SFT Training LLM for Improved Reasoning

ECE5554 Trustworthy Machine Learning | Fall 2025

Team Members: Mack Werner, Gary Ruppert, BJ Janson, Jacob Ramey*

*Equal Contributors

Introduction

The purpose of this assignment is to understand and execute the process of supervised fine-tuning a large language model (LLM) to improve the model's reasoning ability. The steps our group followed to accomplish this task include setting up our computer environments, evaluating the base model and fine-tuned model, creating the dataset we used to train our model, and fine-tuning the model. The team also considered data selection methods since strategic approaches can achieve better performance than the model. Some of these strategies include difficulty-based selection, diversity-based selection, and random selection. In later sections, we will discuss why we chose the data selection strategy that we did and what results we achieved using it.

We will train, fine-tune, and evaluate the Qwen 2.5-3B-Instruct model from HuggingFace.

Baseline Model Evaluation

This section presents the baseline model evaluation results of the Qwen2.5-3B-Instruct model. The model configuration is provided in the GitHub repository. Parameters for the math reasoning models were set to the following:

- *MODEL_NAME = "QWEN/QWEN2.5-3B-INSTRUCT",*
- *DTYPE = BFLOAT16,*
- *TENSOR_PARALLEL_SIZE = $NUM_GPUS,*
- *MAX_MODEL_LENGTH = 32768,*
- *GPU_MEMORY_UTILIZATION = 0.95,*
- *GENERATION_PARAMETERS = {MAX_NEW_TOKENS:32768, TEMPERATURE:0.6, TOP_P:0.95}"*

The evaluation models were aime24, aime25, math_500, gpqa:diamond, and codegeneration. For the general benchmark evaluation, the parameters were the same. The model tested was the mmlu_redux_2.

Baseline Evaluation Results: Mathematical Reasoning

For the Qwen 2.5-3B-Instruct model, we obtained the following results across multiple user platforms and using different developer software:

Group Member	Task	Metric	Value	Stderr
BJ Janson	all	gpqa_pass@k_with_k	0.2929	0.0324
		pass@k_with_k&n	0.317	0.0108
		pass@k_with_k	0.1	0.0557
		codegen_pass@1:16	0.0858	0.0171
	extended:lcb:codegeneration:0	codegen_pass@1:16	0.0858	0.0171
	lighteval:aime24:0	pass@k_with_k	0.1	0.0557
	lighteval:aime25:0	pass@k_with_k&n	0.0	0.0
	lighteval:gpqa:diamond:0	gpqa_pass@k_with_k	0.2929	0.0324
	lighteval:math_500:0	pass@k_with_k&n	0.634	0.0216
Gary Ruppert	all	pass@k_with_k	0.0667	0.0463
		codegen_pass@1:16	0.0933	0.0178
		gpqa_pass@k_with_k	0.2879	0.0323
		pass@k_with_k&n	0.3467	0.0273
	extended:lcb:codegeneration:0	codegen_pass@1:16	0.0933	0.0178
	lighteval:aime24:0	pass@k_with_k	0.0667	0.0463
	lighteval:aime25:0	pass@k_with_k&n	0.0333	0.0333
	lighteval:gpqa:diamond:0	gpqa_pass@k_with_k	0.2879	0.0323
	lighteval:math_500:0	pass@k_with_k&n	0.66	0.0212
Jacob Ramey	all	pass@k_with_k	0.067	0.046
		codegen_pass@1:16	0.104	0.019
		gpqa_pass@k_with_k	0.354	0.034
		pass@k_with_k&n	0.351	0.027
	extended:lcb:codegeneration:0	codegen_pass@1:16	0.104	0.019
	lighteval:aime24:0	pass@k_with_k	0.067	0.046
	lighteval:aime25:0	pass@k_with_k&n	0.033	0.033
	lighteval:gpqa:diamond:0	gpqa_pass@k_with_k	0.354	0.034
	lighteval:math_500:0	pass@k_with_k&n	0.668	0.021
Mack Werner	all	pass@k_with_k&n	0.348	0.027
		codegen_pass@1:16	0.090	0.017

pass@k_with_k	0.100	0.056
gpqa_pass@k_with_k	0.343	0.034

Baseline Evaluation Results: General Benchmarks

Results from one user's general benchmark test are shown below; for brevity, only one evaluation result is provided. Other users obtained similar results.

```
[2025-10-06 22:27:09,784] [    INFO]: --- DISPLAYING RESULTS --- (pipeline.py:432)
```

