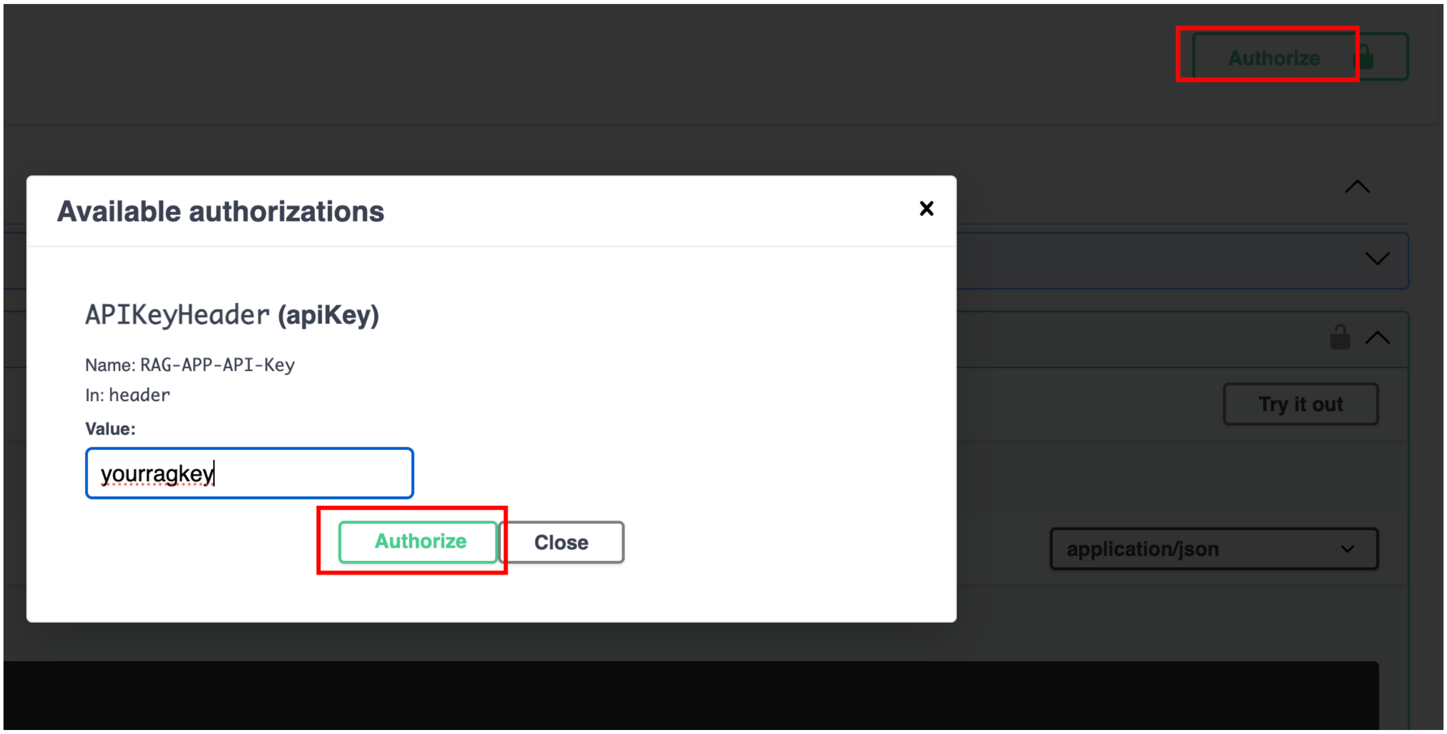
Using RAG Service

Use the RAG API end point provided in the “Useful links” section to open the FAST API Swagger UI.

Step 1: Authorize

Click on the “Authorize” button and top right. Provide the RAG-APP-API-KEY value from the supplied .env file and click “Authorize”.



Available authorizations

APIKeyHeader (apiKey)

