

Setting up Cloudera AI Studios

Date published: 2020-07-16

Date modified: 2025-10-31



Legal Notice

© Cloudera Inc. 2025. All rights reserved.

The documentation is and contains Cloudera proprietary information protected by copyright and other intellectual property rights. No license under copyright or any other intellectual property right is granted herein.

Unless otherwise noted, scripts and sample code are licensed under the Apache License, Version 2.0.

Copyright information for Cloudera software may be found within the documentation accompanying each component in a particular release.

Cloudera software includes software from various open source or other third party projects, and may be released under the Apache Software License 2.0 (“ASLv2”), the Affero General Public License version 3 (AGPLv3), or other license terms. Other software included may be released under the terms of alternative open source licenses. Please review the license and notice files accompanying the software for additional licensing information.

Please visit the Cloudera software product page for more information on Cloudera software. For more information on Cloudera support services, please visit either the Support or Sales page. Feel free to contact us directly to discuss your specific needs.

Cloudera reserves the right to change any products at any time, and without notice. Cloudera assumes no responsibility nor liability arising from the use of products, except as expressly agreed to in writing by Cloudera.

Cloudera, Cloudera Altus, HUE, Impala, Cloudera Impala, and other Cloudera marks are registered or unregistered trademarks in the United States and other countries. All other trademarks are the property of their respective owners.

Disclaimer: EXCEPT AS EXPRESSLY PROVIDED IN A WRITTEN AGREEMENT WITH CLOUDERA, CLOUDERA DOES NOT MAKE NOR GIVE ANY REPRESENTATION, WARRANTY, NOR COVENANT OF ANY KIND, WHETHER EXPRESS OR IMPLIED, IN CONNECTION WITH CLOUDERA TECHNOLOGY OR RELATED SUPPORT PROVIDED IN CONNECTION THEREWITH. CLOUDERA DOES NOT WARRANT THAT CLOUDERA PRODUCTS NOR SOFTWARE WILL OPERATE UNINTERRUPTED NOR THAT IT WILL BE FREE FROM DEFECTS NOR ERRORS, THAT IT WILL PROTECT YOUR DATA FROM LOSS, CORRUPTION NOR UNAVAILABILITY, NOR THAT IT WILL MEET ALL OF CUSTOMER’S BUSINESS REQUIREMENTS. WITHOUT LIMITING THE FOREGOING, AND TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW, CLOUDERA EXPRESSLY DISCLAIMS ANY AND ALL IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY, QUALITY, NON-INFRINGEMENT, TITLE, AND FITNESS FOR A PARTICULAR PURPOSE AND ANY REPRESENTATION, WARRANTY, OR COVENANT BASED ON COURSE OF DEALING OR USAGE IN TRADE.

Contents

- AI Studios Overview (Technical Preview)..... 4**
- Managing RAG Studio.....4**
 - RAG Studio Overview..... 4
 - Key features for RAG Studio.....5
 - Launching RAG Studio within a project..... 6
 - Configuring RAG Studio.....6
- Managing Fine Tuning Studio..... 8**
 - Fine Tuning Studio Overview.....8
 - Key Features for Fine Tuning Studio..... 9
 - Launching Fine Tuning Studio within a project.....9
- Managing AI Studios.....10**
 - Accessing the AI Studios Catalog.....10
 - Viewing AI Studio Deployment Status.....11
 - Accessing the embedded AI Studios application..... 11
 - Redeploying or Resuming AI Studios..... 11
 - Deleting an AI Studio..... 11
- Host names and endpoints required for AI Studios.....12**

AI Studios Overview (Technical Preview)

Cloudera AI Studios is a comprehensive suite of low-code tools designed to simplify the development, customization, and deployment of generative AI solutions within enterprises. This suite empowers organizations to operationalize AI workflows quickly and efficiently by leveraging real-time enterprise data. It provides scalable, cost-effective, and trustworthy AI applications, fostering seamless collaboration between business and IT teams.