Task	Version	Metric	Value		Stderr
all		pass@k_with_k	0.6667	±	0.0456
		acc	0.6391	±	0.0456
lighteval:mmlu_redux_2:_average:0		pass@k_with_k	0.6667	±	0.0456
		acc	0.6391	±	0.0456
lighteval:mmlu_redux_2:abstract_algebra:0		pass@k_with_k	0.4900	±	0.0502
		acc	0.3700	±	0.0485
lighteval:mmlu_redux_2:anatomy:0		pass@k_with_k	0.6200	±	0.0488
		acc	0.6000	±	0.0492
lighteval:mmlu_redux_2:astronomy:0		pass@k_with_k	0.7000	±	0.0461
		acc	0.6500	±	0.0479
lighteval:mmlu_redux_2:business_ethics:0		pass@k_with_k	0.6800	±	0.0469
		acc	0.7300	±	0.0446
lighteval:mmlu_redux_2:clinical_knowledge:0		pass@k_with_k	0.6700	±	0.0473
		acc	0.6500	±	0.0479
lighteval:mmlu_redux_2:college_biology:0		pass@k_with_k	0.6900	±	0.0465
		acc	0.6800	±	0.0469
lighteval:mmlu_redux_2:college_chemistry:0		pass@k_with_k	0.4600	±	0.0501
		acc	0.4400	±	0.0499
lighteval:mmlu_redux_2:college_computer_science:0		pass@k_with_k	0.6000	±	0.0492
		acc	0.4700	±	0.0502
lighteval:mmlu_redux_2:college_mathematics:0		pass@k_with_k	0.4600	±	0.0501
		acc	0.3200	±	0.0469
lighteval:mmlu_redux_2:college_medicine:0		pass@k_with_k	0.7200	±	0.0451
		acc	0.7000	±	0.0461
lighteval:mmlu_redux_2:college_physics:0		pass@k_with_k	0.6300	±	0.0485
		acc	0.4700	±	0.0502
lighteval:mmlu_redux_2:computer_security:0		pass@k_with_k	0.6900	±	0.0465
		acc	0.7000	±	0.0461
lighteval:mmlu_redux_2:conceptual_physics:0		pass@k_with_k	0.6300	±	0.0485
		acc	0.5500	±	0.0500
lighteval:mmlu_redux_2:econometrics:0		pass@k_with_k	0.5000	±	0.0503
		acc	0.5200	±	0.0502
lighteval:mmlu_redux_2:electrical_engineering:0		pass@k_with_k	0.5800	±	0.0496
		acc	0.5900	±	0.0494
lighteval:mmlu_redux_2:elementary_mathematics:0		pass@k_with_k	0.8500	±	0.0359
		acc	0.4800	±	0.0502
lighteval:mmlu_redux_2:formal_logic:0		pass@k_with_k	0.4900	±	0.0502
		acc	0.4700	±	0.0502
lighteval:mmlu_redux_2:global_facts:0		pass@k_with_k	0.4800	±	0.0502
		acc	0.4200	±	0.0496
lighteval:mmlu_redux_2:high_school_biology:0		pass@k_with_k	0.7600	±	0.0429
		acc	0.7600	±	0.0429
lighteval:mmlu_redux_2:high_school_chemistry:0		pass@k_with_k	0.7500	±	0.0435
		acc	0.5800	±	0.0496
lighteval:mmlu_redux_2:high_school_computer_science:0		pass@k_with_k	0.7400	±	0.0441
		acc	0.6900	±	0.0465