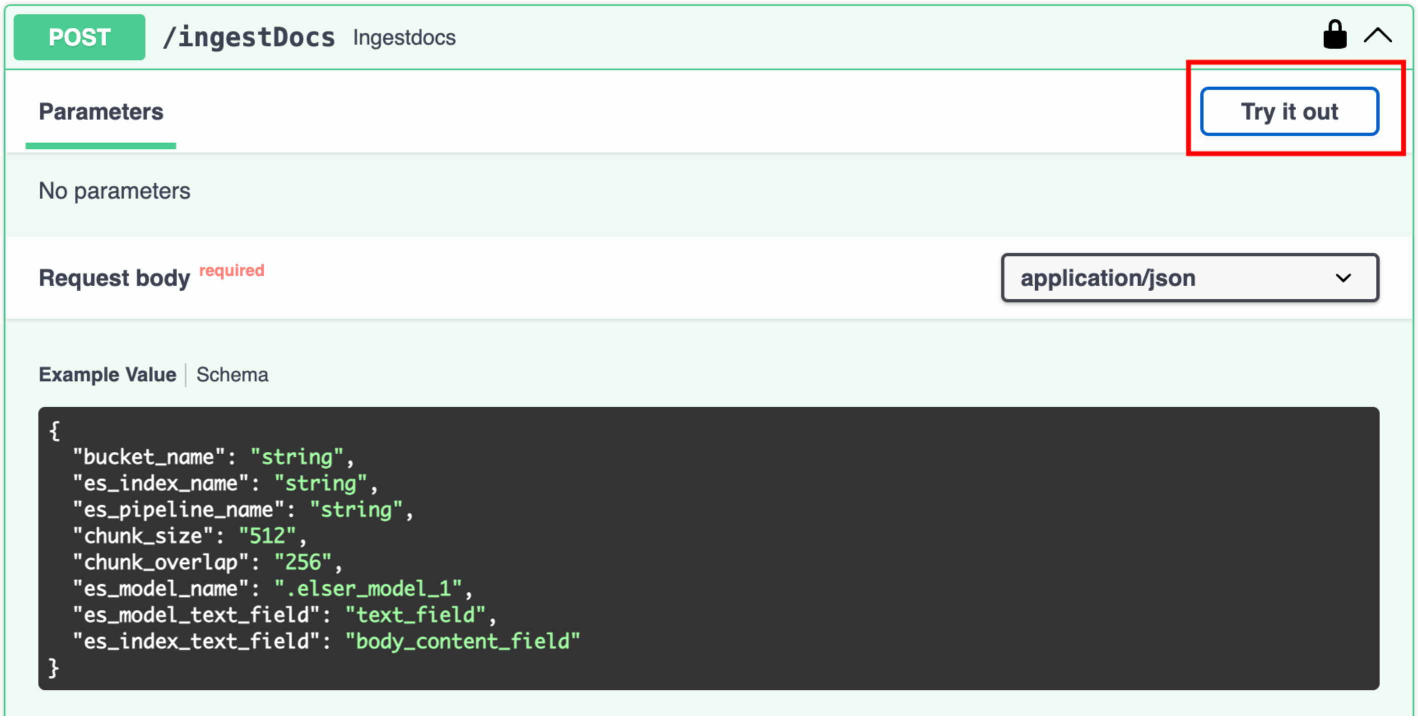
Name: RAG-APP-API-Key

In: header

Value: yourragkey

Step 2: Ingest Documents

Upload the document into the cloud object storage (COS) and use the supplied payload (indexDocsInput.json) as an example to ingest the document. Instruction to access COS is provided in the “Useful links” section below.