Note: This feature is in Technical Preview and not recommended for production deployments. Cloudera recommends that you try this feature in test or development environments.

Key Components of Cloudera AI Studios:

- **RAG Studio:**
 - Purpose: Accelerate the creation of secure, enterprise-grade Retrieval-Augmented Generation (RAG) applications and simple chatbots.
 - Capabilities: Enables organizations to build advanced, context-aware AI chatbots that utilize enterprise data while maintaining data security and compliance.
- **Fine-tuning Studio:**
 - Purpose: Simplify the customization and optimization of large language models (LLMs).
 - Capabilities: Allows users to fine-tune pre-trained models to suit specific business requirements, eliminating the need for extensive technical expertise or computational resources.
- **Synthetic Data Studio:**
 - Purpose: Generate scalable, privacy-compliant synthetic datasets for enterprise use cases.
 - Capabilities: Provides tools to create data that mimics real-world datasets while ensuring privacy and compliance, supporting AI model training and testing with minimal risk.
- **Agent Studio:**
 - Purpose: Design and deploy sophisticated multi-agent AI workflows tailored for enterprise automation.
 - Capabilities: Facilitates the orchestration of AI agents to perform complex tasks collaboratively, streamlining operations, and enhancing efficiency across the organization.

By leveraging Cloudera AI Studios, enterprises can rapidly build and deploy AI-powered applications, enabling innovation, improving decision-making, and driving business value, all while maintaining scalability, compliance, and trustworthiness in their AI initiatives.

Managing RAG Studio

Manage the entire lifecycle of LLMs with the help of RAG Studio, empowering you to optimize AI workloads while enhancing the accuracy and efficiency of your models.

RAG Studio Overview

Retrieval-Augmented Generation (RAG) Studio is a no-code application for creating RAG chatbots that combine the power of retrieval and generation to provide accurate and efficient responses.

Using the Cloudera platform, you can build and deploy RAG chatbots in minutes, without requiring extensive technical expertise. Designed for accessibility, RAG Studio bridges the gap between business and IT teams, fostering collaboration on AI projects. The Retrieval-Augmented Generation studio features a secure, context-aware chatbot that leverages enterprise documents and real-time data ingestion to deliver accurate and efficient responses.

Built on open-source solutions, RAG Studio provides a flexible and customizable framework for developing and deploying chatbots. It allows you to leverage the latest advancements in AI and machine learning while retaining full control over your data and workflows.

RAG Studio is part of the Cloudera AI product portfolio and offers the added benefit of being deployable on premises, providing a secure and controlled environment for developing and deploying AI-powered chatbots. This allows you to integrate with existing data infrastructure and workflows, while also maintaining control over your data.

Key features for RAG Studio

RAG Studio allows you to create an application with a conversational interface (chatbot) that can be used to converse with a connected large language model, and comes with the functionality to enable retrieval-augmented generation to improve the performance and accuracy of the model's responses.

Core Features

- **No-Code Application:** RAG Studio is a no-code application, empowering users to create and deploy RAG chatbots without requiring extensive technical expertise.
- **Retrieval-Augmented Generation:** RAG Studio combines the power of retrieval and generation to deliver accurate and efficient responses, leveraging the strengths of both approaches.
- **Secure and Context-Aware:** Cloudera's chatbot is designed to be secure and context-aware, utilizing enterprise documents and real-time data ingestion to provide precise and relevant responses. Each answer synthesized from your enterprise documents will be automatically scored for faithfulness and relevance and provide reference documents, for easy and reliable human use.
- **Open-Source and Customizable:** RAG Studio is built on open-source solutions, offering a flexible and customizable framework for developing and deploying chatbots that meet your unique needs.

Deployment and Integration

- **On-Premises Deployment:** RAG Studio can be deployed on premises, providing a secure and controlled environment for developing and deploying AI-powered chatbots.
- **Integration with Existing Infrastructure:** RAG Studio seamlessly integrates with existing data infrastructure and workflows, allowing you to maintain control over your data while leveraging the power of AI.