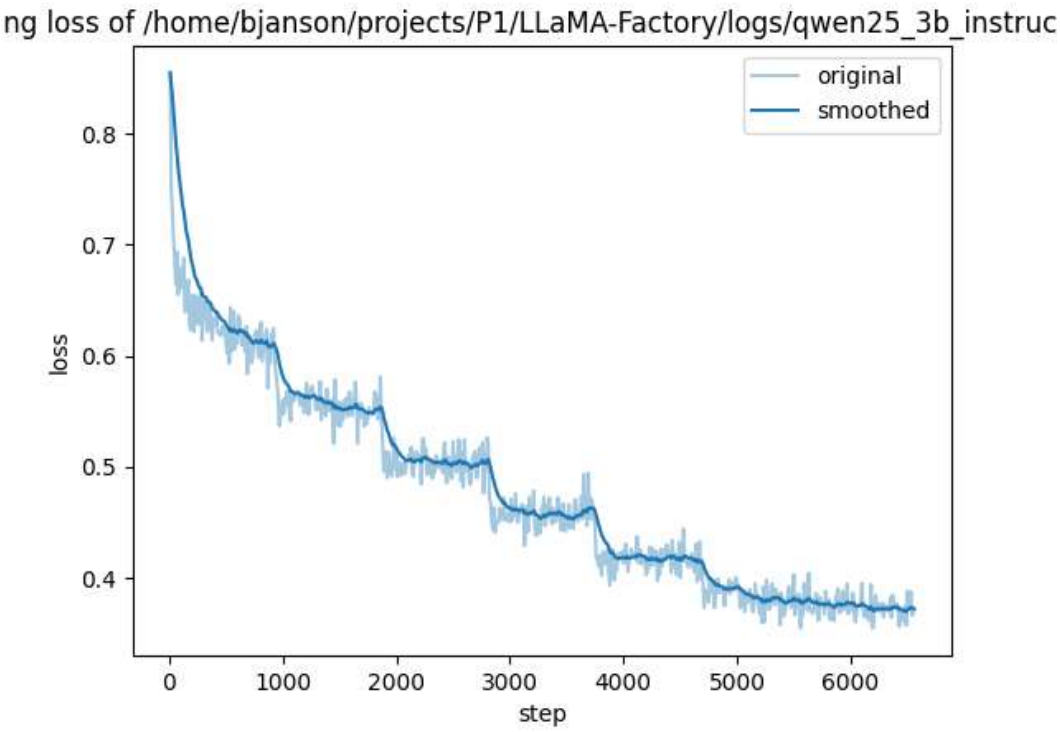
lighteval:mmlu_redux_2:high_school_european_history:0		pass@k_with_k	0.7600	±	0.0429
		acc	0.7600	±	0.0429
lighteval:mmlu_redux_2:high_school_geography:0		pass@k_with_k	0.7000	±	0.0461
		acc	0.7100	±	0.0456
lighteval:mmlu_redux_2:high_school_government_and_politics:0		pass@k_with_k	0.8100	±	0.0394
		acc	0.7900	±	0.0409
lighteval:mmlu_redux_2:high_school_macroeconomics:0		pass@k_with_k	0.6400	±	0.0482
		acc	0.5800	±	0.0496
lighteval:mmlu_redux_2:high_school_mathematics:0		pass@k_with_k	0.5500	±	0.0500
		acc	0.3500	±	0.0479
lighteval:mmlu_redux_2:high_school_microeconomics:0		pass@k_with_k	0.7500	±	0.0435
		acc	0.8200	±	0.0386
lighteval:mmlu_redux_2:high_school_physics:0		pass@k_with_k	0.5900	±	0.0494
		acc	0.4400	±	0.0499
lighteval:mmlu_redux_2:high_school_psychology:0		pass@k_with_k	0.7700	±	0.0423
		acc	0.8000	±	0.0402
lighteval:mmlu_redux_2:high_school_statistics:0		pass@k_with_k	0.7800	±	0.0416
		acc	0.5700	±	0.0498
lighteval:mmlu_redux_2:high_school_us_history:0		pass@k_with_k	0.7200	±	0.0451
		acc	0.7700	±	0.0423
lighteval:mmlu_redux_2:high_school_world_history:0		pass@k_with_k	0.7400	±	0.0441
		acc	0.8200	±	0.0386
lighteval:mmlu_redux_2:human_aging:0		pass@k_with_k	0.6700	±	0.0473
		acc	0.6600	±	0.0476
lighteval:mmlu_redux_2:human_sexuality:0		pass@k_with_k	0.7300	±	0.0446
		acc	0.7400	±	0.0441
lighteval:mmlu_redux_2:international_law:0		pass@k_with_k	0.7200	±	0.0451
		acc	0.7500	±	0.0435
lighteval:mmlu_redux_2:jurisprudence:0		pass@k_with_k	0.7500	±	0.0435
		acc	0.7700	±	0.0423
lighteval:mmlu_redux_2:logical_fallacies:0		pass@k_with_k	0.8600	±	0.0349
		acc	0.8500	±	0.0359
lighteval:mmlu_redux_2:machine_learning:0		pass@k_with_k	0.4600	±	0.0501
		acc	0.4300	±	0.0498
lighteval:mmlu_redux_2:management:0		pass@k_with_k	0.7300	±	0.0446
		acc	0.7700	±	0.0423
lighteval:mmlu_redux_2:marketing:0		pass@k_with_k	0.8900	±	0.0314
		acc	0.8900	±	0.0314
lighteval:mmlu_redux_2:medical_genetics:0		pass@k_with_k	0.7300	±	0.0446
		acc	0.7100	±	0.0456
lighteval:mmlu_redux_2:miscellaneous:0		pass@k_with_k	0.8100	±	0.0394
		acc	0.8400	±	0.0368
lighteval:mmlu_redux_2:moral_disputes:0		pass@k_with_k	0.6100	±	0.0490
		acc	0.6200	±	0.0488
lighteval:mmlu_redux_2:moral_scenarios:0		pass@k_with_k	0.3600	±	0.0482
		acc	0.3500	±	0.0479
lighteval:mmlu_redux_2:nutrition:0		pass@k_with_k	0.6800	±	0.0469
		acc	0.7500	±	0.0435

