How to set up your own LLM with alby

Last Modified on 07/09/2025 3:36 pm EDT

This article will outline the benefits, risks, and limitations to make a decision about whether you should incorporate your own large language model (LLM) account instead of using alby's LLM. When integrating your LLM with alby, your organization will maintain control over the credentials and model access.

alby allows you to integrate the following LLMs:

- OpenAI Enterprise
- Azure Open AI
- Google Vertex AI

	Cost	Security	Ease of maintenance
Use your own LLM	You're required to cover costs related to the LLM calls	You control the data security configuration of the LLM	Varies based on LLM used, but you're required to maintain it
Use alby's LLM	alby controls and manages costs related to LLM calls	alby manages the data sent to its LLM and the security configurations of those services	alby manages and maintains the LLMs used

Benefits

Using your own LLM credentials could offer the following benefits:

- Greater control of the data managed and maintained by the LLM, such as
  - Any completion tasks (such as chat-based functionality)
  - Customer messages and product data discussed in a conversation
  - Encapsulating Personally Identifiable Information (PII) so it won't leave your environment
    - By default, alby automatically scrubs all PII from any user submitted requests
- You may already have a private deployment of an LLM, making integration with alby an easy extension of your existing capabilities.

Risks and limitations

Rate limiting

When using your own LLM provider, you are in charge of your own rate limits, which need to be suitable for your traffic volume.

Low rate limits may cause errors during high-traffic periods, so careful planning is needed to ensure a smooth experience for customers.

It can be challenging to predict required capacity based on factors like:

- Website traffic
- User interaction rate
- Product catalog size

Alby has implemented techniques and technologies to identify and reduce bot traffic, Denial of Service attempts, and other malicious activity that may impact your resource utilization. We have also implemented techniques to cache responses and attempt to reduce LLM API usage. However, high traffic scenarios may still present significant spikes in usage.

Improper planning for resource spikes may result in the unavailability of the alby system.

Azure considerations

If you use Azure OpenAI as an LLM provider, there are some additional requirements to consider.

Through Azure, each model can only be used in one deployment, meaning each model will need separate credentials.

For example, if you set up your Azure OpenAI service with one LLM model and want to change it, when a better model is released, you will need to set up a new configuration with the new model and reconnect your credentials through alby.

This differs from OpenAI Enterprise and Google Vertex AI, where alby would have access to all available models with your credentials.

Prerequisites

Depending on the LLM you use, different providers have different requirements to set up and use with alby.

OpenAI Enterprise

Use your OpenAI API key to configure your LLM integration.

Azure Open AI

To use your Azure OpenAI service, you will need the following information:

- Azure Endpoint
- Base Model
- Deployment Name
- API Version
- Azure OpenAI API version
- API Key

Google Vertex AI

To use Google Vertex AI with alby, you will need to upload a JSON file, or paste your JSON credentials into the input modal.

You also need your Location and Project.

Configuration

To configure your own LLM with alby:

1. Navigate to your alby admin panel by clicking on your brand name.
2. Navigate to **Settings**
3. Under Developer Settings, click **LLM Provider**.
4. Choose your provider.
5. Complete the configuration with your provider.

You can delete and reconfigure your LLM settings as needed.