Accessibility and Collaboration

- **Accessibility:** RAG Studio is designed for accessibility, bridging the gap between business and IT teams and fostering collaboration on AI projects.

Advanced Capabilities

- **Optimized with Vector Search:** Our platform is optimized with vector search, enabling fast and efficient retrieval of relevant information.
- **Automatic Evaluation:** Our platform includes automatic evaluation, allowing you to assess the quality and accuracy of your chatbot's responses.
- **Scalable NiFi Ingestion Pipelines:** Our platform supports scalable NiFi ingestion pipelines, enabling you to handle large volumes of data and ensure seamless integration with your existing infrastructure.
- **Tool Calling:** RAG Studio enables seamless integration with large language models (LLMs) that support function or tool calling, including OpenAI, Azure OpenAI, Amazon Bedrock, and Cloudera AI Inference service. It also provides connectivity to Model Context Protocol (MCP) tools, which can be hosted by RAG studio or used externally.

Foundation Model Integration

- **Seamless Connection:** RAG Studio seamlessly connects to proprietary foundational models or fine-tuned models (if Cloudera AI Inference service is integrated) with your organizational knowledge, empowering users to leverage the right model for every task and switch easily between them.

Launching RAG Studio within a project

RAG Studio is compatible with Cloudera AI Inference service, AWS Bedrock, Azure OpenAI, and Open AI enabling you to select different LLM and embedding models tailored to your specific needs.

Before you begin

RAG Studio integrates with three major enterprise inference services:

- AWS Bedrock
- Cloudera AI Inference Service
- Azure OpenAI



Note: The Reranking model option (Chat settings) is not available with Azure OpenAI.

- Host names: For air-gapped installations that use a proxy setup, it is essential to whitelist the necessary URLs in your firewall rules. For a list of hostnames to whitelist, see [Host names and endpoints required for AI Studios](#).



Note:

Note that in an air-gapped environment, the Cloudera AI Inference service may be the sole option available.

Procedure

1. In the Cloudera console, click the **Cloudera AI** tile.

The Cloudera AI Workbenches page displays.

2. Click on the name of the workbench.

The workbenches Home page displays.

3. Click Projects, and then click New Project to create a new project.

In the left navigation pane, the new **AI Studios** option is displayed.

4. Click AI Studios.

5. Click the Launch button in the **RAG Studio** box. The **Configure Studio: RAG Studio** page is displayed.

6. Select the Runtime version.

7. Click Launch AI Studio.

The **AI Studio Setup Steps** page is displayed.

After launching, you can view the list of tasks being executed as part of the AI studio deployment.

Configuring RAG Studio

After launching RAG Studio, an initial configuration is required, with optional settings available for additional customization. Without this configuration, the Studio will not be able to access any AI models.

Procedure

1. In the Cloudera console, click the **Cloudera AI** tile.

The Cloudera AI Workbenches page displays.

2. Click on the name of the workbench.

The workbenches Home page displays.

3. Click Projects, and then select the required Project.

In the left navigation pane, the new **AI Studios** option is displayed.

4. Click AI Studios and select **RAG Studio**.

The **RAG Studio** page is displayed.

The **Settings** page is displayed when opening up the RAG Studio for the first time. If the **Settings** page is not displayed, click on **Settings** in the top-right corner. The following settings are available:

- a) Enable the Enhanced PDF Processing option for better text extraction, however, note that this feature requires at least one GPU and at least 16GB of RAM and can significantly slow down document parsing.
- b) Select one of the following Metadata Database storage options.
 - Embedded H2 database (default): Use an Embedded H2 database for storing metadata information.
 - External PostgreSQL database: Store metadata in an external PostgreSQL database.
- c) Select one of the following File Storage options.
 - Project Filesystem (Default): Use the Cloudera AI Project filesystem for file storage.
 - AWS S3: Select an existing Amazon S3 bucket and provide the bucket name and prefix to be used for all S3 paths.
- d) Select one of the following Vector Database options.
 - Embedded Qdrant database (Default): Use Qdrant locally as a vector store database.
 - Cloudera Semantic Search: Configure an existing CSS host by providing the host details, namespace and optionally, a username and password for authentication.
 - ChromaDB database: To use it as a vector store, you can either run it locally by setting the host to localhost or provide external endpoint details for remote access.
- e) Select one of the following Model Provider options.

