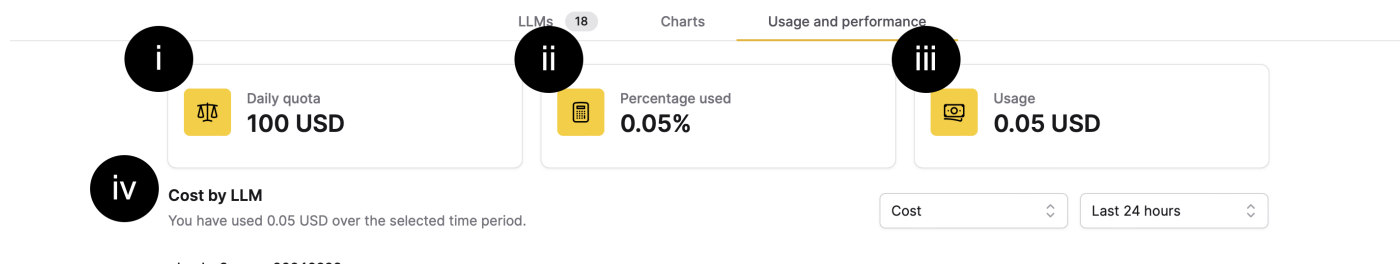
## Usage and performance

**Overview** The **Usage and performance** tab provides an overview of your approximated large language model (LLM) usage in the last 24 hours. The section includes details on quotas, cost estimates, and usage breakdown by specific LLMs.

**Note:** The daily quotas and cost estimates are configured by administrators.

**Instructions** To access the **Usage and performance** tab, consider the following steps:

1. On **Enterprise h2oGPTe** navigation menu, click **Models**.
2. Click the **Usage and performance** tab.
  1. **Daily quota:** The daily allocation for LLM usage.
  2. **Percentage used:** The percentage of the daily quota that has been utilized.
  3. **Usage:** The cost incurred for LLM usage within the last 24 hours.
  4. **[Cost], [Speed], or [Latency] by LLM:** This section covers the cost, speed, or latency of each LLM incurred in a given timeframe.



**Note:** The **[Cost], [Speed], or [Latency] by LLM** section defaults to displaying usage within the last 24 hours. However, you can choose different time frames to see usage information within the **Last hour**, **Last 24 hours**, **Last**

**week, Last month, or All time.**

## Prompts

### Create a prompt template

#### Overview

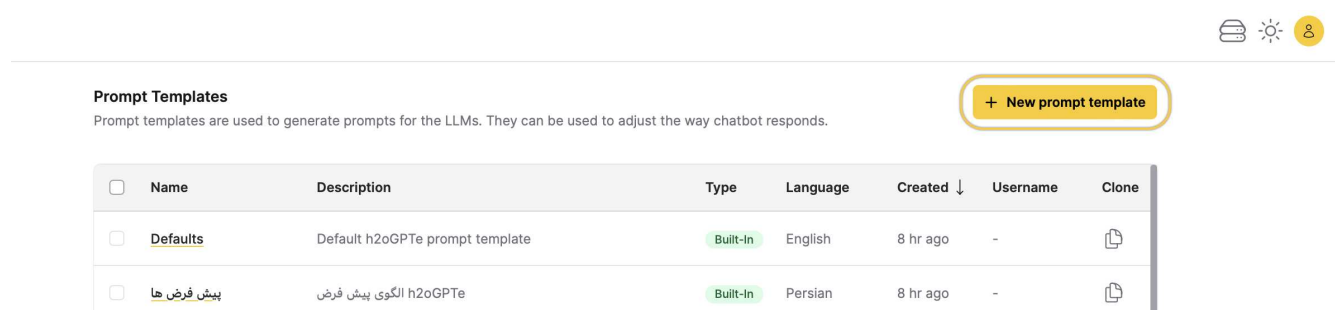
You can create an array of prompt templates for a Collection or Chat session, each designed to guide the language model in generating specific tailored text (for example, tailored text for a Collection’s description or a question from a Chat session).

These templates serve as structured outlines with instructions for the model to follow, ensuring consistent and tailored outputs for your specific needs. In other words, prompt templates enable customization of responses from Enterprise h2oGPTe by providing reusable instructions sent alongside the user’s query/request to the LLM.

#### Instructions

To create a new prompt template, consider the following instructions:

1. On the **Enterprise h2oGPTe** navigation menu, click **Prompts**.
2. Click **+ New prompt template**.



3. In the **Template name** box, enter a name for the prompt template.
4. (Optional) In the following setting sections, make the changes you want:
  - [General](#)
  - [Prompts](#)
  - [Self-reflection](#)
  - [Sample questions](#)
5. Click **+ Create**.

### Create a prompt template for a specific language

Enterprise h2oGPTe features built-in prompt templates for Chinese, Turkish, Russian, Portuguese, Malay, Japanese, Indonesian, Hindi, French, Persian, and Spanish. You can also create your own prompt template instead of using the defaults and select a language from a list of about 180 languages that you can choose from.

The built-in prompt templates support multiple languages by:

1. Writing the System Prompt (personality) in the specified language.
2. Writing all prompt extensions in the specified language (the prompt used before and after the RAG chunks are included in the prompt, self-reflection prompts, etc.).
3. Using an embedding model (the model that converts the documents to embeddings in the vector database) that supports the selected language.

**Caution:** If the LLM you choose to use to answer the questions does not understand the language you have chosen, then you will likely get responses back in English. For example, Mixtral only supports English, French, Italian, German, and Spanish.

### Clone a prompt template

#### Overview

Once you’ve designed a new prompt template, you can duplicate it to generate additional templates with identical or similar configurations. This feature streamlines the process of creating multiple templates tailored to your specific requirements.



## Instructions




To clone a prompt template, consider the following steps:

1. On the **Enterprise h2oGPTe** navigation menu, click **Prompts**.
2. In the **Prompts** table, locate the row where the prompt template you want to clone is located.
3. Click file\_\_copy **Clone**.

**Prompt Templates**

Prompt templates are used to generate prompts for the LLMs. They can be used to adjust the way chatbot responds.

[+ New prompt template](#)

<input type="checkbox"/>	Name	Description	Type	Language	Created ↓	Username	Clone
<input type="checkbox"/>	<u>Defaults</u>	Default h2oGPTe prompt template	Built-in	English	5 days ago	-	
<input type="checkbox"/>	<u>پیش فرض ها</u>	h2oGPTe الگوی پیش فرض	Built-in	Persian	5 days ago	-	
<input type="checkbox"/>	<u>Minimal</u>	Just keeps user prompts	Built-in	English	5 days ago	-	

4. In the **Template name** box, enter a name for the clone prompt template.
5. (Optional) In the following setting sections, make the changes you want:
  - [General](#)
  - [Prompts](#)
  - [Self-reflection](#)
  - [Sample questions](#)
6. Click **+ Clone**.

## Prompt template settings

### General

**Template name** This setting defines the name for the prompt template.

**Language** This setting defines the language for the prompt template.

**Description** This setting defines the description for the prompt template.

### Prompts

**System prompt** This setting defines a system prompt, which refers to a language model's initial input (instruction) to initiate a specific task or interaction. In the context of natural language processing models like Enterprise h2oGPTe, a system prompt is the initial text input given to the model to prompt it to generate a response for a Collection and Document.

System prompts can vary widely depending on the task. They serve as a guide for the model, providing context and direction for generating the desired output. System prompts can range from simple queries or instructions to more complex scenarios or prompts tailored to specific tasks, such as summarization.

In essence, a system prompt acts as the starting point for the language model's generation process, shaping the direction and content of its responses.

**For example:** You are h2oGPTe, an expert question-answering document AI system created by H2O.ai that performs like GPT-4 by OpenAI.

**Note:** The system prompt provides overall alignment and safety and is used for all Chats and Document summarizations. Self-reflection uses its own system prompt.

## Chat prompts before/after document context

1. **Before text box:** This text box allows you to define a prompt **before** the Document(s) contexts within your Collection. This prompt is crucial in constructing the LLM prompt that Enterprise h2oGPTe sends to the Large Language Model (LLM). The LLM prompt serves as the question or instruction you send to the LLM to generate the desired response. You can customize this RAG prompt according to your specific requirements and objectives.

**For example:** Pay attention and remember the information below. You will need to use only the given document context to answer the question or imperative at the end.

2. **After text box:** This text box allows you to define a prompt **after** the Document contexts within your Collection. This prompt is crucial in constructing the LLM prompt that Enterprise h2oGPTe sends to the Large Language Model (LLM). The LLM prompt serves as the question or instruction you send to the LLM to generate the desired response. You can customize this prompt according to your specific requirements and objectives.

**For example:** According to only the information in the document sources provided within the context above,

### Chat prompts before/after document context ⓘ

Pay attention and remember the information below. You will need to use only any chat history, any images given, or any document text in order to answer the question or imperative at the end.

<DOCUMENT CONTEXT>

According to only the information in any chat history, any images given, or any document text provided within the context above,

<USER PROMPT>

**Note:** This setting is the most important set of prompts when doing grounded generation. Grounded generation, within the context of Language Models (LLMs), refers to the generation of text that is coherent and grammatically correct and aligned with a given context, prompt, or grounding information. This grounding information could include textual prompts, keywords, or any other form of input that constrains or guides the generation process.

