lyzr

Fine-Tuning Cheat Sheet

Types of Model Training and Fine-Tuning

Pre-training

Learns language and knowledge. Labs do it once.

Fine-tuning

Learns the customer's domain. Trained on QA pairs.

Post-training

Shapes tone and judgment. Uses preferences and rewards.

Pre-training

The base model is built once by model makers (Qwen, Llama, Gemma, Mistral) on trillions of tokens of general text. Customers never do this; we always start from an existing open-weight model.

Cost to pre-train Qwen3.5 27B

$10M to $16M

Rented compute for the final run alone

$20M to $50M

Roughly, all in

Our estimate. Qwen does not publish its cost.

Data	Qwen3 was trained on 36 trillion tokens and Qwen3.5 on more, so 36 trillion is the floor.
Compute	6 × 27 billion parameters × 36 trillion tokens is about 5.8×10^{24} operations, about 4 million H100 GPU hours at realistic utilization, or roughly six months on 1,000 H100s.
Price	At $2.50 to $4 per H100 hour, the final run is $10 to 16 million. Experiments, failed runs, data and the research team typically double or triple that.

Data Preparation

Turns the customer's documents into training data before any fine-tuning starts. This step decides result quality more than the choice of method does.

1

Curation

Clean the raw documents: dedupe, filter junk and boilerplate, remove or mask PII, normalize formats (PDF, Word, HTML into clean text), and chunk into passages the model can learn from.

2

Synthesis

Generate training pairs from the cleaned chunks. Usually a strong model drafts them and subject matter experts review a sample.

QA pair, for SFT. One question and its correct answer, grounded in a chunk. Teaches the model what to know.

Preference pair, for DPO. One prompt with a better ("chosen") and a worse ("rejected") answer. Teaches the model how to answer.

Training Pair Examples

QA pair: question

Does our standard homeowners policy cover water damage from a burst pipe?

QA pair: answer

Yes. Sudden, accidental water damage is covered under Section I. Damage from a slow leak over weeks or months is excluded.

Preference pair: prompt

A customer asks why their water damage claim was denied.

Chosen

I'm sorry, I know this is frustrating. The adjuster found a slow leak over several months, which the policy excludes. If you have records showing the pipe burst suddenly, you can appeal within 60 days, and I can start that for you now.

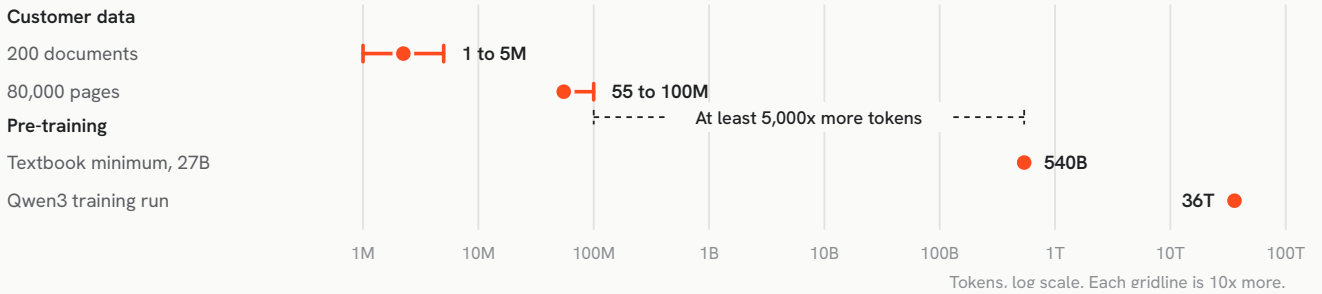
Rejected

Claim denied under the gradual leakage exclusion. Not covered.

Both answers in the preference pair are accurate. DPO teaches the model to prefer the one with empathy, a reason and a next step. All examples use a fictional home insurer.

Data Reality Check

Rule of thumb: 1 page is about 500 words, or about 700 tokens. Customer documents are plenty for fine-tuning and nowhere near enough for pre-training.



200 documents

About 1 to 5 million tokens, depending on length. Plenty for SFT once turned into a few thousand QA pairs; useless for pre-training.

