



# Introduction

May 09, 2025

## Who are we?

CreateAI is a platform designed to support advanced interactions with large language models (LLMs). The platform supports 25+ large language models capable of audio, image and text modality. The documentation provided here aims to offer all necessary details about how to effectively use our APIs, including descriptions of payloads, parameters, response formats, and full API specifications.

### Note

*The DocuBot link on the far right is available for you to chat to our bot about the documentation or any other API related help. If you get a 404 error, kindly contact Siddharth Jain at sjain238@asu.edu regarding access to the bot. Incase the bot does not work well for you feel free to contact your administrator regarding further assistance. Thank you!*

### Warning

This documentation is still under development and might be incomplete. Please contact your administrator incase something goes wrong.

# Usage

## Token

To use our platform, you will first need a token. Please contact [Ayat Sweid](#) (Ayat.Sweid@asu.edu) or [Paul Alvarado](#) (Paul.Alvarado.1@asu.edu) for API access token. You will also be given an API for your department or team to access the platform.

## Request Method

Our platform supports various methods depending on what you want to do. We support both **REST** and **WebSocket** methods for most of the methods, which will be detailed in the corresponding method section below. After the August 29th update, we have a new way of calling

various methods. The method in the URL stays `queryV2` for REST and action stays `queryV2` for WebSocket, but with an additional parameter in the payload called `endpoint`, which can be one of the following: `"speech"`, `"image"`, `"vision"`, or `"audio"`.

## URLs

- **REST:** `https://api-main-beta.aiml.asu.edu/{method}`  
Your access token goes into the authorization header as a bearer token.
- **WebSocket:** `wss://apiws-main-beta.aiml.asu.edu/?access_token={access_token}`

**Note:** Replace `{your_api}` with the appropriate API value. The default value for most users is `main`, but teams such as LE, KE, or Edplus may have their own custom API. The `method` parameter can be one of the following: `query` or `queryV2`, `search`, `embeddings`, or `chunk`.

## Project Owner Token

The project owner token allows you to modify your project settings, upload files,

manage the knowledge base, and more.

## File Upload API

To upload a file to your knowledge base, use the following API call:

```
POST https://api-main-be1
```

Content-Type: application/json

Authorization: Bearer <Your token>

```
{  
    "resource": "data",  
    "method": "upload",  
    "details": {  
        "project_id": "your_project_id",  
        "db_type": "opensearch",  
        "files": [  
            {  
                "file_name": "your_file_name",  
                "search_tags": ["tag1", "tag2"],  
                "metadata": {  
                    "any_metadata": true  
                },  
                "selected": true,  
                "visible": true  
            }  
        ]  
    }  
}
```

This call will return a pre-signed URL that you can use to upload your file.

## Check Upload Status (Optional)

To check the status of uploaded files, use the same URL with the following payload:

```
POST https://api-main-be1
```

Content-Type: application/json  
Authorization: Bearer <Your\_Token>

```
{  
    "resource": "data",  
    "method": "list",  
    "details": {  
        "project_id": "your_project_id",  
        "db_type": "opensearch"  
    }  
}
```

This request is optional and only returns the status of uploaded files.

## Authentication

This API uses Bearer Token for authentication. The token should be included in the Authorization header of the request.

Authorization: Bearer <Your\_Token\_Here>

## Common Issues and Troubleshooting

- **Invalid key-value pair in Authorization header:** Ensure that the token is correctly included and formatted. It should be in the form of a Bearer token.
- **400 Bad Request:** Check if the request body or headers are formatted correctly. Specifically, ensure that all required fields are present and correct.

- **Authorization Issues:** If the token does not have access to the requested resources, you might receive errors related to missing permissions. Ensure the correct token is being used.
- **Incorrect Method (POST/GET):** Some endpoints expect a POST request, not a GET. Ensure the correct HTTP method is used for each endpoint.

## Token Information

Most of the tokens we provide is linked to your project ID, so all payload parameters will be used via the project settings unless you override them. Any search will be done against the project ID.

For example, if your project ID is `your_project_id_here`, the token will be linked to that project and you can check on <https://jwt.io>

## Methods

### 1. POST /search

Perform a search operation.

#### Example Query:

```
{  
  "query": "What is your name?",  
  "search_params": {  
    "collection": "your_project_id_h
```

```
# or collection name if you have
"source_name": "list of your fil
"top_k": "number of search resul
"reranker": "boolean",
"output_fields": ["source_name",
# or whatever you want from quer
"retrieval_type": "chunk" or "ne
"expr": "source_name != 'File_na
}
}
```

## Parameters:

- **source\_name** (array) (Optional):

e.g. ["file\_name1.pdf", "file\_name2.txt"]

You can use this to filter your data before searching with the file names you provided while uploading the data.

