Confidential

Generation Digital

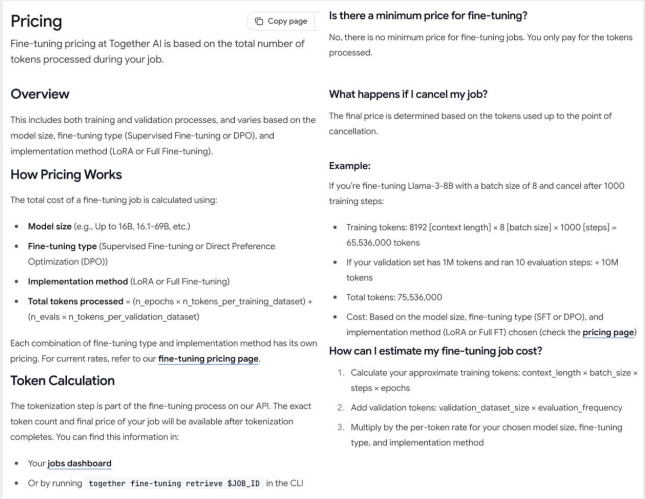
Generation Digital is an award winning consultancy that helps teams reimagine how they work through the power of enterprise AI tools.

As certified partners of multiple AI Products we help organisations unlock the full potential of AI-powered workplace tools, turning everyday systems into engines of efficiency and growth.

- Award-Winning Platinum Solutions Partner
- Global Client Base across Enterprise and Public Sector
- Experts in AI-assisted work, Collaboration, and Change Management

HM Government G-Cloud Supplier

hello@gend.co
+44 020 3951 3986
www.gend.co



Pricing

Copy page

Fine-tuning pricing at Together AI is based on the total number of tokens processed during your job.

Overview

This includes both training and validation processes, and varies based on the model size, fine-tuning type (Supervised Fine-tuning or DPO), and implementation method (LoRA or Full Fine-tuning).

How Pricing Works

The total cost of a fine-tuning job is calculated using:

- **Model size** (e.g., Up to 16B, 16.1-69B, etc.)
- **Fine-tuning type** (Supervised Fine-tuning or Direct Preference Optimization (DPO))
- **Implementation method** (LoRA or Full Fine-tuning)
- **Total tokens processed** = (n_epochs × n_tokens_per_training_dataset) + (n_evals × n_tokens_per_validation_dataset)

Each combination of fine-tuning type and implementation method has its own pricing. For current rates, refer to our **fine-tuning pricing page**.

Token Calculation

The tokenization step is part of the fine-tuning process on our API. The exact token count and final price of your job will be available after tokenization completes. You can find this information in:

- Your **jobs dashboard**
- Or by running `together fine-tuning retrieve $JOB_ID` in the CLI

Is there a minimum price for fine-tuning?

No, there is no minimum price for fine-tuning jobs. You only pay for the tokens processed.

What happens if I cancel my job?

The final price is determined based on the tokens used up to the point of cancellation.

Example:

If you're fine-tuning Llama-3-8B with a batch size of 8 and cancel after 1000 training steps:

- Training tokens: 8192 [context length] × 8 [batch size] × 1000 [steps] = 65,536,000 tokens
- If your validation set has 1M tokens and ran 10 evaluation steps: + 10M tokens
- Total tokens: 75,536,000
- Cost: Based on the model size, fine-tuning type (SFT or DPO), and implementation method (LoRA or Full FT) chosen (check the **pricing page**)

How can I estimate my fine-tuning job cost?

1. Calculate your approximate training tokens: context_length × batch_size × steps × epochs
2. Add validation tokens: validation_dataset_size × evaluation_frequency
3. Multiply by the per-token rate for your chosen model size, fine-tuning type, and implementation method

Together AI uses a flexible, usage-based pricing model rather than simple flat fees.

Customers generally pay as they go for the specific AI services they use — for example, charges for running models (inference) are based on the amount of text processed, and fine-tuning models is billed based on the computing work done.

This approach means organisations pay in line with the scale and type of AI work they need, rather than a one-size-fits-all price.

For larger or enterprise deployments, custom pricing is available in GBP. Please contact Generation Digital for more information.