Fine-Tuning & Serving Open LLMs

LoRA/QLoRA, distillation, small models and vLLM serving on-prem.

DURATION	FORMAT	LEVEL	LABS
40 hours	Onsite, online or hybrid	Advanced	Cloud sandbox included

About this programme

Customise and self-host open-weight language models when prompting and RAG are not enough — covering data preparation, parameter-efficient fine-tuning, evaluation, distillation into small models and high-throughput serving on your own infrastructure.

Who it is for: ML engineers and data scientists with Python and PyTorch experience

What your team will be able to do

- Decide when fine-tuning beats prompting or RAG
- Prepare instruction and preference datasets with quality controls
- Fine-tune open models with LoRA and QLoRA on a single GPU
- Evaluate tuned models against the base model and task baselines
- Serve models efficiently with vLLM and quantisation on-prem or in a private cloud

Before you start

- Python and PyTorch experience
- Understanding of how transformers are trained and evaluated

Course outline

We send the module-by-module outline with your proposal, adapted to your team's tools and goals.

Assessment and certification

How learning is checked

- A short knowledge check on day one and again at the end, so you can see each person's progress
- Graded hands-on labs in every module, reviewed by the trainer
- A final exercise that applies the programme to a problem like your team's own

Certificate and exams

- Participants who complete the programme receive a Cognitive Networks AI certificate of completion.

Formats and cohort size

- Onsite: At your office or a venue you choose, in full days.
- Online, live: Instructor-led over video, usually in half-day sessions so people can keep up with work.
- Hybrid: Some days in a room together, the rest online. Useful when a team works across cities.

Cohort size: 15 to 25 people per trainer. Larger teams run as parallel cohorts on the same outline, so everyone learns the same thing.

Pricing: We quote per cohort, based on duration, format and number of cohorts. There is no per-seat price list: tell us your team and we send a proposal within one business day.

Bring this programme to your team

Request a proposal at cognitivenetworks.ai/contact?programme=fine-tuning-open-llms

reachus@cognitivenetworks.ai · Bangalore +91 72043 05303 · Dallas +1 (303) 827-7415