

SimSage AI settings

You can set up various Large Language Model providers (LLMs) in SimSage. This is to aid Retrieval Augmented Generation (RAG) as supported by SimSage.

This document discusses the settings in the admin UX for enabling this feature. A *token* in the context below is anything SimSage considers to be a separate token. This includes spaces, words, punctuation items, and entities like credit cards etc. Each of such items is considered a distinct token in SimSage.

Navigate to “AI set up” in “the Mind” section for a given knowledge-base.

None

Select “none” to disable the use of AI and large language models (default).

SimSage

Overview User Management demo knowledge-base Document Management The Mind

AI set up Synonyms Semantics Synsets

Large Language Model (LLM)

SimSage AI Setup Guide

LLM Provider

OpenAI

None

OpenAI

Google Gemini

Ollama (local)

Amazon Bedrock

CoPilot

OpenAI model gpt-4o-mini

maximum number of tokens is exceeded

document summarization token count 1000

search box question answering, the number of results to pass to the LLM 5

search box question answering, the number of tokens to pass for each search result 100

the maximum number of tokens per month before the model is disabled and a warning is emailed (if > 0) 0

the estimated number of tokens used this month: 0

Save

Open AI settings

SimSage

Overview

User Management

demo knowledge-base

Document Management

The Mind

AI set up

Synonyms

Semantics

Synsets

Large Language Model (LLM)

LLM Provider
OpenAI

OpenAI key
Enabled (automatically disabled if maximum number of tokens is exceeded)

document Q&A token count
1000

document summarization token count
1000

search box question answering, the number of results to pass to the LLM
5

search box question answering, the number of tokens to pass for each search result
100

the maximum number of tokens per month before the model is disabled and a warning is emailed (if > 0)
0

the estimated number of tokens used this month: 0

Save

OpenAI model
gpt-4o-mini
gpt-3.5-turbo
gpt-3.5-turbo-16k
gpt-4
gpt-4-32k
gpt-4o-latest
gpt-4o-mini
gpt-4.1
o1

SimSage AI Setup Guide

Open AI specific settings are

setting	description
OpenAI key	your API key from OpenAI
OpenAI model	the OpenAI model to use for inference

Google Gemini settings

SimSage

Overview

User Management

demo knowledge-base

Document Management

The Mind

AI set up

Synonyms

Semantics

Synsets

Large Language Model (LLM)

LLM Provider
Google Gemini

Gemini key
Enabled (automatically disabled if maximum number of tokens is exceeded)

document Q&A token count
1000

document summarization token count
1000

search box question answering, the number of results to pass to the LLM
5

search box question answering, the number of tokens to pass for each search result
100

the maximum number of tokens per month before the model is disabled and a warning is emailed (if > 0)
0

the estimated number of tokens used this month: 0

Save

Gemini model
Select Model
gemini-1.5-flash
gemini-1.5-pro
gemini-2.0-flash-lite
gemini-2.0-flash

SimSage AI Setup Guide

setting	description
Gemini key	your Google Gemini API key
Gemini model	the Gemini model you wish to use for inference

Ollama settings

This requires **Ollama** to be running inside your cluster (ask your SimSage representative). This option is the only available Local LLM provider. This means that information is not sent to any external provider and kept strictly inside your SimSage platform. This is the most secure option for your information.

- Overview
- User Management
- demo knowledge-base
- Document Management
- The Mind

AI set up

Synonyms

Semantics

Synsets

LLM Provider

Ollama (local)

NB. Ollama requires a GPU, please ensure you have a GPU available / set up.

- 14b models requires 9GB of GPU RAM
- Smaller models require 4GB of GPU RAM
- CPU only is not sufficient for Ollama

Ollama Models

Select Model

llama3.2:3b

cogito:14b

run

document summarization token count

1000

search box question answering, the number of results to pass to the LLM

5

search box question answering, the number of tokens to pass for each search result

100

the maximum number of tokens per month before the model is disabled and a warning is emailed (if > 0)

0

the estimated number of tokens used this month: 0

Save

Open AI specific settings are

setting	description
Ollama model	the Ollama model to use for inference.