lighteval:mmlu_redux_2:prehistory:0		pass@k_with_k	0.7400	±	0.0441
		acc	0.7500	±	0.0435
lighteval:mmlu_redux_2:professional_accounting:0		pass@k_with_k	0.5100	±	0.0502
		acc	0.4900	±	0.0502
lighteval:mmlu_redux_2:professional_law:0		pass@k_with_k	0.4300	±	0.0498
		acc	0.4300	±	0.0498
lighteval:mmlu_redux_2:professional_medicine:0		pass@k_with_k	0.6900	±	0.0465
		acc	0.6500	±	0.0479
lighteval:mmlu_redux_2:professional_psychology:0		pass@k_with_k	0.7000	±	0.0461
		acc	0.7100	±	0.0456
lighteval:mmlu_redux_2:public_relations:0		pass@k_with_k	0.7300	±	0.0446
		acc	0.7100	±	0.0456
lighteval:mmlu_redux_2:security_studies:0		pass@k_with_k	0.7200	±	0.0451
		acc	0.6900	±	0.0465
lighteval:mmlu_redux_2:sociology:0		pass@k_with_k	0.8200	±	0.0386
		acc	0.8600	±	0.0349
lighteval:mmlu_redux_2:us_foreign_policy:0		pass@k_with_k	0.7900	±	0.0409
		acc	0.8100	±	0.0394
lighteval:mmlu_redux_2:virology:0		pass@k_with_k	0.4400	±	0.0499
		acc	0.4800	±	0.0502
lighteval:mmlu_redux_2:world_religions:0		pass@k_with_k	0.7900	±	0.0409
		acc	0.8200	±	0.0386

Fine-tuned Model Evaluation

This section presents training and evaluation summaries for the fine-tuning of the Qwen2.5-3B-Instruct model. The model was trained on a random subset of 15,000 points taken from the AceReason-1.1-SFT dataset. The model hyperparameters are located at the HuggingFace repository listed at the end of this report. The training configuration Json file is provided in the GitHub repository listed at the end of the report. Seven epochs were run with a cutoff length of 16,384 tokens, and the model was run through an SFT model with full fine-tuning.

A graph of the training loss per step is shown below. The loss dropped to about 0.36 by the end of the run.



The training run generated a Json file summarizing the results:

- "EPOCH": 7.0
- TOTAL_FLOS": 1.30951717556458 62E+19,
- "TRAIN_LOSS": 0.08976114059529591,
- "TRAIN_RUNTIME": 23786.5239,
- "TRAIN_SAMPLES_PER_SECOND": 4.414,
- "TRAIN_STEPS_PER_SECOND": 0.276

Fine-tuned Model Evaluation Results: Mathematical Reasoning

The results of the mathematical reasoning test conducted on one user system are shown below.