The following options are available:

- Cloudera AI: No authentication required, but you will need to obtain the domain name of the Cloudera AI Inference service. Note, that the domain must reside within the same Cloudera environment as the Cloudera AI Workbench where the studio is operating. For more details, see [Preparing to interact with the Cloudera AI Inference service API](#).



Note: For Cloudera AI Workbench 2.0.50-b68 or higher versions, the domain name is optional if the service is hosted within the same environment. However, if the LLM, embedding or reranking endpoints are hosted in a different environment, both the domain name and CDP token are required for proper functionality.

- AWS Bedrock: Requires authentication:
 - AWS Region: Choose the AWS region to use.
 - AWS Access Key ID: Provide the Access Key ID for authentication.
 - AWS Secret Access Key: Provide the Secret Access Key for authentication.
- Azure OpenAI: Requires configuration and authentication:
 - Azure OpenAI Endpoint: Find the endpoint of the Azure OpenAI service in the Azure portal.
 - API Version: Find the API Version of the Azure OpenAI service in the Azure portal.
 - Authentication Azure OpenAI Key: Provide the Azure OpenAI Key for authentication.
- OpenAI
 - API Key: Provide your OpenAI API Key for authentication.
 - Base URL: It is optional to provide the custom base URL for OpenAI-compatible endpoints.

5. Click on Settings in the top-right corner and select Model Configuration.

The models available for the user are listed here. You can check if your models are available and function as

Test

expected by selecting the button next to the model.

You can choose from the following options:

- Embedding Models
- Inference Models
- Reranking Models

6. Click on Tools in the top navigation bar.

In this tab you can view and manage tools available for use in chat sessions. You can add Model Context Protocol (MCP) to access tools that operate locally using the RAG studio or provide the URL for tools hosted externally.

What to do next

After the initial configuration, the RAG Studio application will restart in order to access the model providers. This restart must be completed, after which the studio is ready to use.

Related Information

[Using the RAG Studio](#)

Managing Fine Tuning Studio

Customize and optimize large language models to meet the specific needs of your organization.

Fine Tuning Studio Overview

The Fine Tuning Studio is a comprehensive application and ecosystem that enables you to manage the entire lifecycle of Large Language Models (LLMs), including training, fine-tuning, and evaluation. With a streamlined approach to organizing and dispatching Cloudera AI workloads, this AI studio is specifically designed to support tasks related to LLM training and evaluation.

The Fine Tuning Studio provides a centralized platform for managing the entire process, from data preparation to model deployment, with a particular focus on Cloudera AI Jobs.

The Fine Tuning Studio supports a wide range of use cases, including:

- Ticketing Support Agent: Fine-tune LLMs to provide accurate and efficient support for customer inquiries, reducing the need for human intervention and improving customer satisfaction.
- Text to SQL conversion: Use the Fine Tuning Studio to fine-tune LLMs for text-to-SQL conversion, enabling the creation of accurate and efficient SQL queries from natural language inputs.
- Dataset detoxification: Fine-tune LLMs to detect and remove biased or toxic data from datasets, ensuring that models are trained on high-quality and diverse data.

By providing a comprehensive platform for managing the entire lifecycle of LLMs, the Fine Tuning Studio enables you to optimize your AI workloads and improve the accuracy and efficiency of your models. With its streamlined approach and focus on Cloudera AI Jobs, this AI studio is an essential tool for any organization looking to leverage the power of LLMs.

