GEBO.AI SETUP GUIDE · MONITORING · PAGE 20

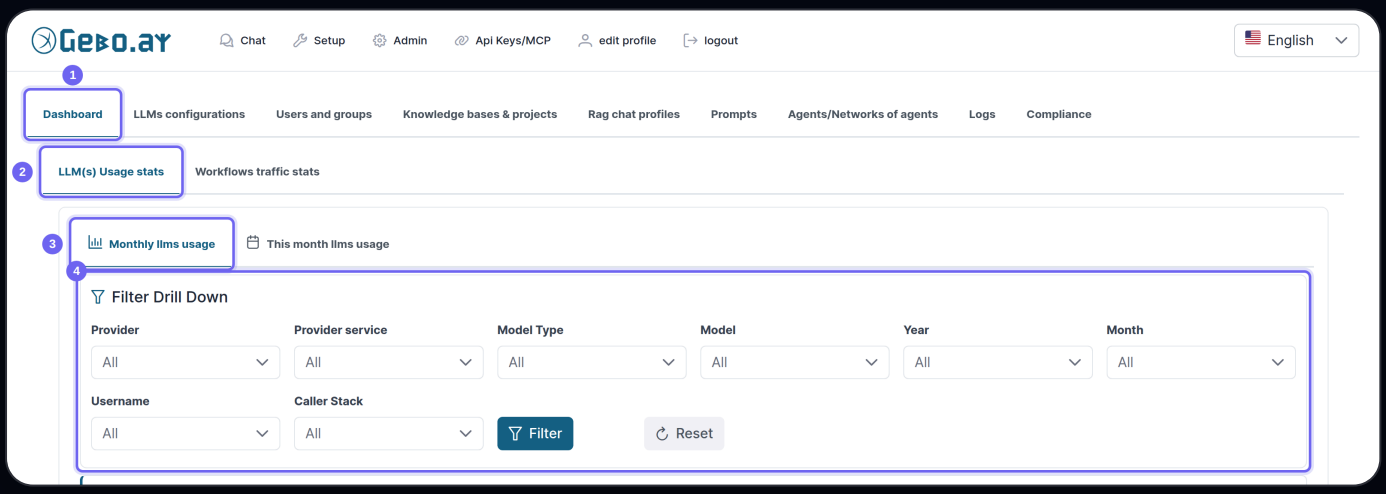
LLM usage and costs

How much each model is used, for whom and by which part of Gebo.ai, how fast it answers and, once the models have prices, what it costs.

WHERE	Admin → Dashboard → LLM(s) Usage stats.
YOU NEED	An administrator account.

1 Open the statistics

In **Admin** the **Dashboard** tab (1) opens on **LLM(s) Usage stats** (2). **Monthly llms usage** (3) shows the months one by one; **This month llms usage** shows the current month day by day.



The usage statistics and their filters (4).

2 Narrow it down

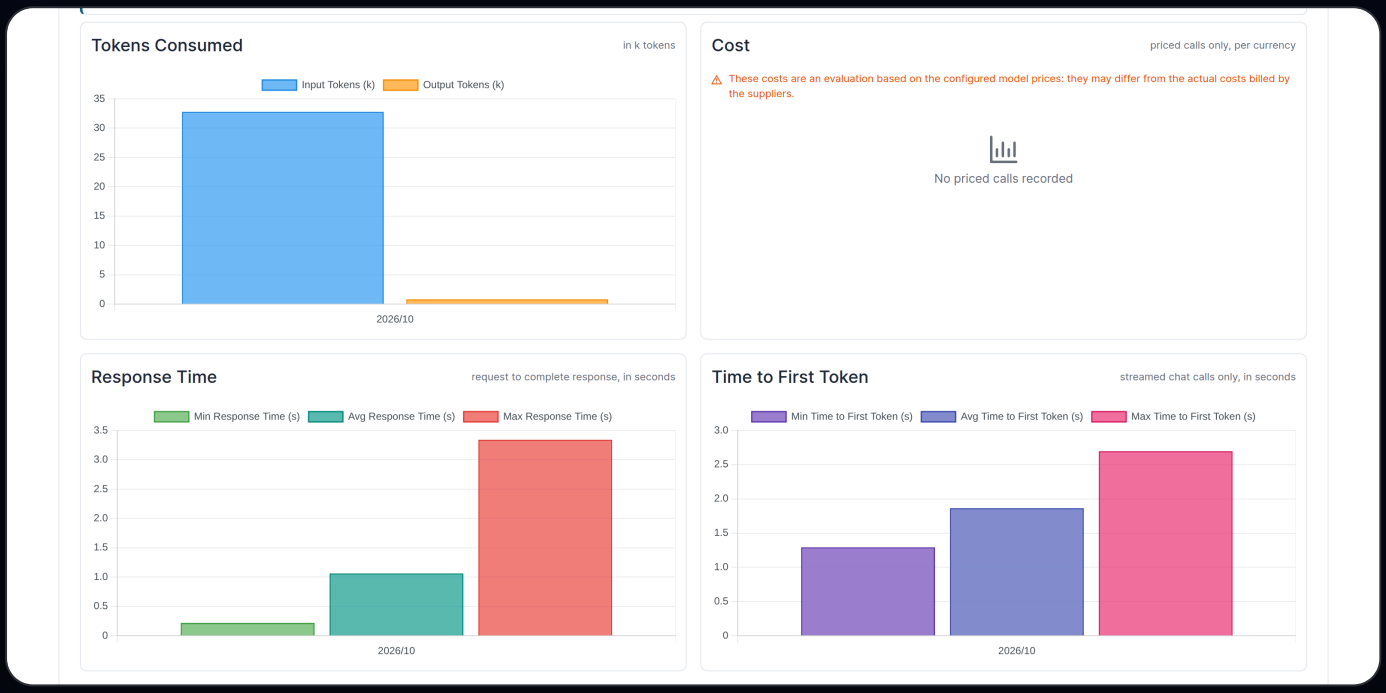
The filters (4) narrow every chart. Choose the values and click **Filter**; **Reset** clears them.

- **Provider** and **Provider service:** the company that runs the model and which of its services was called.
- **Model Type** and **Model:** chat, embedding and the other kinds of model, then the model itself.
- **Year** and **Month**.

- **Username:** who the call was made for.
- **Caller Stack:** which part of Gebo.ai made the call, as the names of its Java packages. It tells the chats apart from the reading of documents, for example.

---

3 Read the charts



Tokens, cost, response time and time to the first token, month by month.

- **Tokens Consumed:** the tokens sent to the models and received from them, in thousands.
- **Cost:** an estimate from the model prices set in **Setup → AI Models, agents & interoperability → Provider deals and model prices**. It adds up the calls of the models with a price, with one set of bars for each currency.
- **Response Time:** from the request to the complete answer, in seconds: shortest, average and longest.
- **Time to First Token:** how long a chat answer takes to start appearing, in seconds.

> **No prices, no cost**
>
> Until the models have prices, **Cost** reads *No priced calls recorded*: page 20b sets them.

← Administrative API keys

Next: Provider deals and model prices →