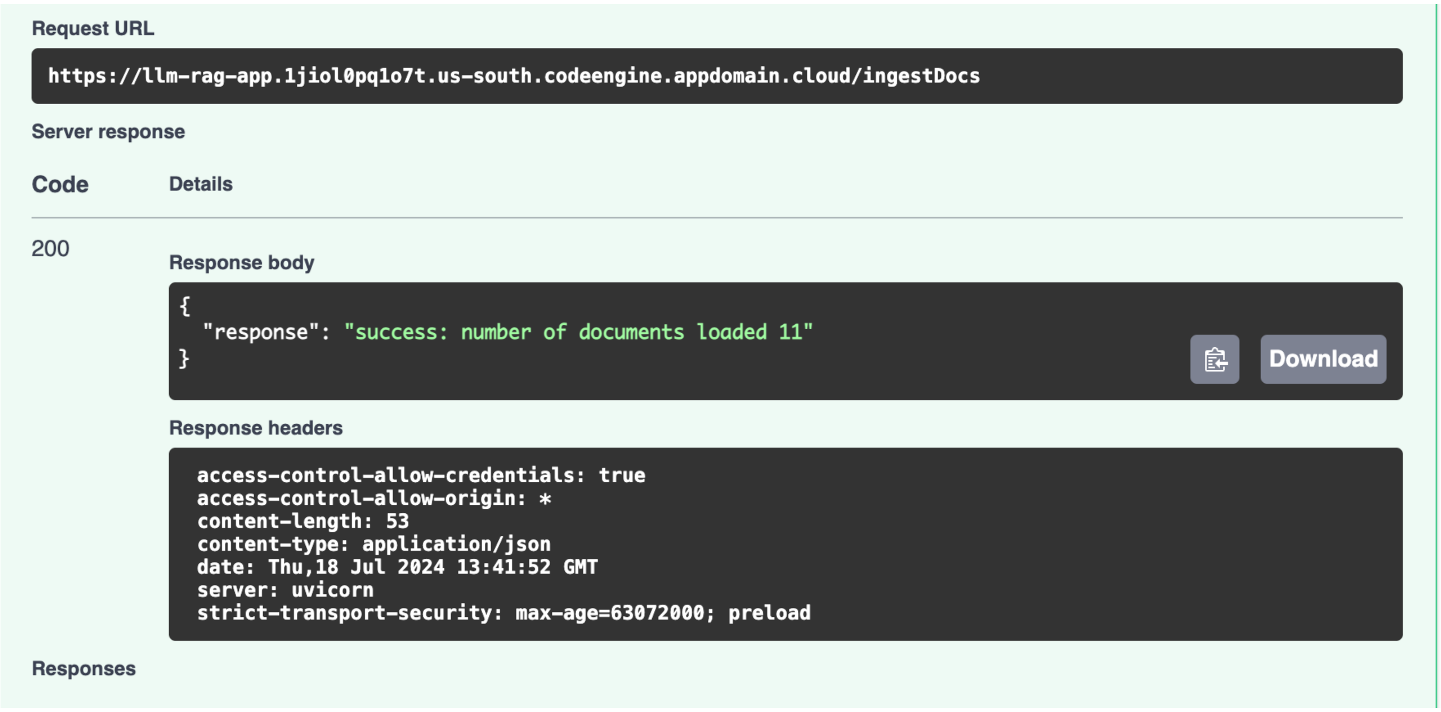


No parameters

Request body required

```
{
  "bucket_name": "string",
  "es_index_name": "string",
  "es_pipeline_name": "string",
  "chunk_size": "512",
  "chunk_overlap": "256",
  "es_model_name": ".elser_model_1",
  "es_model_text_field": "text_field",
  "es_index_text_field": "body_content_field"
}
```

Successful response:


Request URL

```
https://llm-rag-app.1jiol0pq1o7t.us-south.codeengine.appdomain.cloud/ingestDocs
```

Server response

Code	Details
200	Response body

```
{
  "response": "success: number of documents loaded 11"
}
```