Task	Version	Metric	Value		Stderr
all		pass@k_with_k	0.1000	±	0.0557
		pass@k_with_k&n	0.3990	±	0.0381
		gpqa_pass@k_with_k	0.3131	±	0.0330
		codegen_pass@1:16	0.0970	±	0.0181
extended:lcb:codegeneration:0		codegen_pass@1:16	0.0970	±	0.0181
lighteval:aime24:0		pass@k_with_k	0.1000	±	0.0557
lighteval:aime25:0		pass@k_with_k&n	0.1000	±	0.0557
lighteval:gpqa:diamond:0		gpqa_pass@k_with_k	0.3131	±	0.0330
lighteval:math_500:0		pass@k_with_k&n	0.6980	±	0.0206

Fine-tuned Model Evaluation Results: General Benchmarks

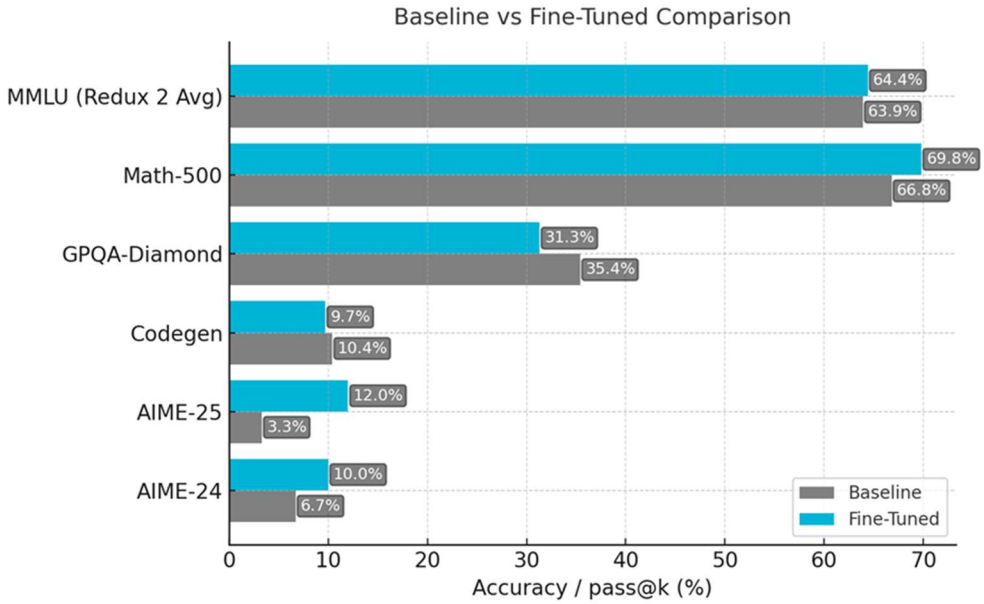
The results from rerunning the general benchmark training on the fine-tuned model are shown below for one user.