Amazon Bedrock

The screenshot shows the 'AI set up' tab for the 'Large Language Model (LLM)' configuration. The interface includes a sidebar with navigation options: Overview, User Management, demo knowledge-base (selected), Document Management, and The Mind. The main configuration area has a 'Amazon Bedrock' dropdown menu. Below it, there are input fields for 'Amazon key id', 'Amazon key' (with a red asterisk), 'Bedrock model' (with a 'Select Model' dropdown showing 'Claude 3 Sonnet' and 'Mistral Large'), and 'Amazon region' (set to 'Asia Pacific (Sydney)'). There are also input fields for 'document Q&A token count' (1000), 'document summarization token count' (1000), 'search box question answering, the number of results to pass to the LLM' (5), and 'search box question answering, the number of tokens to pass for each search result' (100). A 'Serve' button is at the bottom left. A 'SimSage AI Setup Guide' link is in the top right.

setting	description
Amazon key id	the Amazon key id (careful, these can be region specific)
Amazon key	the secret associated with the Amazon key id
Bedrock model	the AWS Bedrock model you wish to use for inference
Amazon region	One of the AWS regions where the model is enabled and hosted

CoPilot

CoPilot is an Azure specific service connecting Azure to OpenAI.

The screenshot shows the 'AI set up' tab for the 'Large Language Model (LLM)' configuration, specifically for the 'CoPilot' provider. The sidebar is the same as the previous screenshot. The main configuration area has a 'CoPilot' dropdown menu. Below it, there are input fields for 'CoPilot key' (with a red asterisk), 'CoPilot endpoint', and 'CoPilot model' (set to 'gpt-4'). There are also input fields for 'document Q&A token count' (1000), 'document summarization token count' (1000), 'search box question answering, the number of results to pass to the LLM' (5), and 'search box question answering, the number of tokens to pass for each search result' (100). A 'Serve' button is at the bottom left. A 'SimSage AI Setup Guide' link is in the top right.

setting	description
---------	-------------

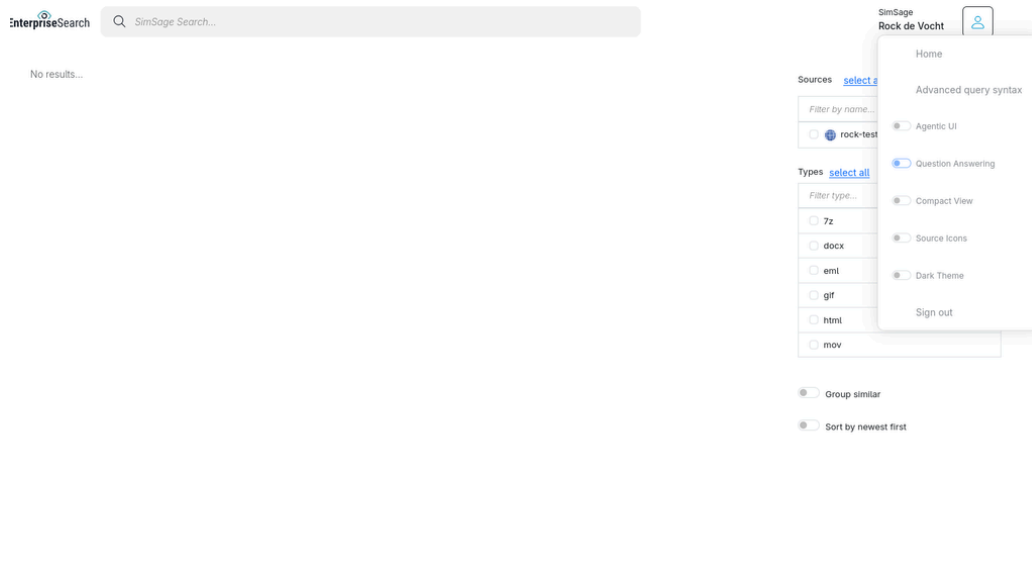
CoPilot key	the CoPilot key to use, provided by Azure once you create a CoPilot instance.
CoPilot endpoint	the Azure endpoint for the CoPilot instance you created inside Azure (provided by Azure)
CoPilot model	the model to use for inference with CoPilot, currently only gpt-4 is available.

