

Qwen3.5-9B LoRA fine-tuning on the full WikiSQL train set: pipeline parallel over four MacBook Pro M5 Max with NukeTorch

Klay Moon, NukeTorch · Report v1.0 · 4 Apple M5 Max, Thunderbolt bridge

Headline

- **5,344.7 trained tokens per second**, pipeline parallel over four Apple M5 Max (NukeTorch's Choo-Choo Relay Train): one pass over the official WikiSQL train set, 56,355 examples in 3,522 steps of 16, 5,126,415 trained tokens (Table 1).
- **Execution accuracy 84.65%** on the official test set (15,878 rows, greedy), scored with WikiSQL's official scorer (Table 2).
- **Against the base model:** on test rows 0–3,999, execution accuracy 82.13% after training against 47.03% for the base model, 1-shot (Table 2).

Table 1. Speed: one pass over the official train set

Measure	Value
Official train set (examples)	56,355
Steps of 16 examples	3,522
Trained tokens	5,126,415
Steps time (s)	959.2
Trained tokens/s, pipeline parallel over four Macs	5,344.7

Trained tokens per second = trained tokens (padding excluded) ÷ the sum of step times. 56,352 of the 56,355 examples are trained: like mlx-lm's own batching, the pass forms full batches of 16 and leaves out the last 3.

Table 2. Accuracy on the official test set (greedy)

Measure	Exact match	Logical form	Execution
Trained, full test set (15,878 rows)	73.05%	77.28%	84.65%
Trained, test rows 0–3,999	69.50%	73.35%	82.13%
Base model, 1-shot, test rows 0–3,999	32.23%	42.98%	47.03%

Scored with WikiSQL's official query and database-engine scorer; held-out loss after training: validation 0.830 (8,421 rows), test 0.837 (15,878 rows). The base model is compared on the first 4,000 test rows only, a slice fixed before any result. One run; higher accuracy was not pursued because this report is a speed benchmark. Energy is not compared for this run: batteries charged during it.

How to reproduce

1. **Machines:** four MacBook Pro 16-inch, Apple M5 Max (18-core CPU, 40-core GPU, 48 GB), joined by a Thunderbolt bridge, on mains power.
2. **Cooling:** No external fans; each Mac's own fans under automatic control (declared by NukeTorch). Our other published benchmarks used external fans; this one did not.
3. **Model:** Qwen/Qwen3.5-9B at revision c20223623576.
4. **Data:** the official WikiSQL splits (salesforce/WikiSQL at cffb4230), every row converted to mlx-community/wikisql's text format with ml-explore/mlx-examples lora/data/wikisql at 796f5b53.
5. **Recipe:** Apple's WWDC26 LoRA settings (global batch 16, 4 per Mac; LoRA rank 8, scale 20, layers 16–31; Adam at $1e-5$), one pass in the maker's batch order, at most 512 tokens per example.
6. **Engine:** NukeTorch pipeline parallel over the four Macs, build d54eebe5 from a clean tree (available to licensees).
7. **Before the run:** every Mac at rest for at least 5 minutes with no thermal warning, then a process check on every Mac.
8. **Measure:** trained tokens $\div$ the sum of step times (Table 1); accuracy with WikiSQL's official scorer (Table 2).

Prepared by NukeTorch Inc. Report v1.0. Author-run measurements; no independent reproduction yet.