ELLM

Managing Models in the eLLM Admin Console

The eLLM Admin Console allows administrators to manage AI models by registering, editing, and controlling their usage and costs. It supports both local and cloud models, enabling flexible management without requiring a system restart.

AUTHOR
Paul Flanders

PUBLISHED
8 Jun 2026

DOCUMENT
InfoHub — Knowledge, news, and updates from EssingtonITS

Managing Models in the eLLM Admin Console

Paul Flanders · 8 Jun 2026

What this product is for

eLLM does not answer questions on its own; it sends them to an AI model. The models page lets administrators register one or more models, describe what each is good at, control who may use it, and manage its cost and behaviour. This gives you a fleet of models the assistant and your users can choose from.

Main features

- Register OpenAI-compatible model backends and remove them.
- Mark each model as **Local** (on your own servers) or **Cloud** (an external provider).
- Give each a display name and capability tags, and set a maximum reply length.
- For cloud models, add a per-model instruction, set costs, restrict to groups, and control auto-routing.
- Edit a model's settings in place, and enable or disable a model.
- Set the default model that the basic chat modes resolve to.

How to use it

Registering a model

- Open the Models page in the admin console.

- Add a model by providing its backend address, the served model name, and an access key if required.
- Choose its tier: **Local** or **Cloud**.
- Give it a friendly display name and, optionally, capability tags and a maximum output length.
  Save. The model is added live and becomes available without a restart.

The table is split into Local and Cloud sections so you can see your fleet at a glance.

Cloud model options

Cloud models have a few extra settings:

- **Pre-context:** a per-model instruction added to the prompt whenever that model answers.
- **Input and output cost:** the price per million tokens, used for spend tracking.
- **Include in auto-routing:** whether the automatic router may pick this model. You can opt a model out.
- **Restrict to groups:** limit the model to chosen groups.
- **Newer reasoning model flag:** tick this for newer reasoning-style models (such as the latest GPT or o-series) that need a different request format; without it they may be rejected.

Editing, enabling, and disabling

- Use the edit (pencil) action to change a model's display name, pre-context, capabilities, maximum output, auto-routing, tier, and group restrictions in place. The fixed identity (backend address, served model name, access key, and pricing) stays as set when registered.
- Use the enable/disable (pause/play) action to take a model out of service temporarily. A disabled model stays registered but is hidden from the chat picker, the automatic router, and the external API, and cannot be selected until re-enabled.

Setting the default model

On the models page you choose which registered model the basic chat modes resolve to: the default the chat picker and the simple external-API modes fall back to. This lets you point the everyday experience at the model you prefer.

Common tasks

- **Add a new on-premises model:** register it as Local with a display name and capabilities.
- **Add an external provider:** register it as Cloud, set its costs, and restrict it to the right groups.
- **Take a model offline for maintenance:** disable it; re-enable when ready.
- **Adjust how a model behaves:** edit its pre-context, capabilities, or output cap in place.
- **Change the everyday default:** set the default model.

Things to know

- Models are managed live: registering, editing, enabling, and disabling all take effect without a restart, within a short refresh window.
- A model restricted to groups only appears for users who belong to one of those groups; administrators see all models.
- A disabled model is fully out of service: hidden from the picker, skipped by the router, unavailable on the external API, and refused if explicitly selected.
- The fixed identity and pricing of a model are set at registration and are not changed by later edits.
- For cloud cost tracking to be accurate, set the input and output costs correctly when registering.

Troubleshooting

- **A newly added cloud model returns errors:** check the served model name is correct, the access key is valid, and, for newer reasoning models, that the reasoning-model flag is ticked.
- **A model isn't appearing for users:** confirm it is enabled and not restricted to groups the users aren't in.
- **"No model available" type errors:** a model that keeps failing may be temporarily set aside by the gateway; fix its configuration, then it will be used again.
- **I can't change a model's backend address:** that identity is fixed at registration. Remove and re-add the model to change it.

Frequently asked questions

What is the difference between Local and Cloud?

Local models run on your own servers; cloud models are external services you connect to. See the Local models and Cloud models articles.

What does disabling a model do?

It keeps the model registered but takes it out of service everywhere until you re-enable it.

What is pre-context?

A per-model instruction added to the prompt whenever that model answers, useful for tailoring a cloud model's behaviour.

Why would I opt a model out of auto-routing?

To keep the automatic router from choosing it, for example an expensive cloud model you only want used when picked deliberately.

Do I need to restart anything after changes?

No. Changes apply live within a short refresh window.

Summary

The models page lets administrators build and manage the fleet of AI models ELLM uses. Register local and cloud models, describe and price them, restrict them to groups, and enable or disable them live. Cloud models add per-model instructions, costs, and routing controls, and you choose the default model that the everyday chat experience falls back to.

> Need assistance navigating the complexities of eLLM? EssingtonITS offers expert guidance and tailored IT solutions to help you succeed. Visit EssingtonITS.co.uk for comprehensive support.