In the context of LLMs like Enterprise h2oGPTe, grounding can help steer the generation process toward producing more relevant and appropriate text for a given task or scenario. For example, providing a prompt about a specific topic or domain can help the model generate more focused and accurate text within that domain.

**HyDE No-RAG LLM prompt extension** This setting defines a prompt extension that helps retrieve more relevant context from the Document(s) during the first large language model (LLM) call when doing HyDE (hypothetical document embedding) based generation.

**Document AI prompts before/after document context** This setting delineates the pre- and post-context prompts used by Enterprise h2oGPTe to generate a Document summary.

1. **Pre-context:** The first part of the setting (first text box) enables you to specify a pre-context prompt involving the provision of specific instructions or queries to guide the language model (LLM) before introducing contextual information. Such a prompt steers the subsequent document summary generated by the LLM, ensuring alignment with the intended context or query.
2. **Post-context:** The second part of the setting (second text box) enables you to specify a prompt after context, which refers to providing a specific instruction or query to guide the language model (LLM) after presenting it with contextual information. This prompt is used to direct the subsequent document summary generated by the LLM, ensuring that it aligns with the provided context or query.

**Document AI prompts before/after document context** ⓘ

In order to write a concise single-paragraph or bulleted list summary, pay attention to any chat history, any images given, or any following text:

<DOCUMENT CONTEXT>

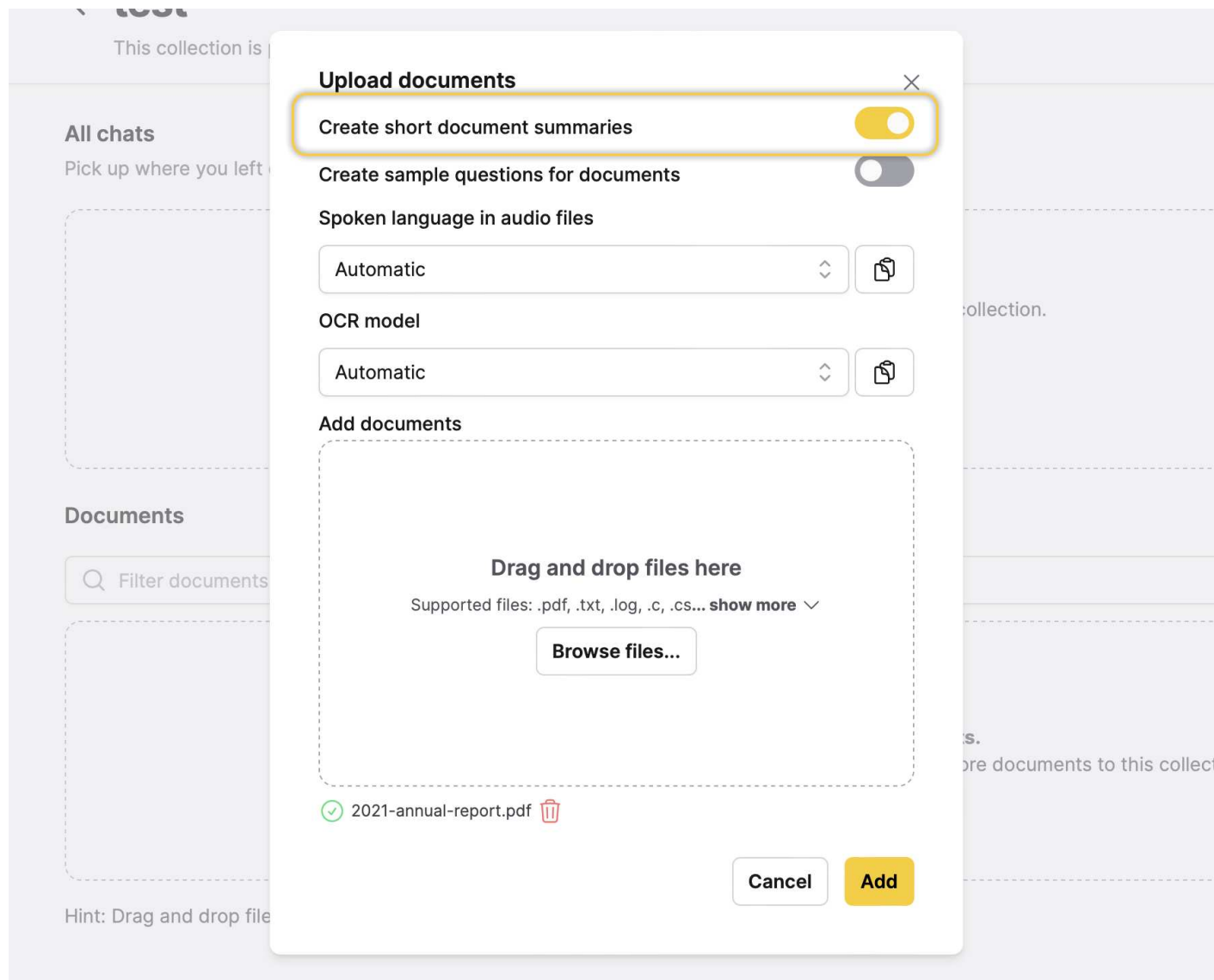
Using only any chat history, any images given, or any text above, write a condensed and concise summary of key results (preferably as bullet points).

**Auto-gen collection description prompt** This setting defines a prompt to automatically generate a description for a Collection. This prompt guides the system as it crafts a concise and informative summary of the content or purpose of a Collection. Customizing this prompt allows you to tailor the generated descriptions to suit your needs, ensuring clarity and relevance for users accessing your Collection.

**For example:** Create a short one-sentence summary from the above context, with the goal to make it clear to the reader what this is about.

**Auto-gen document summary prompts** This setting defines a prompt that Enterprise h2oGPTe utilizes to generate a summary of a Document you import. In particular, this prompt is utilized if you specify to Enterprise h2oGPTe that you want an auto-generated summary of a Document you are about to import.

**Note:** To learn how to specify to Enterprise h2oGPTe how to generate an auto-generated summary for a Document you are about to import, see [Add a Document\(s\) to a Collection](#).



**Prompt per image batch for vision models** This setting defines the prompt used to obtain answers for the user queries from a batch of images when vision mode is enabled.

**Prompt for final image batch reduction for vision models** This setting defines the prompt used to obtain the final answer for the user query from the per-image-batch answers when vision mode is enabled.

### Self-reflection

**Self-reflection system prompt (personality)** This setting defines the self-reflection prompt for the [System prompt](#). In particular, this setting defines the prompt for the language model to self-reflect based on predefined personality traits or characteristics. This prompt encourages the model to generate responses that reflect a particular personality, fostering consistency and depth in its interactions. This setting lets users tailor the language model's responses to align with desired personality traits, enhancing the overall user experience and interaction quality.

**For example:** You are acting as a judge. You must be fair and impartial and pay attention to details.

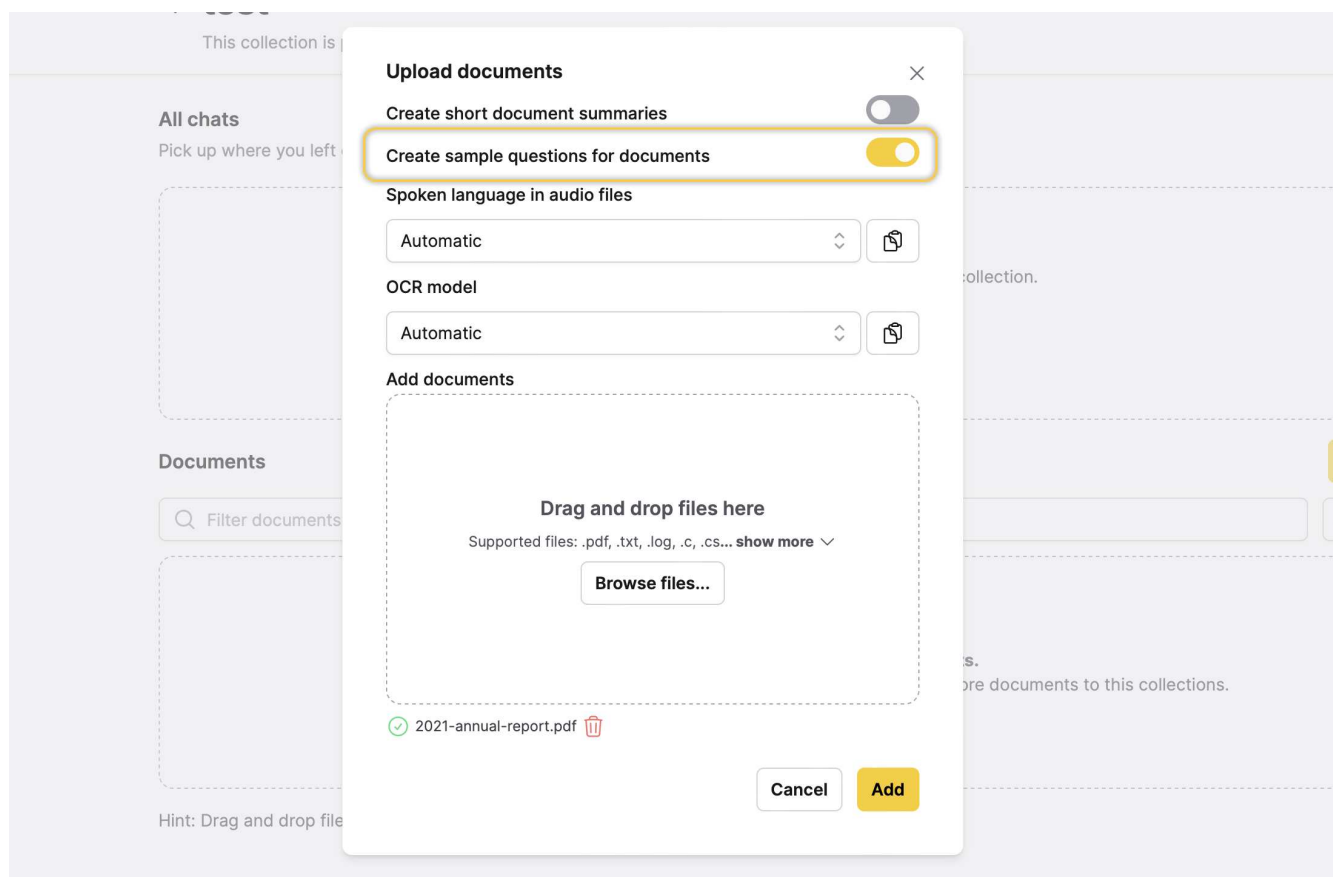
**Self-reflection prompt** This setting defines a self-reflection prompt within the system, prompting the system to engage in self-assessment after processing contextual information. It dictates the wording and format of the prompt, encouraging the system to evaluate the quality and relevance of its responses in light of the provided context. By utilizing this setting, you can guide the system to reflect on its output and enhance its ability to generate coherent and appropriate responses within specific contexts.