SimSage shared settings

setting	description
enabled	If set, the LLM is used. Otherwise the LLM is disabled (not used). This is automatically disabled if the LLM exceeds the maximum number of tokens.
document Q&A token count	the number of tokens passed, starting at the beginning of a document for a user asking questions interactively of a document.
document summarization token count	the number of tokens passed, starting at the beginning of a document for document summarization.
search box question answering, the number of results to pass to the LLM	Used for question answering while searching. This is the number of results to pass to the LLM, starting with the top result.
search box question answering, the number of tokens to pass for each search result	The number of tokens to pass for each search result found, starting at the sentence of each successful search.
Maximum number of tokens per month	If this number is greater than zero (0), this number is the monthly limit for the number of tokens send to / received from the selected LLM. Once exceeded (if greater than zero) the LLM is disabled. This data is reset midnight, the 1st of the Month.
the estimated number of tokens used this month	(read only) what SimSage estimates the number of tokens is sent and received by the selected LLM.

Search UX

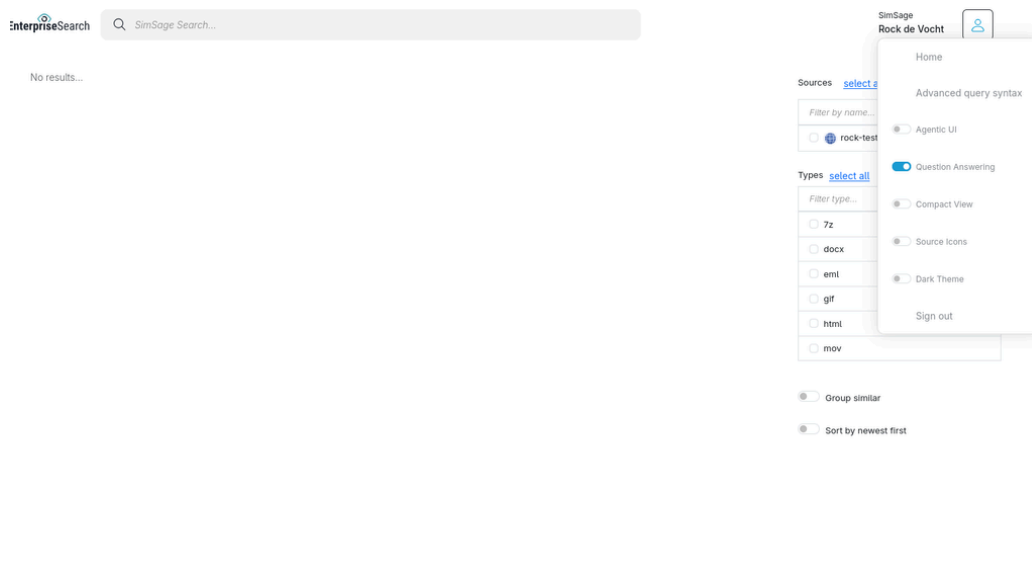
The default SimSage search interface will have extra features once a valid AI set up is set. An additional “Agentic UI” and “Question Answering” are available (these value can be set to be checked by default by SimSage).



Question Answering, when enabled, passes the user’s query and the top results of your search to your LLM to try and provide an answer.

Agentic UI is a different interface / view of SimSage that is more akin to the kind of views used by LLMs like OpenAI or Google Gemini.

Make sure these values are “selected”, depending on what you want to do.




Once selected, SimSage will automatically use the connected AI for question answering if possible (SimSage will detect if you are actually asking a question or are just using keywords).

EnterpriseSearch

SimSage
Rock de Vocht

4 results



integration-tests-2

DRAFT APRIL 6, 2006 NEWSLETTER

<https://dataset.simsage.co.uk/integration-tests-2/068239.pdf>

Sometimes those ideas begin with CDC scientists who have ideas, but lack the resources to implement and **test** them.

[create summary](#)

Group similar

Sort by newest first

Sources

Filter type...