Key Features for Fine Tuning Studio

The Fine Tuning Studio is a powerful tool that enables you to customize and optimize large language models to meet the specific needs of your organization.

With a unified workbench, you can easily import datasets, select base models, launch fine-tuning jobs, evaluate performance, and export models seamlessly.

- **Unified Workbench:** Import datasets, select base models, launch fine-tuning jobs, evaluate performance, and export models in a single, streamlined interface.
- **Direct MLflow Integration:** Track experiments and automate model registry workflows with seamless integration with MLflow.
- **Flexible Interfaces:** Choose from no-code or low-code visual interfaces or use the full-code Python SDK for ultimate flexibility and control.
- **Domain-Specific Applications:** Perfect for domain-specific applications such as healthcare, finance, and legal, where tailored accuracy is crucial.
- **Easy Dataset Upload:** Effortlessly upload datasets to Hugging Face, enabling broader accessibility and facilitating further model training.
- **Cost-Effective:** Fine-tuned smaller models are significantly more cost-effective to operate compared to large foundational models.
- **Improved Performance:** Fine-tuned models can outperform larger foundational models in task-specific scenarios, making them a valuable asset for organizations.

By using the Fine Tuning Studio, you can unlock the full potential of large language models and create customized models that meet the specific needs of your organization.

Launching Fine Tuning Studio within a project

You can launch Fine Tuning Studio on the Cloudera AI Platform to manage the life cycle of LLMs from training, and fine-tuning, to evaluating LLMs.

Before you begin

- A GPU must be available in the cluster. For optimal performance with newer LLMs, it is recommended to use GPUs such as A100 or H100.

Procedure

1. In the Cloudera console, click the **Cloudera AI** tile.
The Cloudera AI Workbenches page displays.
2. Click on the name of the workbench.
The workbenches Home page displays.
3. Click Projects, and then click New Project to create a new project.
In the left navigation pane, the new **AI Studios** option is displayed.
4. Click AI Studios.
5. Click the Launch button in the **Fine Tuning Studio** box.
The **Configure Studio: Fine Tuning** page is displayed.
6. Set the environment variables for the Fine Tuning Studio.
7. Select the Runtime version.

- 8. Click Launch AI Studio.
The **Fine Tuning Studio** page is displayed.
After launching, you can view the list of tasks being executed as part of the AI studio deployment.

Results

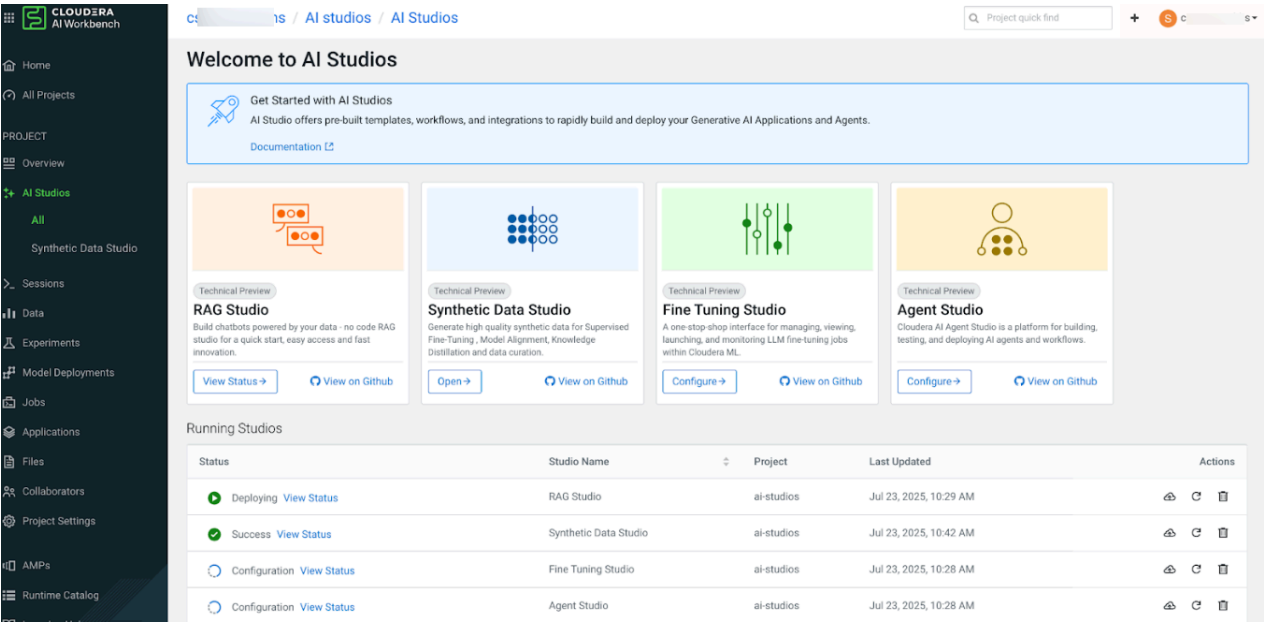
After the configuration, **Fine Tuning Studio** is displayed in the left navigation page under **AI Studios**.
You can train, manage, and evaluate large language models on this page.