- **top\_k** (int, default is 3):

Number of results to retrieve. You can set the number of chunks or documents you want to retrieve here. It can range from 1 to any desired number. Keep in mind, setting a higher number may slow the response time and sometimes reduce quality.

- **output\_fields** (array) (Optional):

e.g. ["source\_name", "content", "source\_type"]

This lets you get the metadata at the end of a response. You need to set the response format to JSON in the additional parameters.

- **retrieval\_type** (string, default: "chunk"):

Options are: 'chunk', 'neighbor', 'document'.

This defines how the data is retrieved.

- **expr** (string) (Optional):

Custom expression to filter results, e.g.  
"metadata1 == 'corresponding\_filter' &&  
metadata2 == 'corresponding\_filter'".

---

## 2. POST /query

Execute a query with specific parameters.

### Example Query:

```
{  
    "action": "query",  
    "request_source": "override_params  
    # if you want to override any para  
    "query": "What is this?",  
    "model_provider": "model_provider"  
    "model_name": "model_name",  
    "session_id": "testingsid123456",  
    "model_params": {  
        "temperature": "model_params['te  
        "system_prompt": "system_prompt"  
        "top_p": "model_params['top_p']"  
        "top_k": "model_params['top_k']"  
        "tools": [  
            {  
                "type": "function",  
                "function": {  
                    "name": "get_weather",  
                    "description": "Get curren  
                    "parameters": {  
                        "type": "object",  
                        "properties": {  
                            "location": {  
                                "type": "string",  
                                "description": "City  
                            }  
                        },  
                        "required": ["location"]  
                    }  
                }  
            }  
        ]  
    }  
}
```

```
        "additionalProperties":  
    },  
    "strict": true  
}  
},  
{  
    "type": "function",  
    "function": {  
        "name": "get_continent",  
        "description": "Get current continent based on location",  
        "parameters": {  
            "type": "object",  
            "properties": {  
                "location": {  
                    "type": "string",  
                    "description": "City or location name",  
                    "example": "New York City"  
                },  
                "required": ["location"]  
            },  
            "additionalProperties":  
        },  
        "strict": true  
    }  
}  
],  
"enable_search": true,  
"search_params": {  
    "collection": "collection",  
    "source_name": "source_name",  
    "top_k": "top_k",  
    "reranker": "reranker",  
    "retrieval_type": "retrieval_type",  
    "output_fields": ["content", "text"],  
    "prompt_mode": "unrestricted",  
    "rerank": true,  
    "reranker_model": "amazon_reranker",  
    "reranker_provider": "aws",  
    "top_n": 3  
},  
"history": [  
    {  
        "query": "What is your name?",  
        "response": "My name is Documenter AI"  
    }  
]
```

```
        },
        {
            "query": "Is AI the solution",
            "response": "No, AI is not t"
        }
    ],
    "enable_history": true,
    "response_format": {"type": "js"
    "enhance_prompt": {"timezone": "
}
}
```

## Parameters:

- **expr** (string) (Optional):  
Defines custom expression filters (e.g., "tags != 'Module 0' && tags != 'Syllabus'").
- **source\_name** (array) (Optional):  
A list of file names (e.g., ["file\_name1.pdf", "file\_name2.txt"]). Use this to filter your data before searching with the file names you provided while uploading the data.
- **top\_k** (int, default is 3):  
Set the number of chunks or documents you want to retrieve. This can range from 1 to the desired number. Keep in mind, higher numbers may slow response time and reduce quality.
- **output\_fields** (array) (Optional):  
e.g., ["source\_name", "content", "source\_type"]. This lets you get the metadata at the end of a response. Set the response format to JSON in the additional parameters.
- **retrieval\_type** (string, default: "chunk"):  
Options are: 'chunk', 'neighbor', 'document'. This defines how the data is retrieved.

- **prompt\_mode** (string):  
e.g., "restricted", "unrestricted", or "custom".  
Defines the mode for the prompt used in search queries.
- **search\_prompt** (string) (Required when **prompt\_mode** is set to "custom"):  
Custom search prompt when prompt\_mode is set to custom (e.g., "This is your new prompt when search is enabled.")
- **enhance\_prompt** (object):  
Parameters to adjust timezone, time, date, and verbosity in the model's responses.
  - **timezone** (string, default: "MST"): Sets the timezone (e.g., "MST").
  - **time** (boolean, default: false): If true, the model will consider the current time in its responses.
  - **date** (boolean, default: false): If true, the model will include the current date in its responses.
  - **verbosity** (string, default: none): Can be set to "brief", "detailed", or "succinct" to control the level of detail in the responses.
- **enable\_search** (boolean, default: false):  
If true, enables the search functionality and allows the model to search through your documents.
- **enable\_history** (boolean, default: false):  
If true and a session ID is provided, enables the model to use the history from the current session.

