onyx

Gen AI Configs › GenAI Overview

Gen AI Configs

GenAI Overview

Overview of the Generative AI functionality in Onyx

LLM Options

Onyx supports a large range of LLM hosting services and local/custom such as:

> For cloud providers, we support `OpenAI`, `Anthropic`, `Azure OpenAI`, `HuggingFace`, `Replicate`, `AWS Bedrock`, `Cohere`, and many others.
>
> For self-hosted, we support Ollama and GPT4All (or any other way you choose to provide an OpenAI-compatible API).

Note: Most of the different LLM support is provided by the **LiteLLM Langchain library** and are configured accordingly (see the following sections for some examples).

What are Generative AI (LLM) models used for?

The Large Language Models are used to interpret the contents from the most relevant documents retrieved via Search. These models extract out the useful knowledge from your documents and generates the **AI Answer**.

What is the default LLM?

Our default recommendation is to use `gpt4` from OpenAI or `Claude 3.5 Sonnet` from Anthropic. These are the most powerful and highest quality models available.

Azure OpenAI, Claude through Bedrock, or self-hosted Llama3.1 70B / 405B are also highly recommended.

Why would you want to use a different model?

- Use a cheaper, faster model (such as `gpt-4o`)
- Use a hosting service with a different data retention policy
  - Currently OpenAI and Azure OpenAI retain data for 30 days for monitoring against misuse
- Host the model yourself for complete control and flexibility
  - The Gen AI is the only feature in Onyx that reaches out to third party controlled service
  - There are options (see below) to avoid this entirely but at the cost of performance
- Use a different model perhaps finetuned or better suited for a particular domain of interest

Onyx LLM Configs

To setup various LLMs, head to the `LLM` page on the Admin Panel. A fun thing about Onyx is that you can setup multiple LLM providers at the same time! This allows you use different models for different assistants and play to each LLM's strengths.

See the next sections for some examples on how to configure different providers.

As always, don't hesitate to reach out to the **Onyx team** if you have any questions or issues!

Chrome Extension OpenAI

onyx

Powered by Mintlify