ELLM

Local Models in eLLM

Discover how local AI models in eLLM keep your organisation's data private by running on your own servers. Learn about their features, setup, and benefits, including cost savings and enhanced data security.

AUTHOR
Paul Flanders

PUBLISHED
8 Jun 2026

DOCUMENT
InfoHub — Knowledge, news, and updates from EssingtonITS

Local Models in eLLM

Paul Flanders · 8 Jun 2026

What this product is for

Many organisations need their information to stay in-house. Local models make that possible: the AI that answers your questions runs on your own hardware, so the content of your questions and documents does not leave your premises. ELLM is built to work fully with local models alone.

Main features

- The AI runs on your own servers, keeping data in-house.
- No per-question charges; local models are free to run once set up.
- They appear under Local models in the chat model picker.
- Automatic routing prefers them where they fit, to keep costs down.

How it works

From a user's point of view

When you open the chat model picker, models are grouped into Local models and Cloud models. Local models are the on-premises ones. You can pick a local model just like any other, or let automatic routing choose one for you.

From an administrator's point of view

Administrators register local models on the admin console's Models page, marking each as the Local tier. Each can be given a friendly name, capability tags describing what it is good at, and a maximum reply length. See the Managing models article.

Common tasks

- **Keep everything on-premises:** use a local model, or use automatic routing, which prefers them.
- **Pick a specific local model:** choose it from the Local models group in the picker.
- **Add a local model (administrators):** register it as the Local tier on the Models page.

Things to know

- Local models keep the content of questions and documents within your organisation.
- They are free to use once running, which is why automatic routing prefers them.
- Their capability and speed depend on your hardware and the model chosen.
- eLLM can run entirely on local models; cloud models are an optional addition.

Troubleshooting

- **No local model appears:** an administrator needs to register one on the Models page.
- **A local model seems slow or limited:** performance depends on your hardware and the model; an administrator can review the setup.
- **I want a more capable answer:** consider a cloud model if your organisation allows it; see the Cloud models article.

Frequently asked questions

What makes a model "local"?

It runs on your organisation's own servers rather than an external provider's.

Does using a local model keep my data private?

Yes. The content stays within your organisation.

Do local models cost anything per question?

No. They are free to run once set up, which is why routing prefers them.

Can ELLM work without any cloud models?

Yes. It is designed to work fully on local models alone.

Summary

Local models run on your own servers, keeping your information private and free of per-question costs. They appear under Local models in the chat picker, and automatic routing favours them. ELLM works fully with local models alone, with cloud models available as an optional extra.

> 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.