ELLM

Scaling eLLM

eLLM is built to grow with you. You can start on a single server and, as demand rises, add more AI capacity in several independent ways without rebuilding the system or moving your data. This article explains the ways eLLM scales: running several local models, reaching out to cloud models, letting eLLM pick the right model automatically, and turning ordinary desktops into a private pool of AI workers with Edge Mesh.

AUTHOR
Paul Flanders

PUBLISHED
8 Jun 2026

DOCUMENT
InfoHub — Knowledge, news, and updates from EssingtonITS

Scaling eLLM

Paul Flanders · 8 Jun 2026

What this product is for

Demand on an AI assistant is rarely steady. More people join, documents pile up, and some tasks are far heavier than others. Rather than forcing you to buy one very large server up front, eLLM lets you add capacity in the place it is actually needed more models, occasional cloud help for the hardest questions, or spare desktop computers for background work. The goal is simple: the more capacity you add, the more the platform can do, and you always keep a sensible fallback so nothing breaks when a piece is busy or switched off.

Main features

- **Many local models at once: r**egister several on-premises models and use them side by side for different needs.
- **Cloud models on tap:** bring in a cloud model for the occasional task that benefits from extra power, under your control and budget.
- **Automatic routing:** let eLLM choose a sensible model for each question, balancing fit, access, and cost.
- **Edge Mesh compute:** turn approved desktops into a private worker pool that takes on background AI work, the more machines you add, the faster that work gets.
- **Always a fallback:** if a model, the cloud, or the mesh is unavailable, eLLM keeps working using your central setup.
- **Grow without rebuilding:** every step above is additive your documents, settings, and history stay exactly where they are.

How it works

Start small, grow without rebuilding

A small deployment can run perfectly well on one server with a single local model. As your needs grow, you add capacity rather than replacing what you have. Each of the methods below can be introduced on its own, in any order, and combined freely. Your knowledge base and configuration are untouched by adding capacity.

Running several local models

An administrator can register more than one on-premises model and have them all available at the same time. This is useful because different models suit different jobs a small, fast model is ideal for quick everyday questions or high volumes of background work, while a larger model handles harder reasoning. Users can pick a model in chat, or let automatic routing choose. Adding models spreads load and lets you match the right tool to each task without sending anything off your premises.

Reaching for the cloud when it helps

Some questions benefit from a very capable cloud model. eLLM lets an administrator register cloud models alongside local ones, with full control: which groups may use them, a monthly spending cap, and sensitive-information safeguards applied before anything leaves your network. Cloud models are entirely optional many organisations run on local models alone but they are there for the occasional heavyweight task. Because local models are free and private, eLLM treats the cloud as a top-up, not the default.

Choosing the right model automatically

With several models available, deciding which to use for each question is a job in itself. Automatic routing does it for you: for each question it considers what you are allowed to use, whether the question includes an image, whether it looks like a coding or reasoning task, and cost preferring free local models where they fit. This means you can keep adding models

and let eLLM make good use of them without anyone having to choose each time. See the Automatic model routing article for detail.

Edge Mesh compute: turning desktops into AI workers

Edge Mesh is the way eLLM scales by using hardware you already own. Approved Windows and Linux desktops install a lightweight worker and join a private pool. To the rest of eLLM this pool looks like a single extra model, but behind it many machines share the work. The mesh takes on background tasks — not your live chat, which always stays fast on the central server such as scoring documents for search quality, generating keywords, topics and summaries for documents, sorting documents into categories, building an overview of a whole document set, running deeper safety checks on what users ask, and even sharing out the work of indexing new documents so a large library imports faster. Administrators choose which of these run on the pool, and can steer heavier jobs to more capable (GPU) machines. Administrators manage all of this from **Configuration → Edge Mesh** — see Edge Mesh managing the worker pool.

The key idea is that the mesh gets more powerful the more machines you add. A large batch of background work is split into pieces and run across all the available desktops at once, so a job that took minutes on one machine finishes far quicker across several. An administrator can set how the mesh is used — how many machines must be available before it is used at all, and how large a job must be to be worth sharing out and every task can be turned on or off individually. If no desktops are available, or the mesh is switched off, the work simply runs on the central server as before.

Edge Mesh can also make safety checks more robust. For sensitive screening, eLLM can ask several worker machines to judge the same query independently and take a majority decision, so no single machine's mistake decides the outcome. As with throughput, adding machines here improves the result more independent opinions mean a safer call.

Common tasks

- **Handle more users:** register additional local models and let automatic routing spread the load.
- **Take on a heavyweight question now and then:** add a cloud model with a spending cap and restrict it to the groups that need it.
- **Speed up background work as you grow:** enrol more desktops into Edge Mesh; batch jobs finish faster the more you add.
- **Keep everyday chat fast:** nothing you do to the mesh affects live chat it always runs centrally.
- **Strengthen safeguarding:** enable safety consensus so several machines vote on sensitive checks.

Things to know

- Every scaling method is optional and additive you can adopt them in any order and combine them.
  Interactive chat is always served centrally and is never slowed by background mesh work.
- Local models are free and private; the cloud is an optional top-up, governed by access rules and a budget cap.
- The mesh only ever processes work that has already passed your access and permission rules; desktops never gain access to anything a user could not already see.
- If a model, the cloud, or the mesh is unavailable, eLLM falls back to your central setup so users still get answers.
- Adding capacity does not move or rebuild your data your documents, settings, and history stay in place.

Troubleshooting

- **Background jobs feel slow:** check how many desktops are connected and available on the admin Edge Mesh page; add more, or confirm the relevant task is enabled.
- **The mesh does not seem to be used:** an administrator may have set a minimum number of machines or a minimum job size before it engages, or the task may be switched off. Small jobs and empty pools fall back to the central server by design.
- **A cloud model is not selected:** it may be restricted to certain groups, opted out of automatic routing, or over its budget cap. Pick it directly or check with an administrator.
- **Everything routes to a local model:** that is intended local models are free and preferred where they fit. Choose a specific model if you want a cloud one.
- **A desktop dropped out:** the mesh notices silent machines, stops sending them work, and re-runs anything they were doing elsewhere, so jobs still complete.

Frequently asked questions

Do I need a bigger server to support more users?

Not necessarily. You can register additional local models, add the occasional cloud model, and enrol spare desktops into Edge Mesh all without replacing your central server.

Does adding more desktops really make it faster?

Yes, for background work. A batch is split across the available machines and run at the same time, so more machines finish the same job sooner. Live chat is unaffected it always runs centrally.

Will using the mesh or the cloud put my data at risk?

The mesh only handles work that has already passed your access rules, and worker desktops keep nothing. Cloud models are optional, restricted by group, budget-capped, and have sensitive-information safeguards applied before anything is sent.

What happens if a piece is switched off or busy?

eLLM falls back to your central setup. The cloud, extra local models, and the mesh are all enhancements layered on top of a system that works on its own.

Do I have to choose which model handles each question?

No. Turn on automatic routing and eLLM picks a sensible model for each question, respecting access and preferring free local models.

Summary

eLLM scales by adding capacity where you need it rather than forcing a single large purchase. Run several local models for everyday breadth, add cloud models for the occasional heavyweight task under strict control, let automatic routing make good use of them, and turn idle desktops into a private Edge Mesh that speeds up background work and even strengthens safety checks the more machines you add. Every method is optional, additive, and backed by a central fallback, so you can grow steadily and safely without rebuilding or moving your data.

> 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.