

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).

The screenshot shows the 'AI set up' tab selected in the top navigation bar. Under the 'Large Language Model (LLM)' section, the 'LLM Provider' dropdown is set to 'None'. A tooltip indicates that the number of tokens is exceeded. The 'OpenAI' provider is highlighted with a blue selection bar. Other providers listed include Google Gemini, Ollama (local), Amazon Bedrock, and CoPilot. Below the provider dropdown, there are fields for 'document summarization token count' (set to 1000) and 'search box question answering, the number of results to pass to the LLM' (set to 5). There is also a field for 'search box question answering, the number of tokens to pass for each search result' (set to 100). At the bottom, there is a note about token limits and a warning message: 'the maximum number of tokens per month before the model is disabled and a warning is emailed (if > 0)' followed by a field containing '0'. A 'Save' button is at the bottom right. A blue callout box labeled 'SimSage AI Setup Guide' points to the 'None' provider dropdown.

Open AI settings

The screenshot shows the 'AI set up' tab of the SimSage interface. On the left sidebar, there are links for Overview, User Management, demo knowledge-base, Document Management, and The Mind. The main content area is titled 'Large Language Model (LLM)' and shows 'OpenAI' as the provider. It includes fields for 'OpenAI key' (with a placeholder of 'XXXXXXXXXXXXXXXXXXXX'), 'Enabled (automatically disabled if maximum number of tokens is exceeded)', 'document Q&A token count' (set to 1000), 'document summarization token count' (set to 1000), 'search box question answering, the number of results to pass to the LLM' (set to 5), 'search box question answering, the number of tokens to pass for each search result' (set to 100), and a note about token usage. A dropdown menu for 'OpenAI model' shows 'gpt-4-mini' selected. A 'Save' button is at the bottom.

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

The screenshot shows the 'AI set up' tab of the SimSage interface. The sidebar and main content area are identical to the OpenAI setup, but the provider is listed as 'Google Gemini'. The 'Gemini model' dropdown menu shows 'Select Model' selected, with options for 'gemini-1.5-flash', 'gemini-1.5-pro', 'gemini-2.0-flash-lite', and 'gemini-2.0-flash'. A 'Save' button is at the bottom.

| 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.

The screenshot shows the SimSage AI setup interface. On the left, there's a sidebar with navigation links: Overview, User Management, demo knowledge-base, Document Management, and The Mind. The main area has tabs: AI set up (selected), Synonyms, Semantics, and Synsets. Under AI set up, it says "LLM Provider: Ollama (local)". A note below states: "NB. Ollama requires a GPU, please ensure you have a GPU available / set up." It lists requirements: "14b models require 9GB of GPU RAM", "Smaller models require 4GB of GPU RAM", and "CPU only is not sufficient for Ollama". There's a dropdown menu "Select Model" with options "llama3.2:3b" and "cogito:14b", where "llama3.2:3b" is selected. Below it, "document summarization token count" is set to 1000. For "search box question answering, the number of results to pass to the LLM", the value is 5. For "search box question answering, the number of tokens to pass for each search result", the value is 100. A note says "the maximum number of tokens per month before the model is disabled and a warning is emailed (if > 0)" with a value of 0. A "Save" button is at the bottom.

Open AI specific settings are

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

Amazon Bedrock

| 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.