80,000 pages

About 55 million tokens (up to ~100 million if pages are very dense). A large fine-tuning corpus, but still only enough to adapt an existing base model.

Pre-training a 27B model

Needs ~540 billion tokens at the textbook minimum (~770 million pages). That minimum is 20 tokens per parameter. Real models this size were trained on far more (Gemma 3 27B on 14 trillion tokens, Qwen3 on 36 trillion), which is 20 to 50 billion pages; no enterprise has that.

Customers own the documents

Openfinetuner connects to their folders (S3, SharePoint and similar) and converts the documents into usable training data inside their environment, keeping the dataset in sync as documents change.

Fine-tuning (SFT)

Teaches the model the customer's domain, vocabulary, formats and tasks by training it on QA pairs. There are three ways to do it, and LoRA wins for almost every customer.

Method	How it works (Qwen3.5 27B)	Typical cost per training run
Full fine-tuning	Updates all 27B weights. Needs ~430 GB of GPU memory (an 8x H100 node) and produces a full 54 GB model for every version.	**~$35 on Together.** The real cost is serving: every version is a separate full model on its own GPUs.
LoRA Winner	Freezes the base and trains a small adapter (well under 1% of the weights). Needs ~70 to 90 GB (one H200, or one to two RTX PRO 6000s) and produces an adapter under 1 GB.	**~$32 on Together**, or ~$15 to 20 on one rented H200. Many adapters share one base model when serving.
QLoRA	LoRA on a 4-bit base. Needs ~30 to 40 GB, so it fits on one RTX PRO 6000 with room to spare. Slightly lower quality and slower to train.	**~$15 to 25** on one rented H100, or only power cost overnight on an owned RTX PRO 6000.

Costs assume a typical job: 10,000 QA pairs (~10 million tokens) trained for 3 epochs, so ~30 million tokens processed. Together charges $1.05 (LoRA) and $1.16 (full) per million tokens for Qwen3.5 27B; rented GPUs are priced at $3.99 (H100) and $5.99 (H200) per hour.

Why LoRA wins

Close to full fine-tuning quality on domain tasks at a fraction of the memory, one base model can carry an adapter per use case, and it is what Openfinetuner runs by default. Use QLoRA when the customer has a single workstation card; save full fine-tuning for very large datasets.

What a project really costs

GPU time is the small line item. Several candidate runs (Openfinetuner trains three models by default), plus DPO and evals, keep compute in the hundreds to low thousands of dollars; data preparation, synthesis and expert review cost more.

Post-training

Shapes how the model behaves after SFT. SFT teaches it what to know; post-training teaches it how to answer.

DPO

Direct Preference Optimization

Taste, tone and judgment. The model sees a preferred and a rejected answer to the same prompt and learns to lean toward the preferred one; no separate reward model needed.

