

Lab 102: Fine-Tuning vs RAG vs Prompting + LoRA on a Small Model

Level: L300 Journey: llm-foundations Time: ~120 min Cost: free-conditional

info "Phase 3 · local first" The required route is deterministic and local. The Real route is optional and includes an estimate and cleanup.

What You'll Learn

Select prompting, RAG, or fine-tuning from evidence Prepare a small LoRA dataset Estimate GPU and Foundry cost before training Compare quality, freshness, latency, and operations Prerequisites

Python 3.10+ and a terminal No credentials for the required route Read the linked official specification before using a hosted adapter

Visual model

[diagram or code sample]

Scenario

OutdoorGear needs a reviewable workshop implementation with deterministic fixtures, explicit trust boundaries, measurable acceptance criteria, and a reversible optional hosted adapter.

Guided exercise

Step 1: score_options

1. Run the deterministic fixture for this step. 2. Record the declared input, output, policy decision, and failure behavior. 3. Change one bounded behavior and run the local test again.

Evidence: Save the command, result, and one sentence explaining why the evidence is trustworthy.

Step 2: prepare_dataset

1. Run the deterministic fixture for this step. 2. Record the declared input, output, policy decision, and failure behavior. 3. Change one bounded behavior and run the local test again.

Evidence: Save the command, result, and one sentence explaining why the evidence is trustworthy.

Step 3: simulate_lora

1. Run the deterministic fixture for this step. 2. Record the declared input, output, policy decision, and failure behavior. 3. Change one bounded behavior and run the local test again.

Evidence: Save the command, result, and one sentence explaining why the evidence is trustworthy.

Step 4: compare_results

1. Run the deterministic fixture for this step. 2. Record the declared input, output, policy decision, and failure behavior. 3. Change one bounded behavior and run the local test again.

Evidence: Save the command, result, and one sentence explaining why the evidence is trustworthy.

Run locally

[diagram or code sample]

Simulation / Real path

Simulation Real path

Use the included deterministic fixtures and local validator. Estimated cost: US\$0.

Replace only the adapter boundary. Set a spending cap before execution, retain the same acceptance tests, and record region, retention, identity, telemetry, and pricing

assumptions. Delete hosted deployments, endpoints, indexes, traces, and test data after validation.

Supporting kit

The site build generates stable PDF and ZIP downloads. CI also keeps a temporary artifact mirror for 30 days.

README.md: quick start and evidence checklist CHEATSHEET.md: decision and review guide

scenario.json: deterministic exercise contract test_scenario.py: local acceptance tests

Download generated PDF Download generated ZIP

Scored knowledge check

question "Why keep a local contract?"

??? success "Answer" It provides a deterministic baseline before adding hosted variability and cost.

question "When is the Real route complete?"

??? success "Answer" Only after validation evidence and cleanup are recorded.

question "How should tool output be treated?"

??? success "Answer" As untrusted data until policy and schema validation succeed.

Exit criteria

[] Local tests pass without credentials. [] The diagram has a text explanation and does not depend on color. [] The scored quiz was reviewed. [] The Real path has a cost cap and cleanup plan. [] Sources and stability status were checked.

Official public sources

Hugging Face PEFT Microsoft Foundry fine-tuning

Supporting file: CHEATSHEET.md

Fine-Tuning vs RAG vs Prompting + LoRA on a Small Model ? Cheat Sheet

Start with the local contract and explicit acceptance criteria. Treat prompts, retrieved content, tool output, and protocol messages as untrusted. Keep identity, authorization, policy, telemetry, cost, and cleanup explicit. Prefer the smallest orchestration or adaptation that satisfies the evidence.

Supporting file: README.md

Fine-Tuning vs RAG vs Prompting + LoRA on a Small Model

python -m unittest discover -v

The required route is local and deterministic.

Supporting file: adaptation_dataset.jsonl

```
{"input": "OutdoorGear request", "expected": "grounded response", "label": "pass"}
```

Supporting file: decision_matrix.md

Prompting vs RAG vs fine-tuning

Need		Prompting		RAG		Fine-tuning		Fresh private facts		Limited		Best fit		Poor fit
	Stable behavior/style		Good		Good		Best fit		Lowest operational cost		Best fit			
Medium		Highest setup cost												

Supporting file: scenario.json

```
{ "lab": "102", "steps": [ "score_options", "prepare_dataset", "simulate_lora",  
  "compare_results" ], "required_route": "local", "credentials_required": false, "real_path":  
{ "optional": true, "cost_cap_required": true, "cleanup_required": true } }
```

Supporting file: test_scenario.py

```
import json import unittest from pathlib import Path
class Phase3ScenarioTests(unittest.TestCase): @classmethod def setUpClass(cls): cls.data =
json.loads(Path("scenario.json").read_text(encoding="utf-8"))
def test_identity_and_steps(self): self.assertEqual(self.data["lab"], "102")
self.assertEqual(len(self.data["steps"]), 4)
def test_free_first_contract(self): self.assertEqual(self.data["required_route"], "local")
self.assertFalse(self.data["credentials_required"])
def test_real_path_guardrails(self): real = self.data["real_path"]
self.assertTrue(real["optional"]) self.assertTrue(real["cost_cap_required"])
self.assertTrue(real["cleanup_required"])
if __name__ == "__main__": unittest.main()
```