INTELLIGENCE · TAUGHT WITH HUGGING FACE

LLM / Fine-tuning

Prompting, retrieval, agents and fine-tuning. You measure a model before changing it, ground its answers in your own data, and put guardrails on it before anyone else uses it.

Intermediate | 5 phases | 23 sessions | 35 contact hours

Syllabus

5 phases, 23 sessions. Each phase ends in something you have built.

1 Driving the Model

Get output you can reproduce instead of a demo that worked once.

4 sessions
$72 alone

1.1 Tokens, Sampling & Context

What the model actually predicts, and what temperature, top-p, and a filling context window do to it.

FREE PREVIEW

1.2 Chat Templates

Apply the tokenizer's own template, then watch a mismatched one quietly wreck output quality.

1.3 Structured Output

JSON schemas, constrained decoding, and validation that fails loudly instead of silently.

1.4 Prompting Reasoning Models

Decomposition, few-shot, and why chain-of-thought prompts now hurt a model that already reasons.

By the end: **A structured-output task that returns schema-valid JSON on fifty consecutive runs.**

2 Evals Before Anything Else

Build the measurement you will judge every later change against.

4 sessions
$72 alone

2.1 Collecting Real Failures
Turn logged mistakes into a labeled eval set instead of inventing test cases at your desk.

2.2 Graders That Can't Argue
Exact match, schema checks, unit tests, and reshaping a task so a grader can verify it.

2.3 LLM as Judge, Honestly
Position, verbosity, and self-preference bias, and checking your judge against human labels.

2.4 Benchmarks & Contamination
Run LightEval, read a leaderboard, and work out whether a score was memorized.

By the end: **An eval harness of forty cases drawn from real failures, with a locked baseline score.**

3 Grounding & Context

Give the model your documents and prove retrieval earns what it costs.

4 sessions
$72 alone

3.1 Embeddings & Vector Search
Turn meaning into geometry, and measure recall@k before blaming the model.

3.2 Retrieval That Holds Up
Chunking, hybrid keyword plus dense search, reranking, and citations a reader can check.

3.3 Long Context vs Retrieval
Run both on your eval, then compare accuracy, latency, and cost per query before you choose.

3.4 Context Engineering
Token budgets, compaction, and why answers degrade long before the window is full.

By the end: **A cited assistant benchmarked against a long-context baseline on the same eval set.**

4 Agents, Tools & Rails

Let a model take real actions without handing an attacker your systems.

5 sessions
$72 alone

4.1	**Tool Calling**	Write tool schemas a model can actually use, and handle the call it gets wrong.
4.2	**MCP & the Protocol Layer**	Connect to MCP servers, publish one of your own, and see where agent-to-agent protocols sit.
4.3	**Code Agents & Sandboxes**	Let the model write Python for its whole plan, then run it where it can't hurt you.
4.4	**Tracing a Failed Run**	Read a full trace, classify the failure, and fix the step that actually broke.
4.5	**Injection & the Four Rails**	Guard input, output, retrieved text, and tool arguments, then red-team your own agent.

By the end: **A sandboxed agent wired to MCP servers, traced end to end, with input and tool-call rails.**

5 Making the Model Yours

Change a model's behavior and prove on your own eval that you improved it.

6 sessions
$72 alone

5.1	**Should You Fine-Tune?**	Rule out prompting and retrieval first, and spot the knowledge problem tuning will never fix.
5.2	**Building the Training Set**	Format the data, mask the prompt out of the loss, and audit what you're about to train on.
5.3	**Adapters Past Plain LoRA**	LoRA, rank-stabilized LoRA, and DoRA compared on your task under one memory budget.
5.4	**Quantized vs Full-Precision Tuning**	QLoRA on one consumer GPU, and the model families where 4-bit training now costs you accuracy.
5.5	**Preference & Reward Tuning**	DPO on preference pairs, then GRPO with a verifiable reward once a grader can score it.
5.6	**Serving It & Signing It**	Quantized inference on vLLM, a measured cost per request, and a model card stating the limits.

By the end: **A published adapter that beats the base model on your eval, served with a model card.**

Tools you will use

Hugging Face Transformers | TRL | PEFT | Unsloth | Sentence Transformers | vLLM | smolagents / MCP | LightEval

What you will build

1. **The Eval Harness**
   Forty real failures, graders, and a baseline to beat
2. **Grounded Agent**
   MCP tools, citations, sandbox, injection rails
3. **Your Own Adapter**
   One GPU, measured gain, published with a model card

Registration

Phase 1: Driving the Model	$72
Phase 2: Evals Before Anything Else	$72
Phase 3: Grounding & Context	$72
Phase 4: Agents, Tools & Rails	$72
Phase 5: Making the Model Yours	$72
Whole track, all 5 phases	**$300**

Buying every phase separately costs $360, so the whole track saves $60.

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