## Sample questions

**Document sample questions prompt** This setting defines a prompt that is used to generate a set of questions for a Collection (in particular, for a Document). In particular, this prompt is used if you specify to Enterprise h2oGPTe that you want to auto-generate a set of questions for the Document you are about to import.

### Note:

- By default, Enterprise h2oGPTe displays four auto-generated questions when Chatting with the Collection containing the Document specified to auto-generate questions. These questions are displayed when the following setting is toggled for the Document(s) being imported: **Create sample questions for documents**.
- To learn how you can specify to Enterprise h2oGPTe that you want to auto-generate a set of questions for the Document you are about to import, see [Add a Document\(s\) to a Collection](#).



**Sample questions** In the **Sample questions** setting section, you can define which sample questions will be available when first Chatting with a Collection using the (this) prompt template.

**Note:** By default, Enterprise h2oGPTe displays four sample questions when Chatting with the Collection using the (this) prompt template. These sample questions are displayed when the following setting was not toggled for the Document(s) being imported: **Create sample questions for documents**.

## Sample questions ⓘ

Create five good questions



Summarize in one sentence



Explain this to a 5-year old



Tell me something interesting

**+ Add question**

Explain this to a 5-year old

Create five good questions

Tell me something interesting

Summarize in one sentence

Ask anything



## Extractors

### Overview

Extractors, defined by JSON schemas, play an important role in document AI by converting unstructured document content into structured, actionable data. They allow users to retrieve information from various document types—such as CVs, invoices, Form 10-Ks, or scanned images—without requiring complex setups or extensive annotations.

### Extractor flow

To use an Extractor, first identify the specific information you want to extract from a document. This information is specified in a JSON schema, which is part of an Extractor and acts as a blueprint for the data, detailing the fields and data types you wish to capture. Once you define this schema, you can apply the Extractor to the document, retrieving the desired information in a structured JSON format. This structured data is useful for individuals and applications that require organized information.


### Create an Extractor

To create an Extractor, consider the following steps:

1. In the **Enterprise h2oGPTe** navigation menu, click **Extractors**.
2. Click **+ New extractor**.
3. In the **Extractor name** box, enter a name for the Extractor.
4. In the **LLM** list, select an LLM.
5. Define the labels for the JSON schema with one of the following two options:

#### Option 1 (UI: JSON schema builder)

To build a JSON schema for the Extractor using the JSON schema builder, define each field, its type, and whether it is required. For example:

Input JSON Schema 

Add Field
Add Table


Field Name

Enter field name

Type






String

Required



Add Field

Current Labels

Field Name	Type	Required	Delete
revenueGrowthRate	number	true	
netProfitMargin	number	true	
currentRatio	number	true	
returnOnEquity	number	true	
debtToEquityRatio	number	true	

#### Option 2 (JSON schema code)

To define the actual JSON schema code for the Extractor, consider the following steps:

1. Click the **Input JSON Schema** toggle.
2. In the **JSON schema** box, enter the a valid JSON schema. For example:

```
{
  "$schema": "http://json-schema.org/draft-07/schema#",
  "type": "object",
  "properties": {
    "revenueGrowthRate": {
      "type": "number",
      "description": "The growth rate of revenue."
    }
  },
}
```

```

"netProfitMargin":{
  "type":"number",
  "description":"The company's profit margin."
},
"currentRatio":{
  "type":"number",
  "description":"The company's liquidity position."
},
"returnOnEquity":{
  "type":"number",
  "description":"The efficiency in generating profit from equity."
},
"debtToEquityRatio":{
  "type":"number",
  "description":"The proportion of debt to shareholders' equity."
}
},
"required":[
  "revenueGrowthRate",
  "netProfitMargin",
  "currentRatio",
  "returnOnEquity",
  "debtToEquityRatio"
]
}

```

**Note:** The JSON schema does not require exact label names to align perfectly with document fields, as the collection's large language model (LLM) can interpret and infer label purposes based on context. This allows the model to understand and map various label names, even if there are minor differences in terminology, to their intended data points. Just as a human might deduce what a field intends to capture, the LLM uses its interpretive capability to accurately match schema labels with relevant content, even when exact terms differ.

6. Click **Save**.

## Run an Extractor

To run an Extractor on a document, consider the following steps:

1. In the **Enterprise h2oGPTe** navigation menu, click **Extractors**.
2. In the **Extractors** table, locate the row of the Extractor you want to run and click auto\_awesome **Run** in that row.
3. In the **Select a collection** list, select a Collection. **Note:** The selected Collection must include the document intended for use with the Extractor. The Extractor will retrieve all requested information from the document according to its JSON schema.
4. Click **Run**. **Note:** Enterprise h2oGPTe creates a **Job** to process the Extractor. The Extractor is completed when its Job is completed.

## View a completed Extractor

Once the Extractor has finished processing, you can access the extracted information of a document by following these steps:

6. In the **Enterprise h2oGPTe** navigation menu, click **Collections**
7. Click the **My Collections** tab.
8. In the **My Collections** table, click the Collection name containing the document used for the Extractor.
9. In the **Documents** table, click the document used for the Extractor.
10. The most recent Extractor is located in the **Recent results** section.



Recent results

1 result

🔮 Summarize, Extract, Process

h2oai/h2o-danube3-4b-chat

...

{ "revenueGrowthRate": 0.1, "netProfitMargin": 0.05, "currentRatio": 1.2, "returnOnEquity": 15.0, "debtToEquityRatio": 0.2 }

6 min. ago / 0.00458 [USD]

Completed

Results

Edit Delete

```
{
  "revenueGrowthRate": 0.1,
  "netProfitMargin": 0.05,
  "currentRatio": 1.2,
  "returnOnEquity": 15,
  "debtToEquityRatio": 0.2
}
```

[Original Results](#) >

# Evaluation

## Overview

The **Evaluation** page enables you to assess model performance through leaderboards created in [H2O Eval Studio](#). These leaderboards showcase a comparison of models using diverse benchmarks and evaluation methods within H2O Eval Studio.

## Instructions

To access the **Evaluation** page, consider the following steps:

- In the **Enterprise h2oGPTe** navigation menu, click **Eval**.

## Add a new leaderboard

### Overview

### Instructions

To add a new **H2O Eval Studio** leaderboard, consider the following steps:

1. In the **Enterprise h2oGPTe** navigation menu, click **Eval**.
2. Click **Open Eval Studio** (**H2O Eval Studio** opens on a new page).
3. In **H2O Eval Studio**, click + **Add new leaderboard**.
4. Define the leaderboard settings. **Note:** To learn more about the settings for creating a new leaderboard, see [Create a leaderboard](#).
5. Click **Create**.

## Feedback

### View submitted feedback (Upvote and Downvote responses)

#### Overview

When chatting with a Collection, also known as a Chat session, you can Upvote or Downvote a response from Enterprise h2oGPTe. In particular, when Upvoting or Downvoting a chat response, you can provide written feedback stating why you consider it a good (Upvote) or bad (Downvote) response. Your feedback helps developers improve the model. Enterprise h2oGPTe stores your feedback on the Feedback page.

On the Feedback page, you can view submitted feedback regarding past responses during a Chat session (also known as chatting with a Collection). For each submitted response feedback, you can specifically view the following:

- The responses that received either an Upvote or Downvote.
- Your comments about either an Upvote or Downvote response.

Additionally, you can view the prompt and LLM settings the model used to generate a response.

