GENERATIVE AI · TAUGHT WITH MISTRAL AI

Fine-Tuning and Post-Training

Dataset construction, LoRA, preference optimization, and reinforcement learning on verifiable rewards. The first lesson is when fine-tuning is the wrong answer, because usually it is.

Advanced | 5 phases | 30 sessions | 45 contact hours

Syllabus

5 phases, 30 sessions. Each phase ends in something you have built.

1 Decide, Then Baseline

Establish whether this problem needs new weights at all, and what number would prove it.

5 sessions
$87 alone

1.1 The Cheaper Answers First

A better prompt, few-shot examples, or retrieval beats a tune more often than anyone admits.

FREE PREVIEW

1.2 What Tuning Actually Buys

Format compliance, tone, latency from a smaller model, and private behavior no prompt encodes.

1.3 Writing the Task Down

Turn 'make it better' into a scored task with a holdout set a stranger could grade the same way.

1.4 Baseline Scorecard

Measure the base model on your task, on IFEval, and on MMLU before touching a single weight.

1.5 Hardware and Cost Math

Bytes per parameter, H100 near 2 to 3 dollars an hour, and what an 8B LoRA run really costs.

By the end: **A go or no-go memo with a scored base model, a frozen holdout set, and a budget in GPU hours.**

2 Data Is the Model

Build a training set you can defend example by example, and prove it is not leaking your eval.

6 sessions
$87 alone

2.1	**Where Examples Come From**	Production logs, paid writers, and teacher models, with the cost and the bias of each priced out.
2.2	**Small and Clean Wins**	A thousand curated examples usually beats a hundred thousand scraped ones, and why that holds.
2.3	**Chat Templates, Exactly**	Tokenizer chat templates and special tokens, plus the mismatch that trains a string you never wrote.
2.4	**Masking the Prompt**	Loss on completion tokens only, and how to print the label tensor to prove the mask is where you think.
2.5	**Synthetic Data and Licenses**	Generating traces from Ministral 3, plus the upstream terms that decide if you can ship the result.
2.6	**Dedup and Decontamination**	Near-duplicate removal and an n-gram check of train against eval, run before you trust any number.

By the end: **A versioned dataset with a data card, a decontamination report, and 50 examples reviewed by hand.**

3 Supervised Fine-Tuning

Get an adapter that beats the base model on your task without wrecking everything else it could do.

7 sessions
$87 alone

3.1 First SFT Run

TRL 1.8 SFTTrainer on Ministral 3 8B, sequence packing, and what the loss curve does not tell you.

3.2 LoRA on Every Layer

Adapters on all linear layers including MLP, not attention only, which is the 2024 advice that aged badly.

3.3 Rank and Learning Rate

Rank 16 to 32 unless capacity-bound, and a learning rate near ten times the full fine-tune value.

3.4 QLoRA and Its Tax

4-bit NF4 base weights, the quality cost you sometimes pay, and when 24GB is simply not enough.

3.5 Unsloth and Axolotl

Unsloth for the fastest single GPU run, Axolotl YAML when the recipe has to survive four people.

3.6 Full Fine-Tune, Multi-GPU

FSDP and ZeRO-3 sharding, and the point where renting eight H100s beats another LoRA sweep.

3.7 Catastrophic Forgetting

Replay mixing and a retention pack, because LoRA reduces forgetting but does not prevent it.

By the end: **A trained adapter with before and after numbers on both the task and a retention pack, published together.**

4 Preference and Verifiable Rewards

Push past what demonstrations can express, using DPO and GRPO where each one actually applies.

7 sessions
$87 alone

4.1 Why SFT Runs Out

Demonstrations cannot say 'this answer is worse', which is where preference and reward data start.

4.2 DPO, Honestly

Chosen and rejected pairs, beta, the reference model, and the length bias that inflates your win rate.

4.3 Rewards You Can Verify

Unit tests, exact match, and schema checks: RLVR only works where a program can grade the answer.

4.4 GRPO in Practice

Group sampling with no value model, KL to the reference, and rollouts that dominate the GPU bill.

4.5 Variants Worth Knowing

GSPO and DAPO as sequence-level and clipping fixes, and why PPO-style RLHF is now rarely the first choice.

4.6 Reward Hacking

The model finds the bug in your grader before you do, so read raw samples every run, not just curves.

4.7 Reasoning Distillation

Sample traces from a large teacher, keep only the verified ones, and train a student that still reasons.

By the end: **A GRPO-trained model with reward curves, saved rollouts, and a written account of one reward hack you caught.**

5 Quantize and Serve

Turn a training artifact into something that serves real traffic at a cost you can defend.

5 sessions
$87 alone

5.1 Merge or Keep Adapters
Multi-LoRA serving off one base model against a single merged checkpoint, and what each choice costs.

5.2 Quantize for Serving
FP8 first on Hopper and Blackwell, AWQ through llm-compressor when memory is the binding constraint.

5.3 Regression Before Rollout
Re-run the scorecard on the quantized artifact, because quantization undoes a tune very quietly.

5.4 The Judge That Lies
Position swaps, a judge from a different model family, and a human-labeled calibration set of your own.

5.5 Was It Worth It
Task lift against the base model, retention deltas, and dollars per successful request, side by side.

By the end: **A vLLM endpoint serving the tuned model with a published quality, latency, and cost regression report.**

Tools you will use

Mistral AI Studio | Ministral 3 8B (Apache 2.0) | TRL 1.8 | Hugging Face PEFT 0.19 | Unsloth | Axolotl | vLLM 0.26 | llm-compressor | lm-evaluation-harness

What you will build

1. **The Go or No-Go Memo**
   Proves you can tell a prompting problem from a weights problem before spending money
2. **Domain Adapter That Holds**
   A LoRA that wins on the task and still passes the retention pack afterward
3. **Verified Reasoning Tune**
   GRPO against a real grader, quantized, served, with the regression report attached

Registration

Phase 1: Decide, Then Baseline	$87
Phase 2: Data Is the Model	$87
Phase 3: Supervised Fine-Tuning	$87
Phase 4: Preference and Verifiable Rewards	$87
Phase 5: Quantize and Serve	$87
Whole track, all 5 phases	**$360**

Buying every phase separately costs $435, so the whole track saves $75.

No payment is taken online. Registering creates an order and payment instructions are sent with it. Access opens once payment is confirmed.