Task	Version	Metric	Value		Stderr
all		pass@k_with_k	0.6430	±	0.0457
		acc	0.4949	±	0.0482
lighteval:mmlu_redux_2:_average:0		pass@k_with_k	0.6430	±	0.0457
		acc	0.4949	±	0.0482
lighteval:mmlu_redux_2:abstract_algebra:0		pass@k_with_k	0.3200	±	0.0469
		acc	0.2400	±	0.0429
lighteval:mmlu_redux_2:anatomy:0		pass@k_with_k	0.5900	±	0.0494
		acc	0.5400	±	0.0501
lighteval:mmlu_redux_2:astronomy:0		pass@k_with_k	0.7300	±	0.0446
		acc	0.5100	±	0.0502
lighteval:mmlu_redux_2:business_ethics:0		pass@k_with_k	0.6400	±	0.0482
		acc	0.7200	±	0.0451
lighteval:mmlu_redux_2:clinical_knowledge:0		pass@k_with_k	0.7100	±	0.0456
		acc	0.5800	±	0.0496
lighteval:mmlu_redux_2:college_biology:0		pass@k_with_k	0.7700	±	0.0423
		acc	0.5900	±	0.0494
lighteval:mmlu_redux_2:college_chemistry:0		pass@k_with_k	0.4200	±	0.0496
		acc	0.3300	±	0.0473
lighteval:mmlu_redux_2:college_computer_science:0		pass@k_with_k	0.5300	±	0.0502
		acc	0.3800	±	0.0488
lighteval:mmlu_redux_2:college_mathematics:0		pass@k_with_k	0.3300	±	0.0473
		acc	0.2300	±	0.0423
lighteval:mmlu_redux_2:college_medicine:0		pass@k_with_k	0.7000	±	0.0461
		acc	0.5200	±	0.0502
lighteval:mmlu_redux_2:college_physics:0		pass@k_with_k	0.5400	±	0.0501
		acc	0.3200	±	0.0469
lighteval:mmlu_redux_2:computer_security:0		pass@k_with_k	0.6800	±	0.0469
		acc	0.5500	±	0.0500
lighteval:mmlu_redux_2:conceptual_physics:0		pass@k_with_k	0.6600	±	0.0476
		acc	0.4700	±	0.0502
lighteval:mmlu_redux_2:econometrics:0		pass@k_with_k	0.4300	±	0.0498
		acc	0.3100	±	0.0465
lighteval:mmlu_redux_2:electrical_engineering:0		pass@k_with_k	0.5600	±	0.0499
		acc	0.5100	±	0.0502
lighteval:mmlu_redux_2:elementary_mathematics:0		pass@k_with_k	0.6300	±	0.0485
		acc	0.3200	±	0.0469
lighteval:mmlu_redux_2:formal_logic:0		pass@k_with_k	0.5600	±	0.0499
		acc	0.3900	±	0.0490
lighteval:mmlu_redux_2:global_facts:0		pass@k_with_k	0.4300	±	0.0498
		acc	0.2600	±	0.0441
lighteval:mmlu_redux_2:high_school_biology:0		pass@k_with_k	0.7500	±	0.0435
		acc	0.5700	±	0.0498
lighteval:mmlu_redux_2:high_school_chemistry:0		pass@k_with_k	0.5800	±	0.0496
		acc	0.4100	±	0.0494
lighteval:mmlu_redux_2:high_school_computer_science:0		pass@k_with_k	0.7800	±	0.0416
		acc	0.4300	±	0.0498
lighteval:mmlu_redux_2:high_school_european_history:0		pass@k_with_k	0.8000	±	0.0402
		acc	0.5300	±	0.0502
lighteval:mmlu_redux_2:high_school_geography:0		pass@k_with_k	0.7200	±	0.0451
		acc	0.5800	±	0.0496
lighteval:mmlu_redux_2:high_school_government_and_politics:0		pass@k_with_k	0.8300	±	0.0378
		acc	0.6500	±	0.0479
lighteval:mmlu_redux_2:high_school_macroeconomics:0		pass@k_with_k	0.6600	±	0.0476
		acc	0.4500	±	0.0500
lighteval:mmlu_redux_2:high_school_mathematics:0		pass@k_with_k	0.3100	±	0.0465
		acc	0.2700	±	0.0446
lighteval:mmlu_redux_2:high_school_microeconomics:0		pass@k_with_k	0.7900	±	0.0409
		acc	0.6100	±	0.0490