#### Instructions

To view past submitted feedback responses, consider the following step:

1. In the **Enterprise h2oGPTe** navigation menu, click **Feedback**.

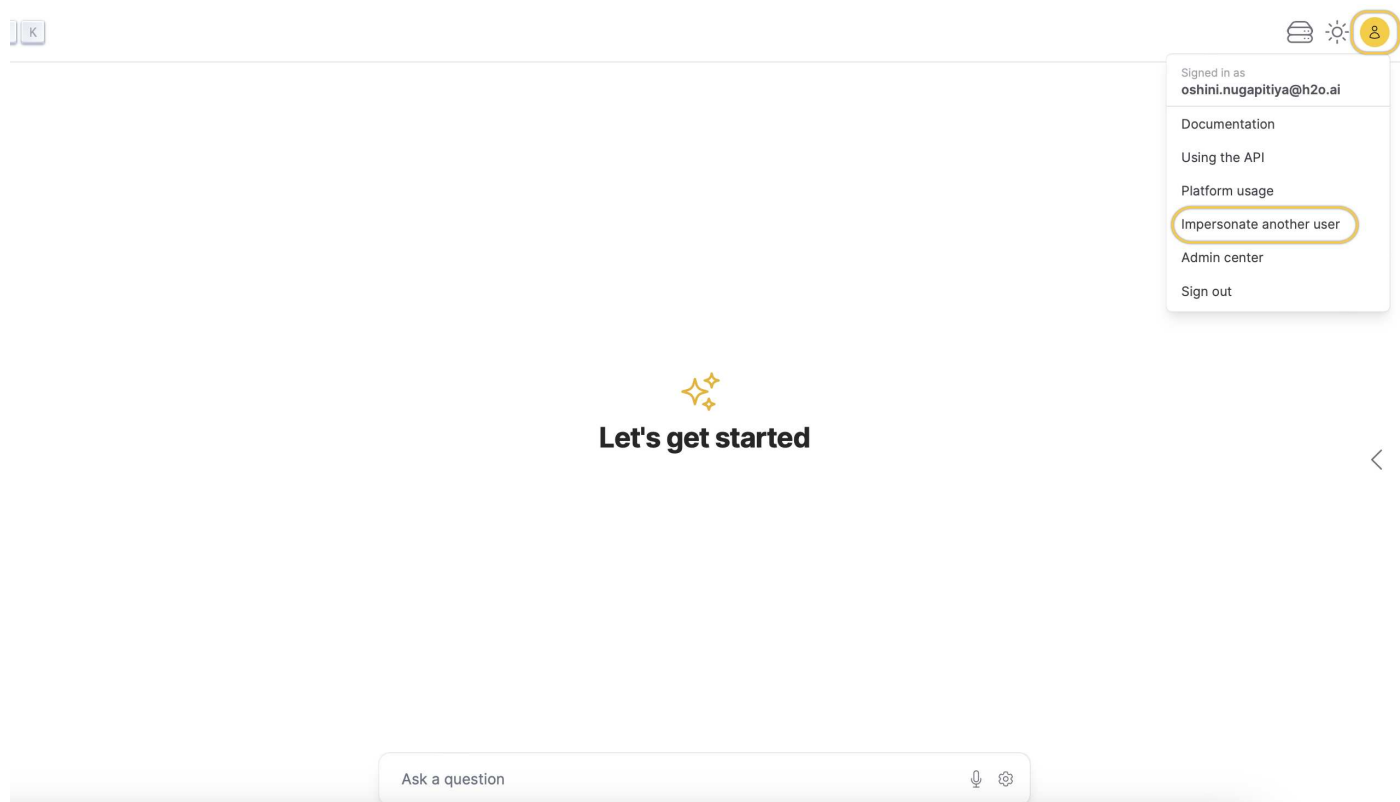
# Impersonate another user

## Overview

User impersonation lets you temporarily sign in as a different user. The [global API key](#) is used to impersonate a user for all API calls.

## Instructions

To impersonate a user, follow these steps: 1. On **Enterprise h2oGPTe**, click account\_circle **Account Circle**. 2. Select **Impersonate another user**.



3. Enter the global API key for the user you want to impersonate. **Note:** Create a new global API key on the **API Key** page. For more information, see [APIs](#).

4. Click **Impersonate**.

## APIs

### Create an API Key

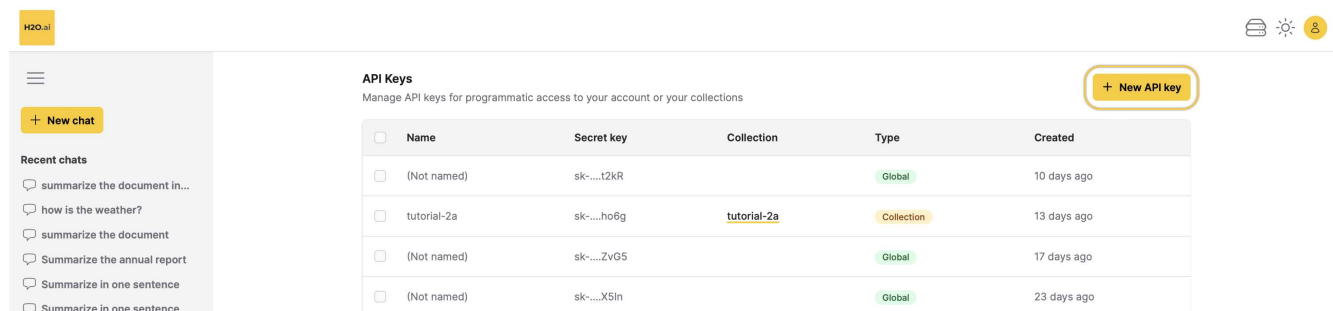
#### Overview

You can create an application programming interface (API) key to obtain programmatic access to your Enterprise h2oGPTe account.

#### Instructions

The following steps describe how to create an API Key:

1. On **Enterprise h2oGPTe**, click `account_circle` **Account Circle**.
2. Select **Using the API**.
3. Click **+ New API Key**.



4. (Optional) In the **Key name** box, enter a name for the API Key.
5. (Optional) In the **Restrict to collection** list, select a Collection. **Note:** Specifying a Collection enables you to create a Collection-specific API Key by indicating the Collection you want to link it with. For more information, see [Types of API keys](#).
6. Click **Generate new key**. **Caution:** Do not share your API Key with others or expose it within the browser or other client-side code.

**Note:** Refer to the [Python Client guide](#) for detailed instructions on how to utilize your API Key with the Enterprise h2oGPTe Python client.

### Delete an API Key

#### Overview

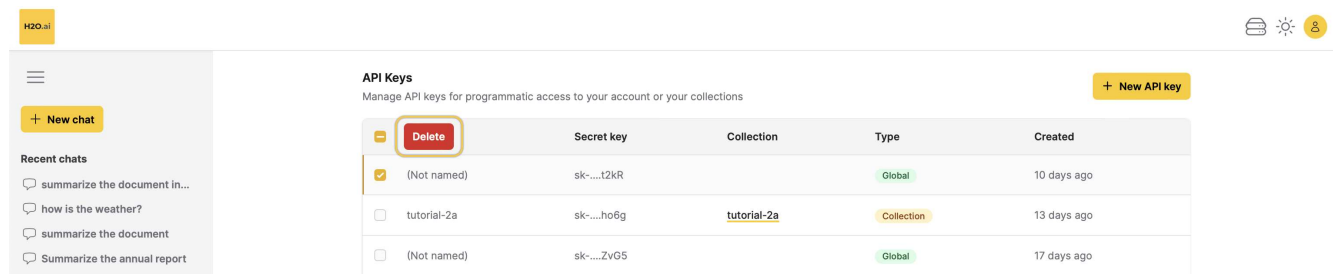
Once an API Key is no longer in use, you can delete it.

**Caution:** Deleting an API Key removes the API Key from the servers permanently. Once an API Key is deleted, there is no way to undo the action or recover the deleted API Key.

#### Instructions

To delete an API Key, consider the following steps:

1. On **Enterprise h2oGPTe**, click `account_circle` **Account Circle**.
2. Select **Using the API**.
3. In the **API Keys** table, locate the checkbox corresponding to the API Key you wish to delete, then select it.
4. Click **Delete**.



5. In the **Are you sure?** card, click **Delete**.

## Types of API Keys

In Enterprise h2oGPTe there are two types of API Keys:

- [Global API Keys](#)
- [Collection-specific API Keys](#)

### Global API Keys

If a Collection is not specified when creating a new API Key, that key is considered to be a global API Key. Use global API Keys to grant full user impersonation and system-wide access to all of your work. Anyone with access to one of your global API Keys can create, delete, or interact with any of your past, current, and future Collections, Documents, Chats, and settings.

### Collection-specific API Keys

Use Collection-specific API Keys to grant external access to only Chat with a specified Collection and make related API calls to it. Collection-specific API keys do not allow other API calls, such as creation, deletion, or access to other Collections or Chats.

