Fireworks AI

Fireworks / Deepseek R1 Distill Llama 8B

accounts/fireworks/models/deepseek-r1-distill-llama-8b

LLM | Tunable | Chat

Details

Llama 8B distilled with reasoning from Deepseek R1

Fine-tuning

Deepseek R1 Distill Llama 8B can be fine-tuned on your data to create a model with better response quality. Fireworks uses low-rank adaptation (LoRA) to train a model that can be served efficiently at inference time.

See the Fine-tuning guide for details.

Fine-tune this model

On-demand deployments

On-demand deployments allow you to use **Deepseek R1 Distill Llama 8B** on dedicated GPUs with Fireworks' high-performance serving stack with high reliability and no rate limits.

See the On-demand deployments guide for details.

Deploy this Base model

Finetune this model

Deploy this model

Model Details

Created by	yingliu@fireworks.ai
Created	1/22/2025
Visibility	Public
Kind	Base model
Model size	8B parameters
Fine-tuning	Supported
Serverless LoRA Deployment	Supported
Function Calling	Not supported
Provider	Fireworks