RAG vs fine-tuning: how to choose

26 October 2026 · 5 min read · blogs.iictp.org/rag-vs-fine-tuning

Both RAG and fine-tuning adapt a model to your needs, but they change different things. RAG changes what the model can see at answer time. Fine-tuning changes how the model behaves.

TL;DR The short version

- RAG changes what a model can see when it answers; fine-tuning changes how it behaves.
- Use RAG for facts that change or are private and when you must show sources; use fine-tuning for consistent style or a narrow repeated task.
- Try a clear prompt with examples first, add retrieval if facts are missing, and fine-tune only after that.

Read online, listen to it, or change the text size: https://blogs.iictp.org/rag-vs-fine-tuning

Use RAG when...

- The answer depends on facts that change or that are private.
- You need to show where an answer came from.
- Different users may see different documents.

Use fine-tuning when...

- You need a consistent style, format or tone.
- The task is narrow and repeated, and prompts alone are not reliable enough.
- You want a smaller model to do one job well.

Maintenance

RAG is updated by updating the index. Fine-tuning is updated by training again, so facts that change quickly are a poor fit. Knowledge baked into weights is also hard to audit or remove.

A practical order

1. Try a clear prompt with examples.
2. Add retrieval if the model lacks facts.
3. Consider fine-tuning only if behaviour is still wrong and you have good examples.
4. Measure each step on your own test set before moving on.

Related credential: AI Practitioner. Explore the exam facts and syllabus.

Sources and further reading

1. Lewis et al. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.” arXiv:2005.11401.

Published by IICTP, an independent technical credential body, on 26 October 2026. The current version of this post is at https://blogs.iictp.org/rag-vs-fine-tuning. More guides: blogs.iictp.org.