Related Information

Using Fine Tuning Studio

Managing AI Studios

The AI Studios feature within Cloudera AI Workbench provides users with a comprehensive interface for managing and interacting with AI Studio deployments. The functionality includes viewing available studios, monitoring deployment status, accessing embedded studio applications, and performing redeployment or resumption of studio workflows.



Accessing the AI Studios Catalog

To access the AI Studios catalog:

Procedure

- 1. In the Cloudera console, click the **Cloudera AI** tile.
The Cloudera AI Workbenches page displays.
- 2. Click on the name of the workbench.
The workbenches Home page displays.
- 3. Click Projects, and then select the required Project.
In the left navigation pane, the new **AI Studios** option is displayed.

4. Browse the available AI Studio templates listed in the catalog.

Viewing AI Studio Deployment Status

The AI Studios page provides visibility into the status of all AI Studio deployments associated with the current project. Each deployment is listed with its corresponding status, allowing users to monitor ongoing and completed activities.

Accessing the embedded AI Studios application


Once an AI Studio deployment has completed successfully, the associated embedded application will be available within the AI Studios sidebar. To launch the application:

Procedure


1. In the Cloudera console, click the **Cloudera AI** tile.
The Cloudera AI Workbenches page displays.
2. Click on the name of the workbench.
The workbenches Home page displays.
3. Click Projects, and then select the required Project.
In the left navigation pane, the **AI Studios** option is displayed.
4. Click on the desired studio entry to open the corresponding embedded application.

Redeploying or Resuming AI Studios

- Redeploying an AI Studio

Click  to re-import and execute tasks based on updates made to the .project-metadata.yaml file. This is useful when changes have been made to the project configuration or workflow definitions.


- Resuming an AI Studio

Click  to continue execution from the point of failure in the last deployment. This is particularly beneficial for recovering from interrupted or failed tasks without restarting the entire workflow.

Deleting an AI Studio

You can delete an AI Studio when it is no longer needed. Be aware that deleting a studio will permanently remove all workloads created during its deployment and delete the associated studio directory.

Procedure

1. In the Cloudera console, click the **Cloudera AI** tile.
The Cloudera AI Workbenches page displays.
2. Click on the name of the workbench.
The workbenches Home page displays.
3. Click AI Studios in the left navigation pane.
4. Click  from the Actions menu to delete the AI Studio.

Host names and endpoints required for AI Studios

For air-gapped installations with a proxy setup, ensure that the following URLs are whitelisted in your firewall rules to enable Cloudera AI Studios.



Important:

- A fully air-gapped installation of AI Studios is not supported if it lacks internet access or a proxy to allow required endpoints.