- integration test 2
- rock test

A user searches for “test” - this is a single keyword and cannot be phrased as a question, so the search box question answer system is not used. However, once a user enters a proper query, this changes. Note the “create summary” button. This button is part of the SimSage AI system and provides a short summary for the given document.

EnterpriseSearch

SimSage
Rock de Vocht

one result



Ideas begin with CDC scientists who lack resources to implement and test them.


Group similar

Sort by newest first

Sources

Filter type...

- integration test 2
- rock test



integration-tests-2

DRAFT APRIL 6, 2006 NEWSLETTER

<https://dataset.simsage.co.uk/integration-tests-2/068239.pdf>


Sometimes those **ideas** begin with **CDC scientists** who have **ideas**, but lack the resources to **implement** and test them.

[create summary](#)

The search query “what ideas begin with CDC scientists?” does look like a question. This activates SimSage’s search box query answer system and responds accordingly.

EnterpriseSearch

3,136 results



test / test2 / glip

Demand For Contract Lawyers To Be Pushed Up By Market Uncertainty


<https://dataset.simsage.co.uk/test/test2/glip/demand-for-contract-lawyers-to-be-pushed-up-by-market-unc...>

Last modified 2015/12/16 09:02:00 | Demand for contract lawyers to be pushed up by market uncertainty

Demand for contract lawyers to be pushed **up** by **market** uncertainty .

[create summary](#)

+



test / test2 / glip

Dla Piper Launches 70 Strong Cannabis Law Team To Target 50bn Global Market

<https://dataset.simsage.co.uk/test/test2/glip/dla-piper-launches-70-strong-cannabis-law-team-to-target-50b...>

Last modified 2020/02/21 04:35:00 | DLA Piper launches 70-strong cannabis law team to target \$50bn global market

DLA Piper launches 70-strong cannabis law **team** to target \$50bn global **market** .

[create summary](#)

+

Group similar

Sort by newest first

Last modified:

1970/01/20 2025/01/20

Sources

Filter type...

- integration test 2
- rock test


Select the “speech bubble” icon after each search result to enter document Question and Answering mode for that specific document.

EnterpriseSearch


markets

SimSage
Rock de
Vocht


3,136 results



Demand For Contract Lawyers To Be Pushed Up By Market Uncertainty
https://dataset.simsage.co.uk/test/test2/glp/demand-for-contract-lawyers-to-be-pushed-up-by-market-uncertainty
Last modified 2015/12/16 09:02:00 | Demand for contract lawyers to be pushed up by market uncertainty .
create summary



DLA Piper Launches 70 Strong Cannabis Law Team To Target \$50bn Global Market
https://dataset.simsage.co.uk/test/test2/glp/dla-piper-launches-70-strong-cannabis-law-team-to-target-50bn-global-market
Last modified 2020/02/21 04:35:00 | DLA Piper launches 70-strong cannabis law team to target \$50bn global market
create summary



Drone Manufacture Market Predicted To Boom Despite Strict Regulation
https://dataset.simsage.co.uk/test/test2/glp/drone-manufacture-market-predicted-to-boom-despite-strict-regulation
Last modified 2015/07/16 10:57:00 | Drone manufacture market predicted to boom despite strict regulation The international drone market is

Converse with this Document

Generative AI can make mistakes. Consider checking important information.

Please ask me any question about "DLA Piper Launches 70 Strong Cannabis Law Team To Target 50bn Global Market" (https://dataset.simsage.co.uk/test/test2/glp/dla-piper-launches-70-strong-cannabis-law-team-to-target-50bn-global-market.pdf)

what is this document about?

This document announces DLA Piper's launch of a 70-person international cannabis law team to capitalize on the growing global medicinal cannabis market, projected to reach at least \$43 billion by 2024. The team, led from Canada and London, will offer services including corporate, IP, real estate, and technology law to clients ranging from investors to established companies. The announcement highlights the increasing legalization of cannabis in various jurisdictions and mentions other law firms pursuing similar opportunities in this market.

Your question...

We can now start a conversation with this document as shown above.