## Jobs

### View a Job

#### Overview

A Job in Enterprise h2oGPTe signifies a single/batch of crawling, indexing or document processing tasks. In particular, the following tasks are referred to as a Job:

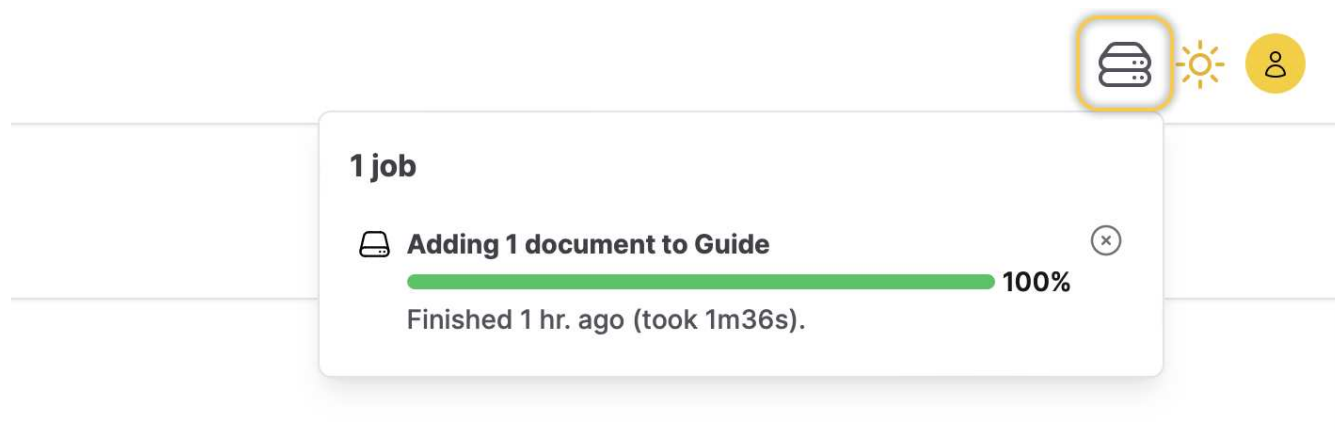
- Ingest plain text
- Ingest a document from the file system
- Ingest from cloud storage
- Upload Files
- Ingest website or files from URL
- Convert files to PDF
- Index Documents
- Update Collection statistics
- Delete Documents
- Delete Documents from a Collection
- Delete Collections
- Import an existing Document to a Collection
- Import all Documents from a Collection to another Collection
- Process a Document (Summarize/Extract/Process)

After starting a Job in Enterprise h2oGPTe, you can access its status, including the percentage of completion.

#### Instructions

To view a Job, consider the following step:

1. In **Enterprise h2oGPTe**, click **Jobs**.



### Terminate a Job

#### Overview

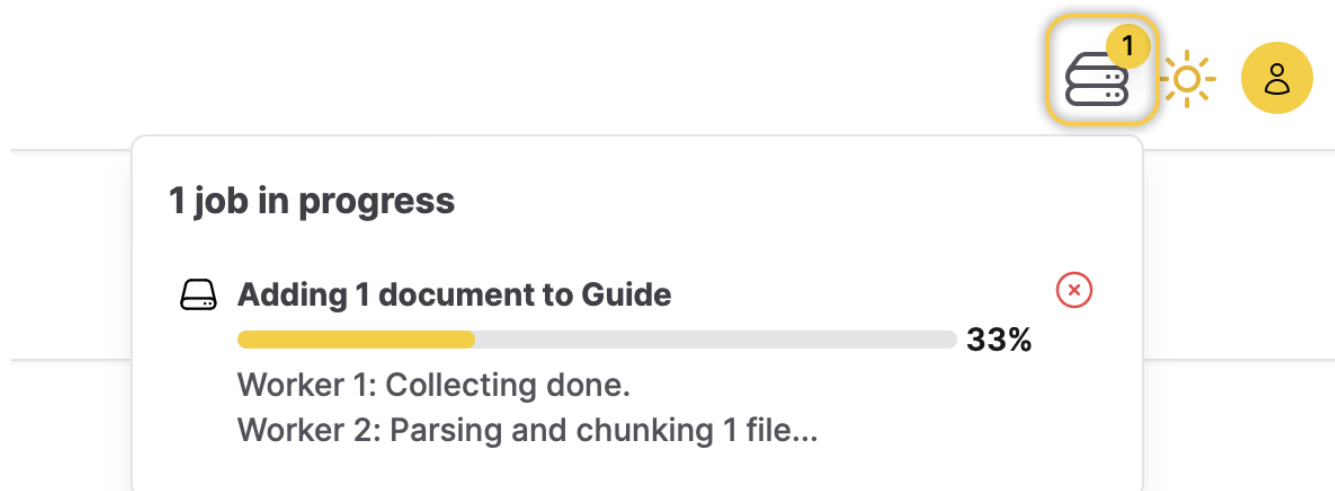
Right after starting a [Job](#) in Enterprise h2oGPTe, you can terminate it.

**Caution:** Terminating a Job removes the Job from the Enterprise h2oGPTe servers permanently. Once a Job is terminated, there is no way to undo the action or recover the deleted Job.

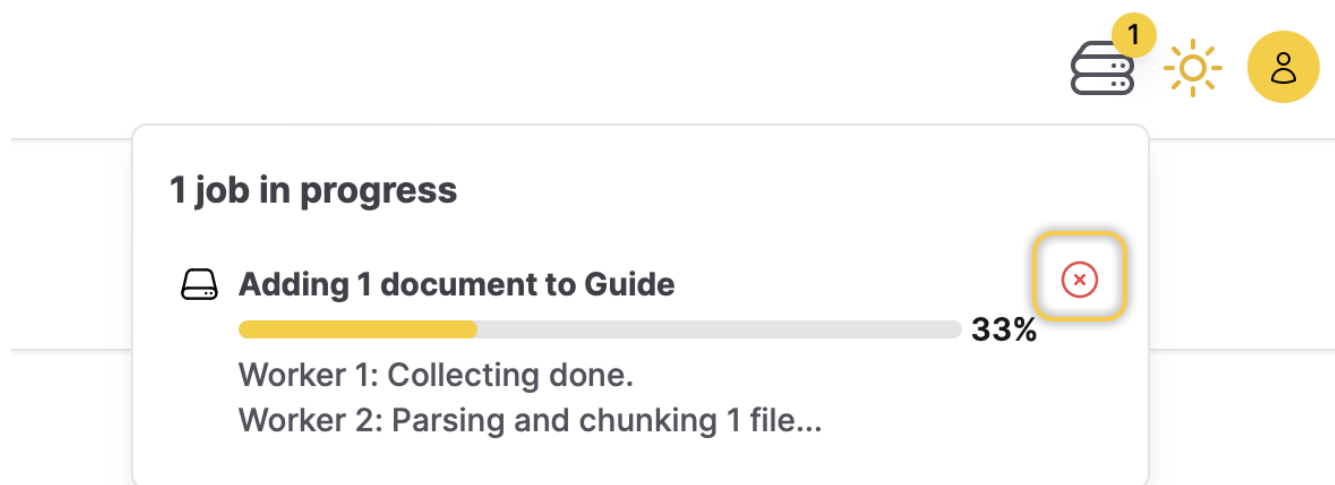
#### Instructions

To terminate a Job right after starting the Job, consider the following steps:

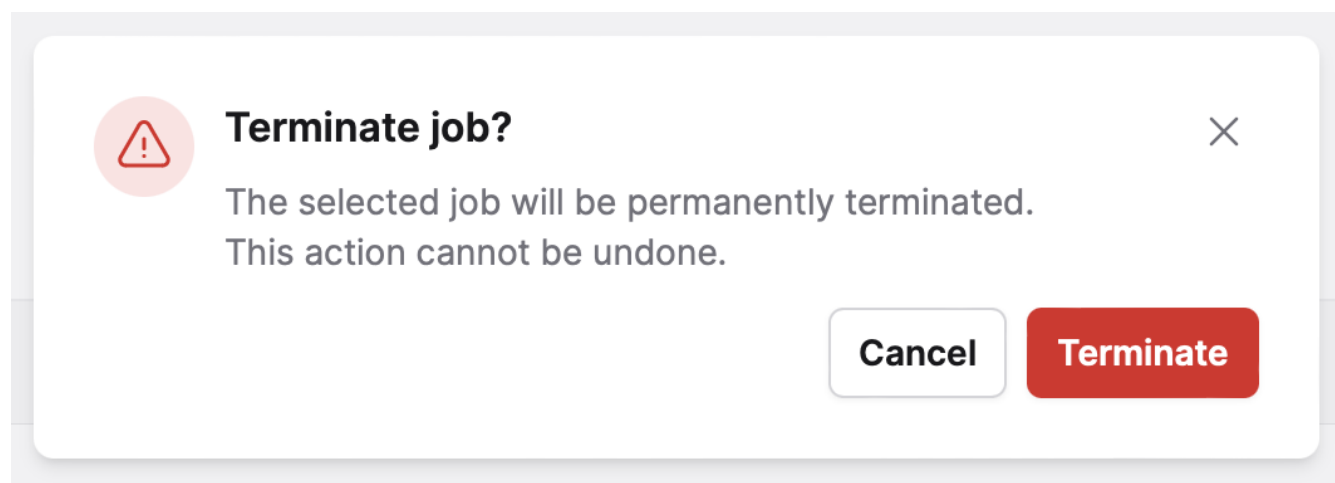
1. In **Enterprise h2oGPTe**, click **Jobs**.