| setting | description |
|---------|-------------|
|---------|-------------|

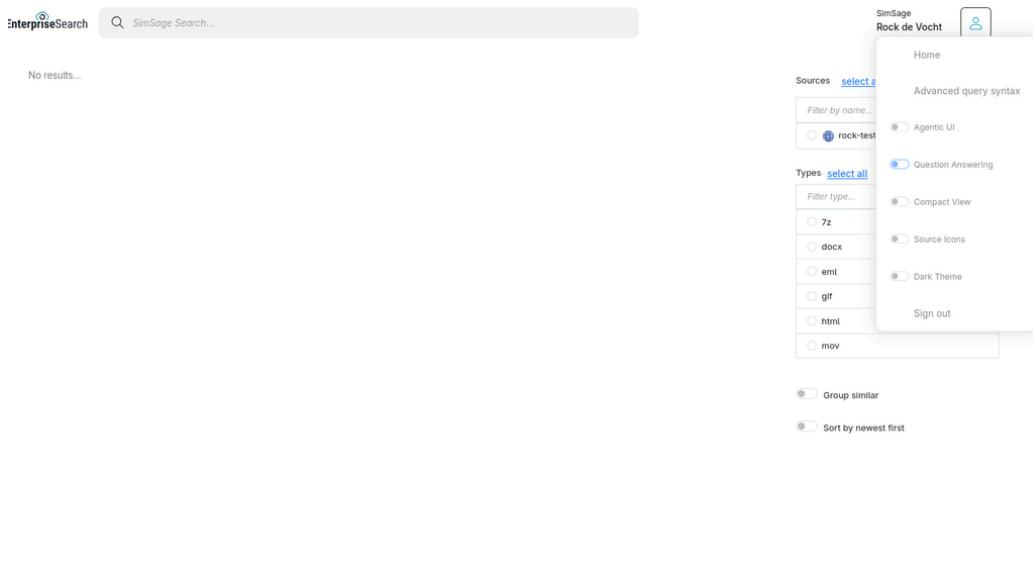
| | |
|------------------|--|
| CoPilot key | the CoPilot key to use, provided by Azure once you create a CoPilot instance. |
| CoPilot endpoint | the Azure endpoint for the CoPilot instance you created inside Azure (provided by Azure) |
| CoPilot model | the model to use for inference with CoPilot, 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 sent 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

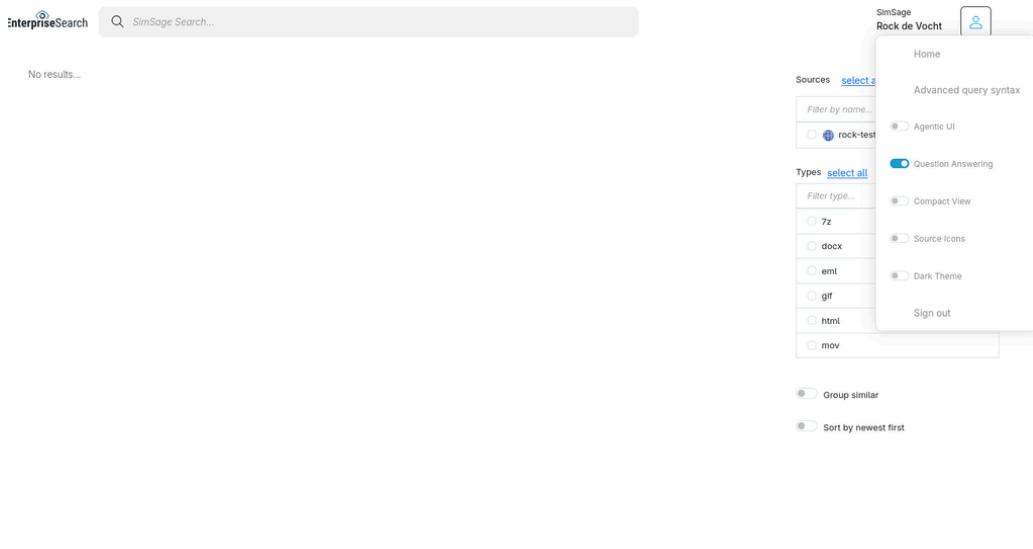
The default SimSage search interface will have extra features once a valid AI set up is set. An additional “Agentic UI” and “Question Answering” are available (these value can be set to be checked by default by SimSage).



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

DRAFT APRIL 6, 2006 NEWSLETTER

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

[create summary](#)

Sources

- 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.

DRAFT APRIL 6, 2006 NEWSLETTER

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

[create summary](#)

Sources

- integration test 2
- rock test

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 SimSage Rock de Vocht

3,136 results

Demand For Contract Lawyers To Be Pushed Up By Market Uncertainty

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

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

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

[create summary](#)

Sources

- Filter
- integration test 2
- rock test

Group similar

Sort by newest first

Last modified: 1970/01/20 2025/C

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

3,136 results



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

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

[create summary](#)



Dla Piper Launches 70 Strong Cannabis Law Team To Market
<https://dataset.simsage.co.uk/test/test2/gip/dla-piper-launches-70-strong-cannabis-law-team-to-market.pdf>
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/gip/drone-manufacture-market-predicted-to-boom-despite-strict-regulation.pdf>
Last modified 2015/07/16 10:57:00 | Drone manufacture market predicted to boom despite strict regulation

Drone manufacture **market** predicted to boom despite strict regulation The international drone **market** is

Converse with this Document X

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/gip/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.