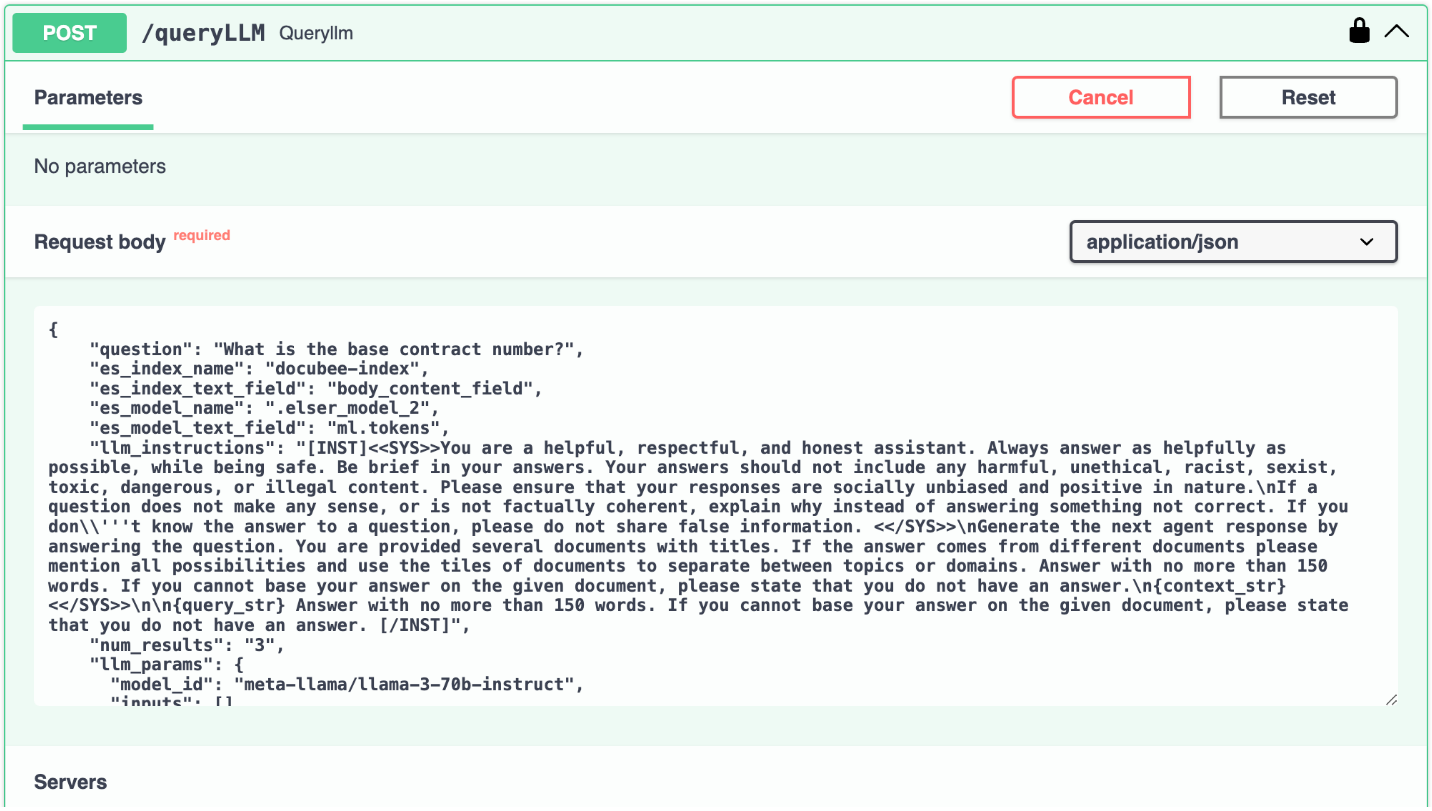
Response headers

```
access-control-allow-credentials: true
access-control-allow-origin: *
content-length: 53
content-type: application/json
date: Thu,18 Jul 2024 13:41:52 GMT
server: uvicorn
strict-transport-security: max-age=63072000; preload
```