2. In the [N] **job in progress** card, click the **X** button for the Job you want to terminate.



3. In the **Terminate job?** card, click **Terminate**.





## Connectors

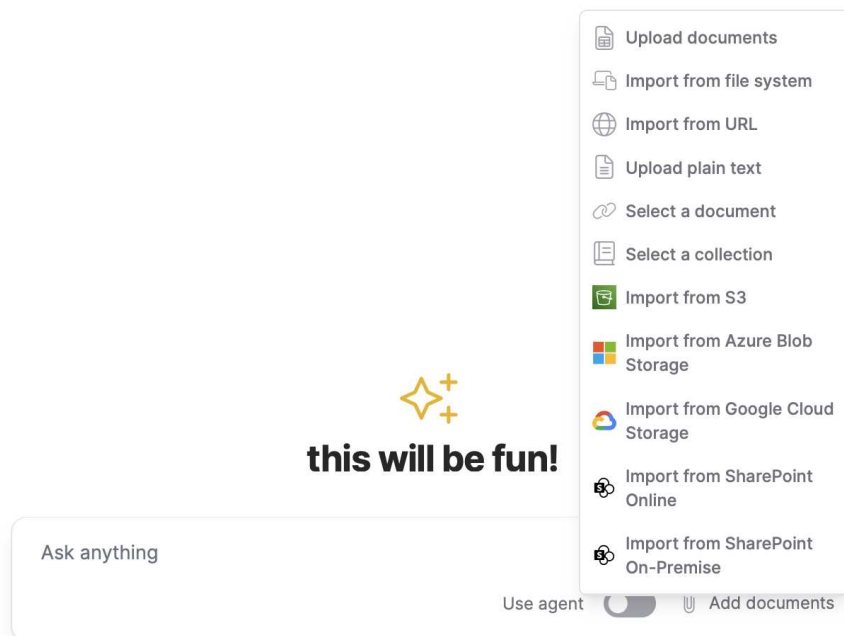
### Overview

**Enterprise h2oGPTe** provides a variety of connectors to integrate and import data from multiple sources. These connectors enable integration with cloud storage platforms, on-premise systems, and document repositories.

### Instructions

To view the available connectors in Enterprise h2oGPTe, consider the following steps:

1. In **Enterprise h2oGPTe**, click **+ New chat**.
2. Click `attach_file` **Attach file**.



### Note:

- To learn about each connector, see [Connectors](#).
- To learn how to add a document to a Collection, see [Add a Document\(s\) to a Collection](#).

## Connectors

- **File-based imports**
  - **Upload documents:** Upload files directly into the system.
  - **Import from file system:** Integrate with a local or networked file system to import data.
  - **Import from URL:** Provide a publicly accessible URL to fetch and process content directly.
  - **Upload plain text:** Manually enter or upload raw text content.
- **Document and collection management**
  - **Select a document:** Choose an existing document from the platform.
  - **Select a collection:** This option allows batch operations on grouped data sets. It allows you to work with predefined collections of documents.
- **Cloud Storage connectors**

- **Import from S3:** Integrate with Amazon S3 buckets for data import. Authentication and bucket access permissions are required.
- **Import from Azure Blob Storage:** Connect to Microsoft Azure's Blob Storage service to fetch and store data efficiently. Access and authentication are required.
- **Import from Google Cloud Storage:** Enable integration with Google Cloud Storage to retrieve data stored in buckets. The appropriate credentials are required.
- **SharePoint integration**
  - **Import from SharePoint Online:** Fetch documents and data from the cloud-hosted version of SharePoint. SharePoint Online credentials and permissions are required.
  - **Import from SharePoint On-premise:** Integrate with an on-premise SharePoint installation. Ensure access to the necessary credentials and permissions.

# Admin center

## Overview

The Admin Center page offers administrators an overview of recent [Jobs](#). A Job signifies a single/batch of crawling or indexing tasks.

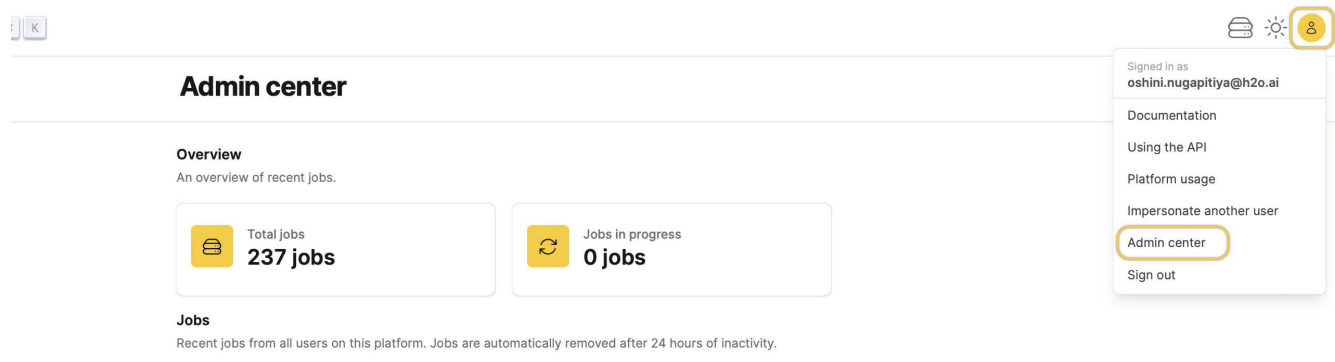
It shows the total number of Jobs and the count of Jobs in progress within the last 24 hours. The Jobs table lists all recent Jobs from all users on the platform.

**Note:** Jobs are automatically removed after 24 hours of inactivity.

## Instructions

To access the Admin center, consider the following steps:

1. In **Enterprise h2oGPTe**, click account\_circle **Account Circle**.
2. Select **Admin center**.




The Jobs table contains the following information:


- **User:** This column displays the email address of the user who created the Job.
- **Job:** This row describes a single or batch crawling or indexing task. For more information, see [Jobs](#).
- **Created:** This column indicates the number of hours since the Job was created.
- **Duration:** This column shows the Job's processing time.
- **Status:** This column displays the Job's completion status.
- **Errors:** This column lists any errors encountered during the Job processing.

### Admin center

Overview

An overview of recent jobs.

Total jobs  
**240 jobs**

Jobs in progress  
**0 jobs**

Jobs

Recent jobs from all users on this platform. Jobs are automatically removed after 24 hours of inactivity.

User	Job	Created	Duration	Status	Errors
sandil.ranasinghe@h2o.ai	Deleting collections	23 hr ago	0s	Completed	
sandil.ranasinghe@h2o.ai	Deleting collections	18 hr ago	1s	Completed	
sandil.ranasinghe@h2o.ai	Deleting collections	6 hr ago	0s	Completed	
sandil.ranasinghe@h2o.ai	Deleting collections	15 hr ago	1s	Completed	
sandil.ranasinghe@h2o.ai	Deleting collections	18 hr ago	1s	Completed	
sandil.ranasinghe@h2o.ai	Deleting collections	6 hr ago	1s	Completed	
sandil.ranasinghe@h2o.ai	Deleting collections	3 hr ago	1s	Completed	

### RBAC and Jobs

Within Role-Based Access Control (RBAC) and Jobs in Enterprise h2oGPTE, RBAC manages who can perform certain Jobs. For example, the following two actions are referred to as a Job:

- Add a Collection
- Delete a Collection

# h2oGPTe REST API: OpenAPI specification file

## Overview

The **h2oGPTe OpenAPI specification file** outlines the structure and functionality of the h2oGPTe REST API, detailing available endpoints, request and response formats, and authentication requirements.

- **OpenAPI specification:** Access the specification in YAML format:
  - [Download OpenAPI Spec \(YAML\)](#)
- **Interactive documentation:** Explore the API through an interactive interface, complete with live endpoint examples:
  - [Interactive API Docs \(Swagger UI\)](#)

Using the OpenAPI specification file for the h2oGPTe REST API, we have generated SDKs in multiple programming languages to help you quickly integrate with the API. Below are the available SDKs:

### 1. Python SDK

- [Download Python SDK](#)

### 2. JavaScript SDK

- [Download JavaScript SDK](#)

### 3. Go SDK

- [Download Go SDK](#)

## How we generated the SDKs

The available SDKs were developed using the OpenAPI Generator CLI, a robust tool designed to create client libraries directly from OpenAPI specifications.

### General CLI structure

```
pip install openapi-generator-cli==7.10.0
```

```
openapi-generator-cli generate \  
-i replace-with-openapi-spec-link \  
-g replace-with-language \  
-o replace-with-output-directory \  
--additional-properties=replace-with-additional-properties
```

- **-i:** The path or URL to the OpenAPI spec (for example, `openapi.yaml` or `https://example.com/openapi.yaml`).
- **-g:** The target programming language for the client (for example, `python`, `javascript`).
- **-o:** The output directory for the generated files (for example, `./python-client`).
- **--additional-properties:** Additional settings for client generation (for example, `usePromises=true` for JavaScript).

For more information about generating SDKs in other languages, visit the [OpenAPI Generator CLI](#) documentation.

# Python SDK

## Overview

With the “openapi-generator-cli” Python package and the OpenAPI specification file for the h2oGPTe REST API, we generated SDKs in multiple programming languages, enabling quick integration with the API. One of the SDKs generated is the Python SDK.

