AI Explained

Open Weight v Closed Weight Models

Part 2 – Running AI in-house and using your own documents

By Dr Peter Fields

THE 2-MINUTE SUMMARY

In-house hosting can mean the firm's own servers or a cloud account it controls. The model is one part of the setup. The firm also needs to know who operates the app and where the files, search index, logs and backups go.

A tool can look up the firm's documents as it answers, using an open- or closed-weight model. The model's weights stay the same. Fine-tuning can change how the model responds, for example by teaching it the firm's way of classifying contract issues.

A firm may gain an advantage from the quality of its records, its judgement about which examples are worth reusing, and how well it turns those examples into better day-to-day work. It can do that with a model other firms can buy.

Start with one job and a set of documents the firm is entitled to use for it. Check whether the tool improves the work, what it costs after review time is counted, and whether it respects who can see the material.

What would you want it to know?

A contract tool can find the notice period in a draft. Now ask it something harder. "What have we agreed in similar deals, and when did we insist on a different clause?" To answer that, the tool needs the right precedents and some record of why earlier deals turned out as they did.

Part 1 covered access to the model itself. Here the firm has two decisions to make. Where should the tool run, and how should it use the firm's own records? You can bring those records into the answer whether the model has open or closed weights.

What running it in-house involves

An in-house setup can run on the firm's servers or in a cloud account it controls. OneAdvanced and AWS describe running Meta's open-weight Llama 4 Maverick in OneAdvanced's London AWS account, with separate document storage, search and screening. The firm's control depends on who runs and can access each part, including the logs and backups, and which requests call an outside service. [1]

Someone has to keep the software and its safeguards up to date and make sure it works when several people use it at once. When comparing that with a managed service, count the equipment or cloud charges, staff time and the cost of checking the answers.

What does it mean to learn from a document?

Suppose the firm has a precedent with a note explaining when it agreed a longer notice period. The tool can find that note and pass it to the model with the question. This is retrieval-augmented generation, or RAG. The model's weights stay the same. Whether the app keeps a copy of the note, prompt or reply is a separate question. Update the precedent and the index, and the next search can use the new version. See Knowledge, Search and Tools for more on retrieval. [2]

Fine-tuning uses further training to change how a model tends to respond. A firm could use checked examples to teach it how to flag departures from its standard terms. A supplier may offer fine-tuning for a closed model. With access to weights and the right licence, the firm can also adapt a copy itself. One method, LoRA, trains a small set of additional weights and leaves the originals fixed. [3]

Geek Box – Can the firm carry on training the model?

With access to the weights, suitable equipment and permission, a firm can train an existing model on more text. Gururangan and colleagues found gains on some classification tasks with smaller, earlier models. Mistral now offers training on firms' own material through Forge. That shows the option exists. Whether it improves a particular job is something to test. [4]

What if we put everything through it?

Put every email and file through the system and it will see abandoned drafts alongside sound advice. A signed contract records where the parties ended up. An approved note explaining why they accepted an unusual clause may be worth more to the next team than the entire thread.

A UK firm reusing personal data needs a lawful basis and must consider the rules on further use. Staff monitoring raises questions of necessity, proportionality, transparency and impact. For solicitors and firms it regulates, the SRA expects client data to be kept out of model training unless that use is both explicitly authorised and appropriate. [5–7]

A search can check who is asking and restrict results to the matters they may see. It can also stop returning a withdrawn document. A shared model trained on client material cannot be relied on to apply those permissions to what it learnt from its training. Removing a file from the document store leaves the trained weights as they were. That is a reason to keep sensitive material in a permission-controlled search collection. [2, 7]

Where might the advantage lie?

Suppose two firms use the same model for contract review. One can give it a current precedent and a considered account of when an exception was justified. The other can give it the draft agreement alone. The first has a better basis for advice, provided the right material reaches the model and someone checks the result.

Morgan Stanley used OpenAI's closed-weight models for its adviser tools. OpenAI describes testing answers with sample questions and improving retrieval as the document collection grew. The companies' accounts show how an externally supplied model can make use of a firm's own records. The durability of any advantage is a separate question. [8]

Brynjolfsson, Li and Raymond studied 5,172 customer-support agents using a tool fine-tuned on past conversations, with extra weight given to those of top performers. Agents resolved 15% more issues per hour on average. Less experienced and lower-skilled agents benefited most. The most skilled saw little speed gain and some decline in quality. That is evidence about one support operation. [9]

How to use it better

Hold the model fixed and give it the same draft contracts twice, once with the firm's precedent search and once without. Use examples held back from development. Count missed issues, wrong citations and the time a reviewer spends putting things right.

If access to precedents improves the work, measure whether the gain justifies the cost of maintaining the collection. If the tool keeps mishandling the same task even with the right material in front of it, there is a specific reason to test fine-tuning.

Sources and notes

1. OneAdvanced and AWS (12 August 2026), How OneAdvanced deployed over 50 AI agents on UK-sovereign AWS. Joint case study

2. Patrick Lewis, Ethan Perez, Aleksandra Piktus and colleagues (2020), Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, NeurIPS, §§2 and 4.5. Paper

3. Edward J. Hu, Yelong Shen, Phillip Wallis and colleagues (2022), LoRA: Low-Rank Adaptation of Large Language Models, ICLR, §§4.1–4.2. Paper

4. Suchin Gururangan, Ana Marasović, Swabha Swayamdipta and colleagues (2020), Don't Stop Pretraining, ACL, §§3–4. Mistral (17 March 2026), Introducing Forge. Paper • Forge

5. UK GDPR, arts 5, 6 and 8A. See the changes made by s. 71 of the Data (Use and Access) Act 2025, in force from 5 February 2026. Legislation

6. Information Commissioner's Office, Monitoring workers, including its guidance on monitoring emails and messages. ICO guidance

7. Solicitors Regulation Authority, Code of Conduct for Firms, paras 6.3 and 6.5, and Misuse of AI warning notice (17 August 2026), "Client confidentiality". Code • Warning notice

8. Morgan Stanley (2024), Launch of AI @ Morgan Stanley Debrief, and OpenAI, Morgan Stanley uses AI evals to shape the future of financial services. Morgan Stanley • OpenAI

9. Erik Brynjolfsson, Danielle Li and Lindsey Raymond (2025), Generative AI at Work, Quarterly Journal of Economics, 140(2), pp. 889–942. Published paper