Endpoints for AI Studios Infrastructure

Ensure the following endpoints are allowed for AI Studios to function properly:

- https://raw.githubusercontent.com/cloudera/AI-Studios/* - Used to fetch the AI Studio catalog file and studio card thumbnails.
- https://github.com/cloudera/* - Used to pull studio resources during launch.

Table 1: Host names and endpoints required for AI Studios

Endpoints	AI Studios			
	RAG Studio	Fine Tuning Studio	Synthetic Data Studio	Agent Studio
release-assets.githubusercontent.com	For workloads to download essential GitHub release assets and build dependencies (such as the protoc library) into Node modules during application setup and building.	For workloads to download essential GitHub release assets and build dependencies (such as the protoc library) into Node modules during application setup and building.	For workloads to download essential GitHub release assets and build dependencies (such as the protoc library) into Node modules during application setup and building.	For workloads to download essential GitHub release assets and build dependencies (such as the protoc library) into Node modules during application setup and building.
*azure.com		For registering a model If we use different models we register it accordingly.		For using Azure OpenAI as an endpoint. Different endpoints must be enabled based on specific service is being used for inferencing.
*bedrock.us-east-1.amazonaws.com	For AWS Bedrock service in US East 1 region		For AWS Bedrock service in US East 1 region	
*bedrock.us-west-2.amazonaws.com	For AWS Bedrock service in US West 2 region		For AWS Bedrock service in US West 2 region	
*bedrock-runtime.us-east-1.amazonaws.com	For AWS Bedrock runtime in US East 1		For AWS Bedrock runtime in US East 1	
*bedrock-runtime.us-west-2.amazonaws.com	For AWS Bedrock runtime in US West 2		For AWS Bedrock runtime in US West 2	
*cdn-lfs.hf.co/		For Large file serving (models, tokenizers)		
*cloudera.com		For querying the workbench from an application, to fetch models		For querying the workbench from an application, to fetch models
*corretto.aws/downloads/latest/amazon-corretto-21-x64-linux-jdk.tar.gz	For Java			
*files.pythonhosted.org	For Python packages	For Python package hosting	For Python package hosting	
*github.com	For GitHub repository hosting	For GitHub repository hosting	For GitHub repository hosting	For GitHub repository hosting

Endpoints	AI Studios			
*github.com/qdrant/ qdrant/releases/download/ v1.11.3/qdrant-x86_64- unknown-linux-musl.tar.gz	For Qdrant			
*github.com/cloudera/ CML_AMP_RAG_Studio/ releases/latest/download/ rag-api.jar *github.com/cloudera/ CML_AMP_RAG_Studio/ releases/latest/download/ fe-dist.tar.gz https://github.com/ cloudera/ CML_AMP_RAG_Studio/ releases/latest/download/ node-dist.tar.gz *github.com/cloudera/ CML_AMP_RAG_Studio/ releases/download/ model_download/ craft_mlt_25k.pth *github.com/cloudera/ CML_AMP_RAG_Studio/ releases/download/ model_download/ latin_g2.pth	For RAG Studio artifacts			
*huggingface.co/		For Hugging Face		
*hf.co/		For Hugging Face		
*nodejs.org			For node packages	For node packages
*nodejs.org/dist/v22.15.0/ node-v22.15.0-darwin- arm64.tar.xz	For node packages			
*npmjs.org			For node packages	For node packages
*npmjs.com			For node packages	For node packages
*objects.githubusercontent.com		For GitHub raw content storage	For GitHub raw content storage	
*pypi.org	For python packages	For Python Package Index	For Python Package Index	For python packages
*registry.npmjs.org			For NPM package registry	
*raw.githubusercontent.com/ nvm-sh/nvm/v0.40.1/ install.sh	For NVM			
*raw.githubusercontent.com		For GitHub raw content direct access	For GitHub raw content direct access	
*registry.npmjs.org				For installations
*vitejs.dev			For node packages	