# R.a.g. autotune

How close a document fragment must be to a question for the chats to use it is the similarity threshold. The right value depends on the embedding model and on your documents: R.a.g. autotune measures it on your own documents and keeps it current as they change.

| | |
|---|---|
| WHERE | The configuration file of the installation; Setup → Knowledge & chat → R.a.g. autotune for the results. |
| YOU NEED | An embedding model with documents indexed (steps 3-5), a chat model for the internal services, and access to the server. |

## 1 What it does

- It samples about thirty fragments representative of your documents and has the internal services chat model write up to twelve questions on them.
- It tries similarity thresholds, searches the questions with each of them and has the model rate what each threshold retrieves.
- It keeps, for each embedding model, the best rated threshold that still answers at least two thirds of the questions.

The chats with knowledge bases use the tuned threshold from their next search, as do the knowledge base tools of the agents and the deep search; a chat profile with its own thresholds keeps them (page 30).

It runs by itself, at startup, every four hours and after each ingestion, and tunes an embedding model when it has no result yet, when its last result is three days old, or when its indexed fragments changed by 5% or more. A run makes a few hundred short calls to the internal services chat model: on a small knowledge base it takes a few minutes.

## 2 Switch it on

In the configuration file of the installation, as on page 28b, set:

```
ai.gebo.rag-threashold-autotune.config.enabled: true
```

With Docker Compose the `application.yml` the image ships already has the line, set to `false`: change it in your copy. In an `application.properties` file (Linux package, Windows) add:

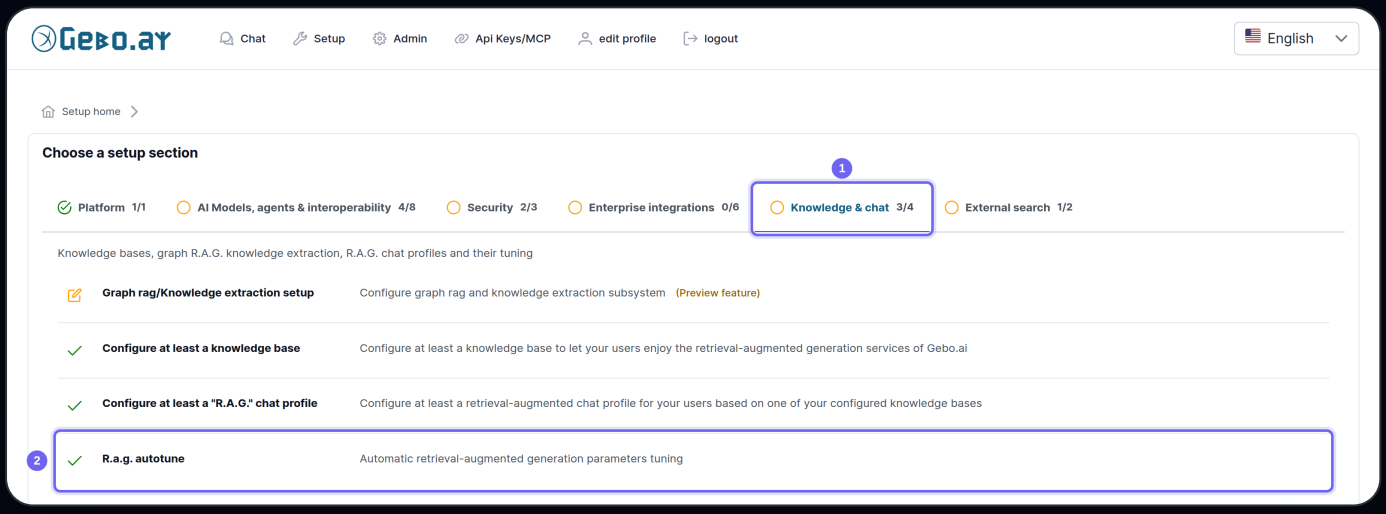
```
ai.gebo.rag-threashold-autotune.config.enabled=true
```

Restart Gebo.ai: the first tuning starts a few seconds later.

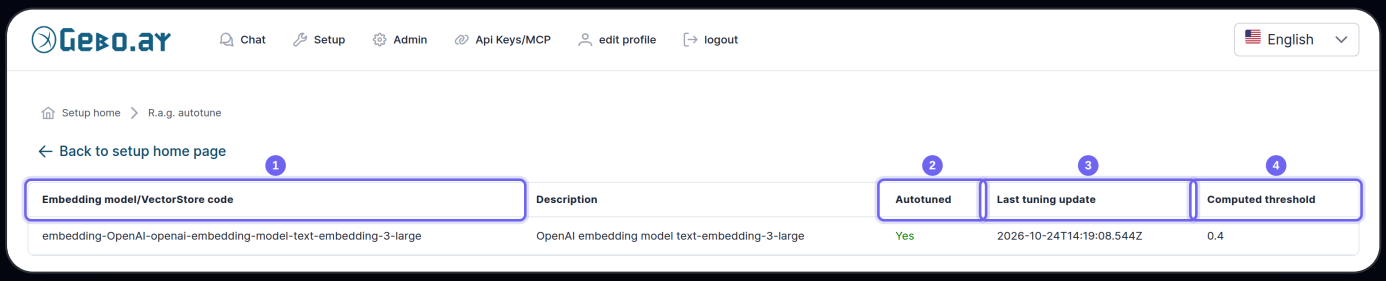
---

## 3 The results

In **Setup** open **Knowledge & chat** (1) and click **R.a.g. autotune** (2).




The Knowledge & chat tab (1) and R.a.g. autotune (2).




One row for each embedding model (1): whether it is tuned (2), when (3) and the threshold computed (4).

Each embedding model has its row (1), with **Autotuned** (2), the **Last tuning update** (3) and the **Computed threshold** (4). The section shows its green check once every embedding model is tuned.

## 4 Fine tuning

Every setting has a default that suits most installations; to change one, add it under the same prefix in the configuration file, for example:

```
ai.gebo.rag-threashold-autotune.config:
  enabled: true
  day-elapsed-without-tuning: 7
  autotune-sample-fragments: 40
```

- `day-elapsed-without-tuning` (3): the days after which an embedding model is tuned again.
- `documents-cardinality-added-percent-trigger` (5): the change of the indexed fragments, in percent, that tunes it again.
- `autotune-sample-fragments` (30): the fragments the questions are written from; more fragments, a wider sample and more calls.
- `autotune-max-generated-questions` (12): the questions written and rated.
- `minimum-answered-questions-share` (0.66): the share of the questions a threshold must still answer to be chosen.
- `sample-fragments-min-token-length` (100): the shortest fragment, in tokens, taken in the sample.
- `sample-candidates-per-fragment` (3), `sample-representativeness-probe-top-k` (20): how widely the representative fragments are chosen.
- `autotune-question-chunks-per-call` (5): the fragments sent to the model in one question writing call.
- `autotune-samples.phrases`, with `autotune-samples.default-threashold` (0.5): phrases of your domain to sample the fragments around, instead of sampling across all the documents.

← Chat profiles

Next: Deep search →