lighteval:mmlu_redux_2:high_school_physics:0		pass@k_with_k	0.5200	±	0.0502
		acc	0.2600	±	0.0441
lighteval:mmlu_redux_2:high_school_psychology:0		pass@k_with_k	0.8200	±	0.0386
		acc	0.7400	±	0.0441
lighteval:mmlu_redux_2:high_school_statistics:0		pass@k_with_k	0.6500	±	0.0479
		acc	0.4000	±	0.0492
lighteval:mmlu_redux_2:high_school_us_history:0		pass@k_with_k	0.8400	±	0.0368
		acc	0.4700	±	0.0502
lighteval:mmlu_redux_2:high_school_world_history:0		pass@k_with_k	0.7700	±	0.0423
		acc	0.4800	±	0.0502
lighteval:mmlu_redux_2:human_aging:0		pass@k_with_k	0.6500	±	0.0479
		acc	0.4400	±	0.0499
lighteval:mmlu_redux_2:human_sexuality:0		pass@k_with_k	0.7200	±	0.0451
		acc	0.6300	±	0.0485
lighteval:mmlu_redux_2:international_law:0		pass@k_with_k	0.7000	±	0.0461
		acc	0.6400	±	0.0482
lighteval:mmlu_redux_2:jurisprudence:0		pass@k_with_k	0.7500	±	0.0435
		acc	0.5600	±	0.0499
lighteval:mmlu_redux_2:logical_fallacies:0		pass@k_with_k	0.7500	±	0.0435
		acc	0.6600	±	0.0476
lighteval:mmlu_redux_2:machine_learning:0		pass@k_with_k	0.5400	±	0.0501
		acc	0.3700	±	0.0485
lighteval:mmlu_redux_2:management:0		pass@k_with_k	0.7500	±	0.0435
		acc	0.6100	±	0.0490
lighteval:mmlu_redux_2:marketing:0		pass@k_with_k	0.8600	±	0.0349
		acc	0.7600	±	0.0429
lighteval:mmlu_redux_2:medical_genetics:0		pass@k_with_k	0.7500	±	0.0435
		acc	0.5700	±	0.0498
lighteval:mmlu_redux_2:miscellaneous:0		pass@k_with_k	0.8100	±	0.0394
		acc	0.7300	±	0.0446
lighteval:mmlu_redux_2:moral_disputes:0		pass@k_with_k	0.6400	±	0.0482
		acc	0.3800	±	0.0488
lighteval:mmlu_redux_2:moral_scenarios:0		pass@k_with_k	0.3600	±	0.0482
		acc	0.2900	±	0.0456
lighteval:mmlu_redux_2:nutrition:0		pass@k_with_k	0.7000	±	0.0461
		acc	0.6000	±	0.0492
lighteval:mmlu_redux_2:philosophy:0		pass@k_with_k	0.6500	±	0.0479
		acc	0.4600	±	0.0501
lighteval:mmlu_redux_2:prehistory:0		pass@k_with_k	0.7200	±	0.0451
		acc	0.5600	±	0.0499
lighteval:mmlu_redux_2:professional_accounting:0		pass@k_with_k	0.5200	±	0.0502
		acc	0.4500	±	0.0500
lighteval:mmlu_redux_2:professional_law:0		pass@k_with_k	0.3900	±	0.0490
		acc	0.3400	±	0.0476
lighteval:mmlu_redux_2:professional_medicine:0		pass@k_with_k	0.4700	±	0.0502
		acc	0.5000	±	0.0503
lighteval:mmlu_redux_2:professional_psychology:0		pass@k_with_k	0.6800	±	0.0469
		acc	0.5300	±	0.0502
lighteval:mmlu_redux_2:public_relations:0		pass@k_with_k	0.6400	±	0.0482
		acc	0.5900	±	0.0494
lighteval:mmlu_redux_2:security_studies:0		pass@k_with_k	0.6700	±	0.0473
		acc	0.6300	±	0.0485
lighteval:mmlu_redux_2:sociology:0		pass@k_with_k	0.8300	±	0.0378
		acc	0.6800	±	0.0469
lighteval:mmlu_redux_2:us_foreign_policy:0		pass@k_with_k	0.7400	±	0.0441
		acc	0.5700	±	0.0498
lighteval:mmlu_redux_2:virology:0		pass@k_with_k	0.4900	±	0.0502
		acc	0.3400	±	0.0476
lighteval:mmlu_redux_2:world_religions:0		pass@k_with_k	0.8200	±	0.0386
		acc	0.7000	±	0.0461

Comparison of Results

The figure below compares the baseline performance of the Qwen2.5-3B-Instruct model before and after supervised fine-tuning (SFT). The fine-tuned model demonstrates improvements across most benchmark categories, with the most significant gain observed in AIME and Math-500 mathematical evaluations.