- Download Python SDK

## Steps to generate and test the Python SDK

1. Create a folder with the following name: `python-sdk`
2. Download the OpenAPI specification file of the h2oGPTe REST API and move it to the “python-sdk” folder: [Download api-spec.yaml](#).
3. In the “python-sdk” directory, set up a Python environment with Python 3.8 or later and install the [OpenAPI Generator CLI v7.10.0](#): `sh python3 -m venv venv source venv/bin/activate pip install openapi-generator-cli==7.10.0`
4. In the “python-sdk” directory, run the following command (CLI) to create the Python SDK using the OpenAPI Generator CLI: `sh openapi-generator-cli generate \ -i rest_api_spec_h2ogpte.yaml \ -g python \ -o sdk \ --additional-properties=packageName=h2ogpte_rest_client,packageVersion=1.6.2,projectName=h2ogpte-rest-client`  
**Note:** The CLI creates a folder named “sdk” containing the Python SDK.
5. In the “sdk” directory, install the dependencies of the Python SDK: `sh cd sdk pip install -r requirements.txt`
6. In the “sdk” directory, install the Python SDK:  
`sh pip install setuptools python setup.py install`
7. In the “sdk” directory, create the following file (script) to test the newly created Python SDK (in particular, let’s create a Collection to test the SDK):

```
“py title=“create_a_collection.py” showLineNumbers # This example with the Python SDK creates a new Collection
with the following # name and description (while testing whether the Python SDK was properly installed): # Name
= The name of my Collection # Description = The description of my Collection import h2ogpte_rest_client import
os from h2ogpte_rest_client.rest import ApiException from pprint import pprint

configuration = h2ogpte_rest_client.Configuration( # This line specifies the URL where Enterprise h2oGPTe is
hosted. # The address should be the location where the API key was created. host=“https://h2ogpte.genai.h2o.ai/”
)

configuration = h2ogpte_rest_client.Configuration( access_token=os.environ[“BEARER_TOKEN”] )

with h2ogpte_rest_client.ApiClient(configuration) as api_client: api_instance = h2ogpte_rest_client.Collections
Api(api_client)

collection_create_request = h2ogpte_rest_client.CollectionCreateRequest(
    name=“The name of my Collection”,
    description=“The description of my Collection”
)

try:
    api_response = api_instance.create_collection(collection_create_request)
    print(“The response of CollectionsApi->create_collection:\n”)
    pprint(api_response)
except Exception as e:
    print(“Exception when calling CollectionsApi->create_collection: %s\n” % e)

““
```

8. Before you can run the Python script, you need to export a global API key within the “sdk” directory: `sh export BEARER_TOKEN=“sk-XX”` **Note:** Global API key

- The value exported for the “BEARER\_TOKEN” should be a global API key.
- To learn more about what is a global API key and how to create one, see [APIs](#).

9. In the “sdk” directory, run the “create\_a\_collection.py” file:

```
python3 create_a_collection.py
```

The response of CollectionsApi->create\_collection:

```
Collection(id='091c6ab1-a031-4eb7-9e3e-7d0939911d00', name='The name of my Collection', description='The
```

# JavaScript SDK

## Overview

With the “openapi-generator-cli” Python package and the OpenAPI specification file for the h2oGPTe REST API, we generated SDKs in multiple programming languages, enabling quick integration with the API. One of the SDKs generated is the JavaScript SDK.

- Download JavaScript SDK

## Steps to generate and test the JavaScript SDK

1. Create a folder with the following name: `javascript-sdk`
2. Download the OpenAPI specification file of the h2oGPTe REST API and move it to the “javascript-sdk” folder: [Download api-spec.yaml](#).
3. Inside the “javascript-sdk” folder, set up a Python environment with Python 3.8 or later and install the [OpenAPI Generator CLI v7.10.0](#): `sh python3 -m venv venv source venv/bin/activate pip install openapi-generator-cli==7.10.0`
4. In the “javascript-sdk” directory, run the following command (CLI) to create the JavaScript SDK using the OpenAPI Generator CLI: `sh openapi-generator-cli generate \ -i rest_api_spec_h2ogpte.yaml \ -g javascript \ -o sdk \ --additional-properties=packageName=h2ogpte-rest-client,packageVersion=1.6.2,projectName=h2ogpte-rest-client,npmName=h2ogpte-rest-client,npmVersion=h2ogpte-rest-client` **Note:** The CLI creates a folder named “sdk” containing the JavaScript SDK.
5. To use the library locally without publishing to a remote npm registry, first install the dependencies: `cd sdk npm install`
6. Now, link the library globally in npm: `sh npm link`
7. In the “sdk” directory, create a Node.js project to test the JavaScript SDK: `sh cd .. mkdir my-javascript-sdk-project`
8. In the “my-javascript-sdk-project” directory, initiate a new Node.js project: `sh cd my-javascript-sdk-project npm init -y`
9. To use the link you just defined in the “sdk” folder, navigate to the directory where you want to use your `h2ogpte-rest-client`. For this example, we will run it in the `my-javascript-sdk-project` directory: `sh npm link /path/to/<THE_SDK_FOLDER>`
10. Now, build the module in the “sdk” directory: `sh cd .. cd sdk npm run build`
11. In the “my-javascript-sdk-project” directory, create the following file to test the newly created JavaScript SDK (in particular, let’s create a Collection to test the SDK):

```
“javascript title=“create-a-collection.mjs” showLineNumbers // This example with the JavaScript SDK creates a
new Collection with the following // name and description (while testing whether the Python SDK was properly
installed): // Name = The name of my Collection // Description = The description of my Collection

let defaultClient = H2ogpteRestClient.ApiClient.instance;

let bearerAuth = defaultClient.authentications[‘bearerAuth’];

// The access token should be a global API key. bearerAuth.accessToken = “sk-
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX”

let apiInstance = new H2ogpteRestClient.CollectionsApi(); let collectionCreateRequest = new H2ogpteRestClient.Col-
lectionCreateRequest( “The name of my Collection”, // name parameter “The description of my Collection” //
description parameter );

apiInstance.createCollection(collectionCreateRequest, (error, data, response) => { if (error) { console.error(error); }
else { console.log(‘API called successfully. Returned data:’ + data); } }); ““
```

**Note:** Access token



- All URIs are relative to <https://h2ogpte.genai.h2o.ai/>. The global API key must be created on this platform
  - To learn more about what is a global API key and how to create one, see [\[APIs\]\(#apis\)](#).
12. In the “my-javascript-sdk-project” directory, run the JavaScript file: 

```
sh cd .. cd my-javascript-sdk-project node create-a-collection.mjs sh
```

 API called successfully. Returned data: [object Object]

## Go SDK

### Overview

With the “openapi-generator-cli” Python package and the OpenAPI specification file for the h2oGPTe REST API, we generated SDKs in multiple programming languages, enabling quick integration with the API. One of the SDKs generated is the Go SDK.

- Download Go SDK

### Steps to generate and test the Go SDK

1. Create a Go project named “my-project” and initiate it: `sh mkdir my-project cd my-project go mod init my-project`
2. Download the OpenAPI specification file of the h2oGPTe REST API and move it to the “my-project” folder: [Download api-spec.yaml](#).
3. In the “my-project” directory, set up a Python environment with Python 3.8 or later and install the [OpenAPI Generator CLI v7.10.0](#): `sh python3 -m venv venv source venv/bin/activate pip install openapi-generator-cli==7.10.0`
4. In the “my-project” directory, run the following command (CLI) to create the Go SDK using the OpenAPI Generator CLI: `sh openapi-generator-cli generate \ -i rest_api_spec_h2ogpte.yaml \ -g go \ -o h2ogpte_rest_client \ --additional-properties=packageName=h2ogpte_rest_client,packageVersion=1.6.2,projectName=h2ogpte_rest_client,goModule`
5. In the “my-project” directory, create a work file to add the “h2ogpte\_rest\_api” module: `sh go work init go work use ./h2ogpte_rest_client`
6. In the “my-project” directory, create the following file to test the newly created Go SDK (in particular, let’s create a Collection to test the SDK):

```
touch create_a_collection.go
```

```
“py title=“create_a_collection.go” showLineNumbers package main
```

```
import ( “context” “fmt” “os” openapiclient “github.com/GIT_USER_ID/GIT_REPO_ID/h2ogpte_rest_client” )
```

```
func main() { collectionCreateRequest := *openapiclient.NewCollectionCreateRequest(“The name of my Collection”, “The description of my Collection”) // CollectionCreateRequest |
```

```
configuration := openapiclient.NewConfiguration()
```

```
configuration.AddDefaultHeader(“Authorization”, “Bearer sk-XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

```
apiClient := openapiclient.NewAPIClient(configuration)
```

```
resp, r, err := apiClient.CollectionsAPI.CreateCollection(context.Background()).CollectionCreateRequest
```

```
if err != nil {
```

```
    fmt.Fprintf(os.Stderr, “Error when calling `CollectionsAPI.CreateCollection`: %v\n”, err)
```

```
    fmt.Fprintf(os.Stderr, “Full HTTP response: %v\n”, r)
```

```
}
```

```
// response from `CreateCollection`: Collection
```

```
fmt.Fprintf(os.Stdout, “Response from `CollectionsAPI.CreateCollection`: %v\n”, resp)
```