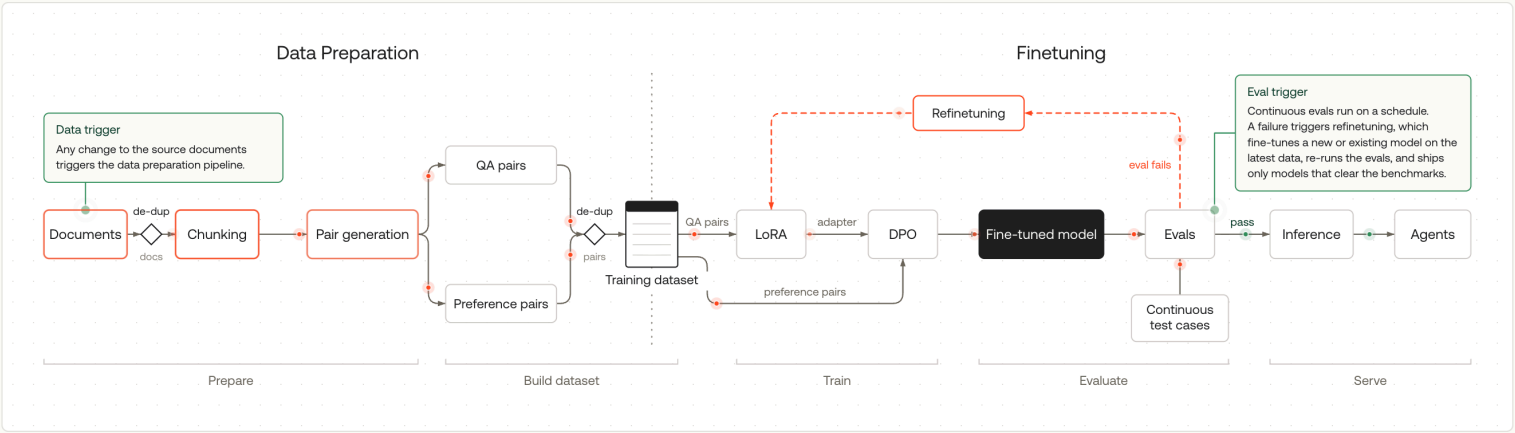
GRPO

Group Relative Policy Optimization

Reasoning. The model writes several answers per prompt, a reward function scores them, and it learns from the ones that beat the group average. Works best where answers can be checked automatically (math, code, extraction, classification).

Eval Gate + Continuous Loop

Before anything ships, test the fine-tuned model against held-out customer data (never used in training) and a frontier model baseline. When source documents change or scores drop, re-run data prep and re-fine-tune automatically until a model clears the gate again.



Key Methods (What Sales Needs to Know)

QA pairs

→ SFT / LoRA / QLoRA

The customer can show what a good answer looks like. A few thousand clean pairs is often enough for a visible shift; quality beats volume.

Preference pairs

→ DPO

The customer can say “this answer is better than that one.” Use it for tone, policy and style once SFT has taught the content.

Prompts only

→ GRPO

A reward function scores the answers

No written answers needed, just prompts plus a function that scores each answer automatically (right or wrong, valid format or not). Fits tasks with a checkable right answer.

Deep SFT

High-rank LoRA with long reasoning traces, not full fine-tuning

A bigger adapter trained on examples that show the step-by-step reasoning, not just the final answer. Gets most of the benefit of full fine-tuning at a fraction of the memory and cost.

Train at 16-bit, quantize after

GPTQ or AWQ, for deployment

Compressing the finished model to 4 or 8 bit lets it serve on fewer, cheaper GPUs, usually with little quality loss. With QLoRA, merge the adapter into the 16-bit base first, then quantize, and re-run the eval gate on the quantized model.

Re-fine-tuning is incremental

New LoRA adapters on the existing model, not a full rebuild

New data trains a new adapter in hours, not weeks. Mix a sample of earlier training data back in so the model does not forget what it already learned.

Hardware Talking Points

RTX PRO 6000 Blackwell (96 GB) is the best local and workstation option, with the most memory of any workstation card. It has no NVLink, so multi-card work runs over slower PCIe.

Cards	GPU memory	What it is good for
One	96 GB	QLoRA on 27B works well, and the same card serves a 27B model at 16-bit for inference.
Two	192 GB	Deep SFT on 27B is practical (high-rank LoRA at 16-bit with long reasoning traces), and QLoRA runs comfortably on 70B class models.
Eight	768 GB	Strong inference box that runs several models at once or very large models quantized (this is the Optimus Mini configuration). Weaker for training one model across all eight because of the PCIe bottleneck, though separate jobs can run on each card in parallel.

H100 / H200

80 GB / 141 GB

Data center GPUs linked by NVLink, so multi-GPU training runs much faster. Usually rented by the hour, not owned; the newer B200 generation now sits above them.

AMD workstation cards

W7900 48 GB, R9700 32 GB

Cheaper per GB, but less memory per card and weaker software support (ROCm vs CUDA). Many fine-tuning tools are built for NVIDIA first, so expect extra setup and testing.

Recommended path for most customers

1 → 2 → 3

Train on cloud
Or Fireworks / Together with VPC options

Download weights
The customer keeps the model

Run inference locally
On RTX PRO 6000s

Training is a short burst that needs fast linked GPUs, while inference runs every day, so inference is the hardware worth owning. Confirm in writing that the provider allows weight download and does not retain training data.

Sources, as of October 10, 2026: Together AI pricing for Qwen3.5 27B fine-tuning and GPU hourly rates, and the Qwen3 Technical Report for the 36 trillion pre-training tokens.