
ELLM

Automatic Model Routing in eLLM

ELLM can choose the best AI model for a question on your behalf. This is called automatic routing. This article explains what it does, when to use it, and how ELLM decides.

AUTHOR

Paul Flanders

PUBLISHED

8 Jun 2026

DOCUMENT

InfoHub — Knowledge, news, and updates from EssingtonITS

Automatic Model Routing in eLLM

Paul Flanders · 8 Jun 2026

What this product is for

When more than one model is available, picking the right one for each question takes knowledge and effort. Automatic routing removes that burden: you let eLLM choose, and it weighs each model's strengths, access rules, and cost to pick a good fit. It is ideal when you would rather not think about models and simply want a sensible answer.

Main features

- eLLM selects a suitable model for each question automatically.
- It respects which models you are allowed to use.
- It favours models that match the question, such as code or reasoning tasks.
- It prefers on-premises models, which are free, over paid cloud models, and lower-cost options where it can.
- It can pick a model that understands images when your question includes one.

How it works

Turning it on

In chat, simply pick automatic routing if your administrator has made it available, or use the everyday default. Through the external API, choose the `ellm-auto` mode. eLLM then selects the model for you.

How eLLM chooses

For each question, eLLM considers, in order:

- **Access:** it only considers models you are allowed to use, and skips any an administrator has opted out of routing or disabled.
- **Images:** if your question includes an image, it requires a model that can understand images.
- **Fit:** for questions that look like coding or reasoning tasks, it leans towards models tagged as good at those.
- **Cost:** it prefers on-premises local models (which are free) over cloud models, then favours lower-cost options.

If no suitable model can be found, eLLM falls back to the everyday default model so you still get an answer.

Common tasks

- Get a good answer without choosing a model: use automatic routing.
- Ask a coding or reasoning question: let routing favour a capable model automatically.
- Send an image: routing will pick a model that can read it.
- Keep costs down: routing prefers free local models where they fit.

Things to know

- Routing only ever chooses from models you are permitted to use.
- Administrators can opt a model out of routing, so it is only used when picked deliberately, which is useful for expensive models.
- Disabled models are never chosen.

- Routing aims for a sensible balance of fit and cost; it is not guaranteed to pick the single most powerful model every time.
- If you want a specific model, you can always pick it yourself instead of using routing.

Troubleshooting

- Routing keeps choosing a local model: that is intended: local models are free and preferred where they fit. Pick a specific model if you want a particular cloud one. An image question used an unexpected model: routing must pick an image-capable model; only those were eligible.
- I expected a particular cloud model: it may be opted out of routing or restricted to other groups. Select it directly, or ask your administrator.
- Routing fell back to the default: no eligible model matched, so eLLM used the default so you still got an answer.

Frequently asked questions

Does routing always pick the best model?

It picks a good fit by balancing capability, access, and cost. It favours free local models, so it won't always choose the most powerful or most expensive option.

Can I override it?

Yes. Choose a specific model yourself instead of using automatic routing.

Will it use a model I'm not allowed to use?

No. It only considers models your groups permit.

What happens with images?

Routing selects a model that can understand images so your picture is handled properly.

Summary

Automatic routing lets eLLM choose a sensible model for each question, respecting your access, favouring capable models for the task, and preferring free local models to keep costs down. Use it when you would rather not pick a model yourself, and switch to a specific model whenever you want direct control.

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.