```
} “ Note: API key
```

- The value of the Authorization key should contain a global API key. The API key should be prefixed with Bearer, followed by the API key. For example: `configuration.AddDefaultHeader(“Authorization”, “Bearer sk-XXX`
- Make sure to replace `sk-XXX` with your actual API key.
- To learn more about what is a global API key and how to create one, see [APIs](#).
- All URIs are relative to `https://h2ogpte.genai.h2o.ai/`. The global API key must be created on this platform to ensure successful requests.

7. In the “my-project” directory, run the “create\_a\_collection.go” file: 

```
sh go run create_a_collection.go sh
```

  
Response from `CollectionsAPI.CreateCollection`: 

```
&{e6926802-9529-419f-9ddd-cbb262137a8a The
name of my Collection The description of my Collection BAAI/bge-large-en-v1.5 0 0 2024-12-11
23:43:19.972726 +0000 UTC 0 false sergio.perez@h2o.ai 0 <nil>}
```

## Key terms

Enterprise h2oGPTe uses several key terms across its documentation, and each, in turn, is explained on this page.

### Collection

A Collection refers to a group of related Documents. A Collection lets you aggregate documents in one location. You can utilize Collections to group particular sets of material (content) to later explore individually through Chats (asking questions to a Collection).

Enterprise h2oGPTe supports Retrieval Augmented Generation (RAG) when getting responses from an LLM, which allows for contextualizing the question to the LLM with information from documents, audio transcriptions, and other data. Users can create one or more data Collections from which they want to get answers or generate new content. When a user interacts with an LLM, the user's prompt is compared with the Collection of documents to find similar chunks of information. This information is then sent to the LLM.

There are many strategies for importing and creating Collections so that you get the best responses for your use case. For more information, see [Collections usage overview](#).

### Job

A Job signifies a single/batch of crawling or indexing tasks. In particular, the following tasks are referred to as a Job:

- Ingest plain text
- Ingest a Document from the file system
- Ingest from cloud storage
- Ingest (add) a Document from upload
- Ingest (crawl) a website
- Conver files to a PDF
- Index Document(s)
- Update a Collection's statistics
- Delete a Document(s)
- Delete a Document(s) from a Collection
- Delete a Collection(s)
- Import a stored Document to a Collection
- Import all Document(s) from a Collection to another Collection
- Summarize a Document
- Process a Document(s)

### Document

A Document refers to one of your imported files to Enterprise h2oGPTe (for example, a PDF or web page).

### Chat

A Chat session is an interaction between you and Enterprise h2oGPTe that consists of a series of prompts and answers.

### API Key

An application programming interface (API) key is a unique identifier to authenticate to the h2oGPTe API.

### Extractors

Extractors, defined by JSON schemas, play a crucial role in document AI, transforming unstructured document content into structured, actionable data. With Extractors, you can effortlessly retrieve information from any document—whether it's a CV, invoice, Form 10-K, or scanned image—without the need for complex setups or annotations. Just specify the data you need with an intuitive JSON schema builder (UI), upload your documents, and receive structured data instantly.

### PII Detection

Personally Identifiable Information (PII) detection is the process of recognizing and classifying sensitive data within a dataset that can be used to identify a specific individual. This includes information like social security numbers, credit

card numbers, bank account numbers, and passport numbers. Non-sensitive PII includes information like names, addresses, and phone numbers.

## Evaluators

Evaluators are tools and metrics used to assess the performance and quality of large language models (LLMs) and Retrieval-Augmented Generation (RAG) models. They also evaluate a Collection's performance, reliability, security, fairness, and effectiveness.

## Vector database

Enterprise h2oGPTe uses a vector database to store document chunks for future retrieval. The following section describes the currently supported options.

### Embedded Vex vector database

This option is the default embedded vector database that ships out of the box with the software.

Vex == uvector + hnswlib + sqlite

[Hnswlib \(fast approximate nearest neighbor search\)](#) is projected to encounter performance degradation at roughly 10M-50M entries.

**Note:** A separate (hnswlib+sqlite) instance is created per Collection, so those limits apply at the Collection level.

### External vector databases

The following external vector databases are currently experimentally supported:

- ElasticSearch
- Milvus
- Qdrant
- Redis

Vector database == uvector + [ ElasticSearch / Milvus / Qdrant / Redis ]

**Note:** Please contact us if your preferred vector database is not listed above (these can be added based on customer interest).

## Major releases

Version	Release time
v1.5.x	July/August 2024
v1.4.x	April/May 2024
v1.3.x	February/March 2024

## FAQs

Enterprise h2oGPTE is an AI-powered search assistant that helps you find answers to questions about your documents, websites, and workplace content - powered by H2O LLM.

- With Enterprise h2oGPTE, finding what you need has never been easier
- Ask a question, and Enterprise h2oGPTE will provide relevant results from across provided materials
- Whether you're looking for a specific document or trying to understand a complex concept, Enterprise h2oGPTE can help you save time and effort

---

The following sections provide answers to frequently asked questions. If you have additional questions, please send them to [cloud-feedback@h2o.ai](mailto:cloud-feedback@h2o.ai).

## General

### How does Enterprise h2oGPTE handle embedded images in PDFs?

Enterprise h2oGPTE supports vision and language models out of the box and automatically determines which to use to answer a user's question. When using language models, optical character recognition (OCR) is used to extract text from any images. When using vision models, each page of a document is viewed as an image by the model.

### Does Enterprise h2oGPTE have an API?

Yes. You can interact with the **h2oGPTE API** through its **Python Client Library** and **TypeScript API**. These libraries simplify making API requests and handling responses, allowing you to integrate the API's functionality seamlessly into your Python or TypeScript applications.

Sign up/in at **Enterprise h2oGPTE** and create an API key to use the client libraries. To learn more, see [APIs](#).

### Which Embedding Models are supported in Enterprise h2oGPTE?

When creating a new collection of data, users can choose from [bge-large-en-v1.5](#), [bge-m3](#), or [instructor-large](#) as the embedding model for their collection, with [bge-large-en-v1.5](#) as the default. This decision is usually based on the language of their data and their users.

[bge-large-zh-v1.5](#), [multi-lingual-e5-large](#), or [instructor-xl](#) are also supported but need to be enabled by an admin.

We regularly update our platform with the latest embedding models to ensure optimal performance and state-of-the-art capabilities.

### What is the Vector Database for Enterprise h2oGPTE (1) and can I customize it (2)?

1. The Vector Database for Enterprise h2oGPTE is called "Vex" and it is custom built by H2O.ai. Vex is a standalone vector "database" specifically designed for running similarity searches for Enterprise h2oGPTE. It serves as a storage and retrieval system for vector and text data.

The current nearest neighbor algorithm used in Vex is [hnsplib](#). Indexes are stored in hnsplib's native format and are created or loaded by the server on demand. This allows for efficient and fast similarity searches. The underlying text representation for each vector is stored in **SQLite**.

Vex is architecturally similar to Chroma, which also utilizes hnsplib for thread-safety during index-writes. However, Vex includes several performance improvements compared to Chroma:

1. It does not include unnecessary bindings to Clickhouse and DuckDB.
  2. It uses barebones Uvicorn (the core of FastAPI) instead of FastAPI, eliminating unnecessary overhead.
  3. It uses MsgPack instead of JSON for communication, reducing the bandwidth required for each request. This is particularly beneficial for compressing numpy arrays and UUIDs.
  4. It minimizes allocations in the query fast-path, avoiding unnecessary steps like Pydantic, and directly calling the library for binary deserialization.
2. Customizability of Vex is limited, as it primarily focuses on delivering efficient similarity searches using its existing features. However, Enterprise h2oGPTE provides additional connectivity options to other Vector and full-text databases, allowing for more customization possibilities.



Starting from **Enterprise h2oGPTe 1.3.8**, users can connect Enterprise h2oGPTe to **hnswlib**, **Elastic Search**, **Milvus**, **Redis**, and **Qdrant** for storing vector data, and **SQLite**, **Redis** and **Elastic Search** for storing text data. By default, **HNSW** and **SQLite** are used, other external databases are currently in experimental mode.

**Note:** To learn how to customize the Vector Database, contact your admin.

## Collections

### Can Enterprise h2oGPTe refer to several Collections to answer a query?

No. Enterprise h2oGPTe is designed to access a specific Collection at a time, allowing users to ask detailed questions and receive precise answers based on the targeted data contained within that Collection. This focused approach ensures accuracy and relevance in the responses provided and ensures data security in the platform.

### What file types (documents) does Enterprise h2oGPTe Collections support?

You can upload the following file types (documents) to a Collection:

## Documents

### Are uploaded documents cleared after the user session, or do they persist?

The uploaded documents are persisted until the user or admin explicitly deletes the document.

### How are entitlements managed so that users are not querying each other's documents or the model is not picking chunks from other documents?

Documents are private and accessible only to the user who uploaded them or via a publicly shared collection. An authenticated user client can only query resources they have access to.

**Note:** If you want to make documents public to all users accessing the system, change the collection from Private to Public (either in the UI or in the Python API). For more information, see [Make a Collection public](#).

## Chats

### Can I chat with a Collection in any language?

Yes. For non-English chatting or documents, we recommend creating your collection with the **Multilingual (bge-m3)** embedding model which will be applied automatically to ingested data and user queries. You can also change the the Prompt Template to one of the 15+ supported languages for Retrieval Augmented Generation or create your own template for new languages or use cases.