# Models List

ID	Provider	Is Active	Capabilities
claude3_opus	aws	True	query, vision
llama3-405b	aws	True	query
nova-pro	aws	True	query, vision
llama3_2-3b	aws	True	query
claude2_1	aws	True	query
claude3_sonnet	aws	True	query, vision
claude4_sonnet	aws	True	query, vision
claude4_opus	aws	True	query, vision
nova-lite	aws	True	query, vision
nova-micro	aws	True	query
mistral-7b	aws	True	query
llama3_2-90b	aws	True	query, vision
amazon_rerank	aws	True	reranker
llama3-70b	aws	True	query
llama3-8b	aws	True	query
claude3_haiku	aws	True	query, vision
mistral-large	aws	True	query
claudeinstant	aws	True	query
claude3_7_sonnet	aws	True	query, vision

claude2	aws	True	query
llama3_2-1b	aws	True	query
titang1lite	aws	True	query
titang1express	aws	True	query
mistral-8x7b	aws	True	query
claude3_5_sonnet	aws	True	query, vision
cohere_rerank-3_5	aws	True	reranker
llama3_2-11b	aws	True	query, vision
geminipro1_5	gcp-deepmind	True	query, vision, audio
geminiflash2-lite	gcp-deepmind	True	query, vision, audio
geminiflash2	gcp-deepmind	True	query, vision, audio
imagen3	gcp-deepmind	True	image
o1	openai	True	query, vision
gpt4-32k	openai	True	query
gpt3_5-16k	openai	True	query
gpt4o-transcribe	openai	True	audio
gpt4_1-nano	openai	True	query, vision
o3-mini	openai	True	query

o3	openai	True	query, vision
tts1hd	openai	True	speech
gpt4_1	openai	True	query, vision
gpt4_1-mini	openai	True	query, vision
tts1	openai	True	speech
gpt4o_mini	openai	True	query, vision
gpt5	openai	True	query, vision
gpt5_chat	openai	True	query, vision
gpt5_mini	openai	True	query, vision
gpt5.nano	openai	True	query, vision
dalle3	openai	True	image
gpt4	openai	True	query
gpt4o	openai	True	query, vision
gpt4o_mini-tts	openai	True	speech
gpt3_5	openai	True	query
gpt4o_mini-transcribe	openai	True	audio
whisper-1	openai	True	audio
o4-mini	openai	True	query, vision
gpt4turbo	openai	True	query, vision

## Audio Model Example

```
{
  "action": "query",
  "endpoint": "audio",
```

```
        "query": "What is this audio?",  
        "audio_file": "",  
        "model_provider": "openai",  
        "model_name": "whisper-1",  
        "model_params": {  
            "translation": true,  
            "language": "en"  
        },  
        "response_format": {  
            "type": "json"  
        }  
    }
```

**model\_params** (dict) (Optional): This contains the model-specific parameters. You can set the **translate** boolean here (default is False). This is only applicable for the Whisper model (see payload for example). Additionally, a **language** parameter has been added to specify the language for transcriptions (default is "en" for English ISO Code). See **Language** for various ISO codes and supported languages.

### Language (ISO Code) Examples:

- Afrikaans (af), Arabic (ar), Armenian (hy), Azerbaijani (az), Belarusian (be), Bosnian (bs), Bulgarian (bg), Catalan (ca), Chinese (zh), Croatian (hr), Czech (cs), Danish (da), Dutch (nl), English (en), Estonian (et), Finnish (fi), French (fr), Galician (gl), German (de), Greek (el), Hebrew (he), Hindi (hi), Hungarian (hu), Icelandic (is), Indonesian (id), Italian (it), Japanese (ja), Kannada (kn), Kazakh (kk), Korean (ko), Latvian (lv), Lithuanian (lt), Macedonian (mk), Malay (ms), Marathi (mr), Maori (mi), Nepali (ne), Norwegian (no), Persian (fa), Polish (pl), Portuguese (pt), Romanian (ro),

Russian (ru), Serbian (sr), Slovak (sk), Slovenian (sl), Spanish (es), Swahili (sw), Swedish (sv), Tagalog (tl), Tamil (ta), Thai (th), Turkish (tr), Ukrainian (uk), Urdu (ur), Vietnamese (vi), Welsh (cy).