Responses

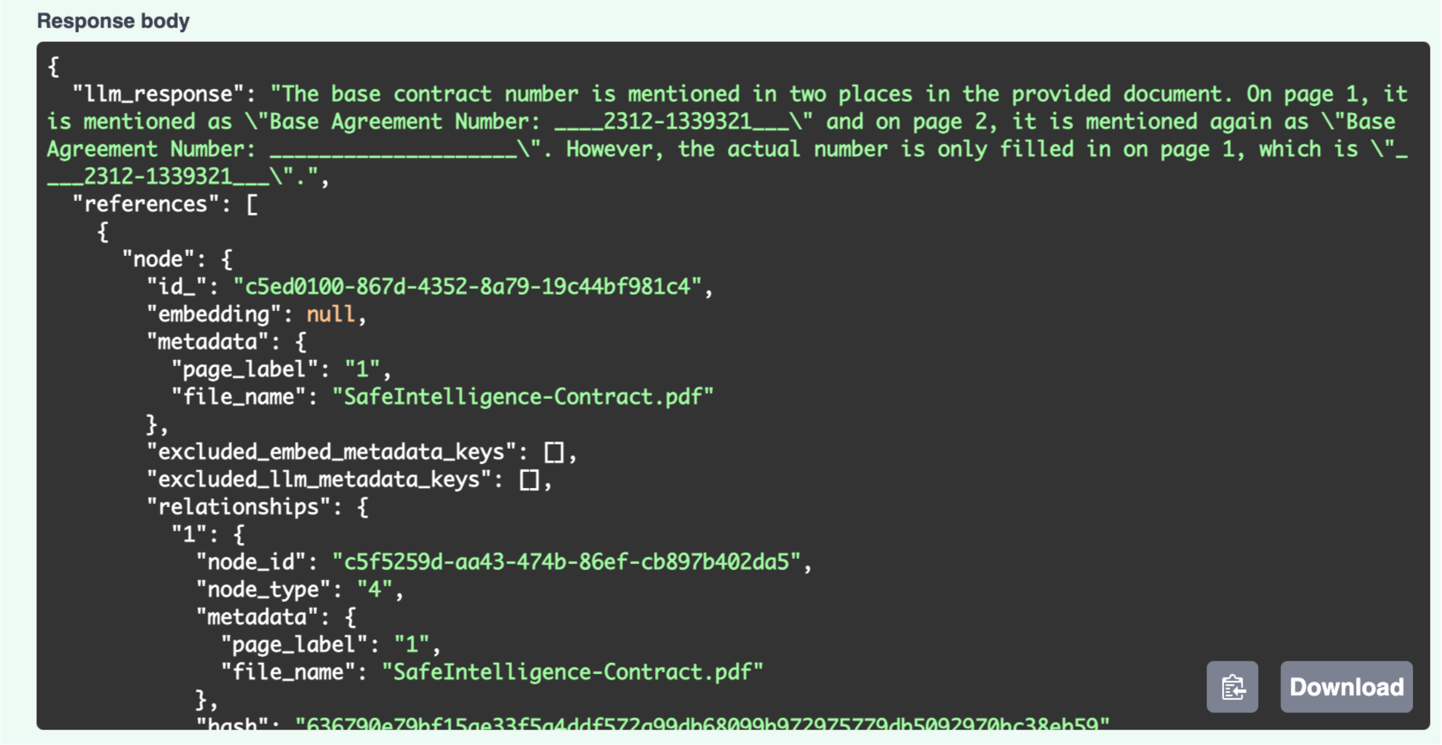
Step 3: Query LLM

Use the “queryLLM” end point to run a semantic search against the Elasticsearch database and then send the returned relevant passages to LLM. Use the supplied payload “queryLLMInput.json” as an example to run this API.



```
{
    "question": "What is the base contract number?",
    "es_index_name": "docubee-index",
    "es_index_text_field": "body_content_field",
    "es_model_name": ".elser_model_2",
    "es_model_text_field": "ml.tokens",
    "llm_instructions": "[INST]<<SYS>>You are a helpful, respectful, and honest assistant. Always answer as helpfully as
possible, while being safe. Be brief in your answers. Your answers should not include any harmful, unethical, racist, sexist,
toxic, dangerous, or illegal content. Please ensure that your responses are socially unbiased and positive in nature.\nIf a
question does not make any sense, or is not factually coherent, explain why instead of answering something not correct. If you
don\\'''t know the answer to a question, please do not share false information. <</SYS>>\nGenerate the next agent response by
answering the question. You are provided several documents with titles. If the answer comes from different documents please
mention all possibilities and use the tiles of documents to separate between topics or domains. Answer with no more than 150
words. If you cannot base your answer on the given document, please state that you do not have an answer.\n{context_str}
<</SYS>>\n\n{query_str} Answer with no more than 150 words. If you cannot base your answer on the given document, please state
that you do not have an answer. [/INST]",
    "num_results": "3",
    "llm_params": {
      "model_id": "meta-llama/llama-3-70b-instruct",
      "inputs": []
```

Successful response:



```
{
  "llm_response": "The base contract number is mentioned in two places in the provided document. On page 1, it
is mentioned as \"Base Agreement Number: ____2312-1339321___\" and on page 2, it is mentioned again as \"Base
Agreement Number: ___________________\". However, the actual number is only filled in on page 1, which is \"_
___2312-1339321___\".",
  "references": [
    {
      "node": {
        "id_": "c5ed0100-867d-4352-8a79-19c44bf981c4",
        "embedding": null,
        "metadata": {
          "page_label": "1",
          "file_name": "SafeIntelligence-Contract.pdf"
        },
        "excluded_embed_metadata_keys": [],
        "excluded_llm_metadata_keys": [],
        "relationships": {
          "1": {
            "node_id": "c5f5259d-aa43-474b-86ef-cb897b402da5",
            "node_type": "4",
            "metadata": {
              "page_label": "1",
              "file_name": "SafeIntelligence-Contract.pdf"
            },
            "hash": "636790e79bf15ae33f5a4ddf572a99db68099b972975779db5092970bc38eb59"
```

Useful links

1. RAG API end point:
   https://llm-rag-app.1jiol0pq1o7t.us-south.codeengine.appdomain.cloud/docs
2. Elastic Search Kibana UI:
   https://itz-watson-apps-eyykdg6d-kibana-app.1ji4dzrczuu1.us-south.codeengine.appdomain.cloud
3. All credentials including the IBM Cloud account information is available in the Techzone reservation:
   https://techzone.ibm.com/my/reservations
4. Cloud Object storage:
   Look for storage service under Storage category in IBM Cloud Console > Resource List
   https://cloud.ibm.com/resources
5. Watsonx Studio/Prompt Lab:
   https://dataplatform.cloud.ibm.com/wx/home?context=wx
6. RAG Service Source Code:
   https://github.com/ibm-build-lab/RAG-LLM-Service