Task	Baseline	Fine-Tuned	Δ (Change)
AIME-24	6.7%	10.0%	+3.3 pp
AIME-25	3.3%	12.0%	+8.7 pp
Codegen	10.4%	9.7%	-0.7 pp
GPQA-Diamond*	35.4%	31.3%	-4.1 pp
Math-500	66.8%	69.8%	+3.0 pp
MMLU (Redux 2 Avg)	63.9%	64.4%	+0.5 pp

** Results from Jacob Ramey's baseline versus BJ Janson's fine-tuned models*

We make the following observations based on the evaluation and results:

- Fine-tuning improved mathematical reasoning tasks (Math-500, AIME-24/25) by 3–9 percentage points.
- General factual knowledge (MMLU) improved slightly by +0.5 pp.
- GPQA-Diamond dropped slightly (-4 pp), indicating domain-specific variation.
- Code generation performance remained stable.

Overall, fine-tuning with reasoning-focused data enhanced the model's quantitative reasoning capability without degrading general knowledge performance.

Advanced Data Selection Strategy

For this section of the project, we decided to try out the LIMOPro (Large-scale Instruction-following Model based on Prompt-response Optimization) approach. The idea behind this method is that you can carefully curate a large dataset to produce a small number of training samples that have a notable impact on the training.

The task leveraged the GitHub repository of files provided by the developers and authors of the paper. The software approach converted our JSON dataset to a format suitable for leveraging their code. Their programs would process the data and enhance it with several metadata fields. Additionally, we would convert our JSON to individual files and run them through causal information estimation (CIE). CIE processing measures changes in perplexity for sentence groups when reasoning steps are removed. Those metrics allow for a weighted pruning of the data to a smaller and more impactful dataset.

After working through a period of development, we ran our code through the sequence of steps and further trimmed the data, keeping only those that had answers filled in. The thought was that this would improve the training. While the result of Qwen2.5 training with this dataset did have a more significant final training loss of 0.17, the evaluation metrics were poor. This result turned out to be an error in our steps. The training loss was likely due to using a smaller dataset that was unchallenging for the model.

We reassessed our approach and found that numerous metadata fields were not populated as expected. Despite continued work, project schedule constraints prevented us from breaking through the checks and flags within the code that would have allowed data to populate. An additional challenge, after we made progress, was that the CIE calculation expects to process each sentence group of the chat output, rather than the whole text we processed the first time. This change would add significant processing time, and resources became scarce toward the end of the project's due date. We were also unable to resolve a problem adding metadata that would accurately allow pruning/trimming of the text.

If we have been successful in previous steps, we would retain the JSON entries in the 50-90% CIE score percentiles. As suggested by the paper, those metrics would have

allowed us to select 1,000 to 2,000 entries that offered the right complexity to challenge the training model and improve benchmarks above baseline.

Conclusion

This report presented evaluation and training results for the Qwen2.5-3B-Instruct model after training with an AceReason data subset. Our results show that baseline model performs better than average overall, per the lighteval tests, while the fine-tuned model performs just slightly better. The team was unable to resolve developmental issues to capture a better data subset using the LIMO method. Test artifacts are provided in the data repositories listed at the end of this document.

Links to Repositories

Github Repository: rameyjm7/SFT-Training-LLM: supervised fine-tuning (SFT) a large language model, Qwen2.5-3B-Instruct, to improve reasoning

Baseline Training Finetuned Model:
https://huggingface.co/BJJ5555/ECE6514_models/tree/main

References

DeepWiki. (2025, October 9). *LLaMa-Factory*. Retrieved from DeepWiki: https://deepwiki.com/hiyouga/LLaMA-Factory

Geeks for Geeks. (2025, July 23). *Supervised Fine-Tuning (SFT) for LLMs*. Retrieved from Geeks for Geeks: https://www.geeksforgeeks.org/artificial-intelligence/supervised-fine-tuning-sft-for-llms/

Huggingface. (2025, October 10). *SFT Trainer*. Retrieved from Huggingface: https://huggingface.co/docs/trl/main/en/sft_trainer

Just, H. A. (2025, October 9). *Project-Reasoning-SFT-LLM*. Retrieved from Github: https://github.com/reds-lab/Project-Reasoning-SFT-LLM

Werner, M., Ruppert, G., Janson, B., & Ramey, J. (2025, October 10). *SFT-Training-LLM*. Retrieved from Github: https://github.com/rameyjm7/SFT-Training-LLM

Zheng, Y. (2025, October 9). *LLaMA-Factory*. Retrieved from GitHub: https://github.com/hiyouga/LLaMA-Factory