## Vision Model Example

```
{  
    "action": "query",  
    "endpoint": "vision",  
    "query": "What is in this image?",  
    "image_file": "",  
    "model_provider": "openai",  
    "model_name": "gpt4o",  
    "response_format": {  
        "type": "json"  
    }  
}
```

## Speech Model Example

```
{  
    "action": "query",  
    "endpoint": "speech",  
    "query": "This is a test sentence.",  
    "voice": "alloy",  
    "model_provider": "openai",  
    "model_name": "tts1"  
}
```

**Note:** The `voice` parameter (default is "alloy") accepts: "alloy", "nova", "echo", "fable", "onyx", "shimmer".

## Image Generation Model Example

```
{  
    "action": "query",
```

```
        "endpoint": "image",
        "query": "This is a test sentence.",
        "model_provider": "openai",
        "model_name": "dalle3"
    }
```

## Reranker Model Example

```
{
    "query": "query",
    "model_provider": "aws",
    "model_name": "reranking_models",
    "model_params": {
        "top_n": 3,
        "documents": [ "list of string of
    }
}
```

## /embeddings method

```
{
    "query": "Hi what is your name?",
    "embeddings_provider": "openai",
    "embeddings_model": "ada",
    "dimensions": 1356,
    "normalize": true
}
```

### #### Parameters

- **query** (string): Your query for getting the embeddings.
- **embeddings\_provider** (string): Options include "openai", "aws", "azure", "gcp". Specifies the provider for the embeddings.

- **embeddings\_model** (string): The corresponding model for the selected provider.
- **dimensions** (int, default is 1024):  
For example, 256, 512, or 1024 (only applicable for the titan2 model).
- **normalize** (boolean, default is false):  
Whether to normalize the embeddings (only applicable for the titan2 model).

## Supported Models

The following models are supported across different providers:

- **model\_provider = openai, model\_name = te3l**
- **model\_provider = openai, model\_name = ada**
- **model\_provider = openai, model\_name = te3s**
- **model\_provider = aws, model\_name = titan**
- **model\_provider = aws, model\_name = titan2**
- **model\_provider = aws, model\_name = ce-english**
- **model\_provider = aws, model\_name = ce-multilingual**
- **model\_provider = gcp, model\_name = gecko**
- **model\_provider = gcp, model\_name = gecko-2**
- **model\_provider = gcp, model\_name =**

### gecko-3

- **model\_provider = gcp, model\_name = gecko-multilingual**
- 

## /chunks method

This endpoint shows how your document will be chunked and tokenized. It supports various chunking strategies and displays the resulting tokens.

```
{  
    "text": "Text you want to chun  
    "enable_chunks": true,  
    "enable_tokenizer": true,  
    "chunk_params": {  
        "chunk_strategy": "character  
        "chunk_size": 256  
    },  
    "tokenizer_params": {  
        "tokenizer_provider": "opena  
        "tokenizer_name": "gpt2" //  
    }  
}
```

### Parameters:

- **text** (string):  
The text that you want to chunk and tokenize.
- **enable\_chunks** (boolean):  
Whether to enable chunking of the document (True/False).
- **enable\_tokenizer** (boolean):

Whether to enable tokenization of the document (True/False).

- **chunk\_params** (object):  
Defines how the text will be chunked.
  - **chunk\_strategy** (string): Options are "character", "words", or "sentence". Defines the strategy used for chunking.
  - **chunk\_size** (int): The size of each chunk (e.g., 256 characters, words, or sentence tokens).
- **tokenizer\_params** (object):  
Defines the tokenizer configuration.
  - **tokenizer\_provider** (string): The provider for tokenization (e.g., "openai").
  - **tokenizer\_name** (string): The specific tokenizer model to use (e.g., "gpt2").

---

## Additional Information

When testing an endpoint (for example using Postman):

1. Set the method to POST.
2. Use the URL:
  - For the query endpoint: <https://api-main-beta.aiml.asu.edu/query>
  - For the search endpoint: <https://api-main-beta.aiml.asu.edu/search>

3. In the Headers section, add: **Authorization: Bearer <Your\_Token\_Here>**
4. In the Body section, select raw and JSON, then add your payload.

Important: Always verify that you are using the correct HTTP method, a valid Bearer token, and that all required parameters are provided.

## Additional Features

The query endpoint supports audio, vision, and speech models. By adding the parameter `"request_source": "override_params"` in your payload, you can override any project settings. This allows you to provide custom system prompts and adjust model parameters that may otherwise be set on the platform.

Adding the parameter `"request_source": "override_params"` you can override any project settings you have enabled. This means that even if you have a system prompt in your project settings, you can override it by sending this flag and the new system prompt in the payload. You can also override the models and use any models that we offer and not available to select via the UI for example, speech, audio and vision models.

Designed with ❤️ by the [AI Acceleration Team](#) for developers  
by developers

