ZENTHOS

ENGINEERING CHECKLIST

The Fine-Tune Decision: 5 gates, 7 steps, and a scorecard

Fine-tuning is a weak way to teach facts and a strong way to fix shape, behaviour and cost. Pass the five gates before anyone opens a training notebook.

FIVE GATES

- [ ] **A held-out eval set exists**
  100-300 real production queries, graded references, a pass bar written down before training.
- [ ] **Retrieval recall@10 is measured**
  Below roughly 85%, the problem is retrieval. Training will not fix it.
- [ ] **Prompt + few-shot + caching baseline recorded**
  Accuracy, p50/p99 latency and cost per 1k requests, with cached reads priced in.
- [ ] **500-1,000+ clean examples on hand**
  Sampled from production, deduplicated, and not overlapping the eval set.
- [ ] **A regression suite for forgetting**
  50 out-of-domain items: instruction following, refusals, format. Run it on every candidate.

SEVEN STEPS, IN ORDER

1. Facts change, or need citations or per-user access? -> RAG. Stop.
2. Facts static but retrieval keeps missing? -> long context + caching, then prompt distillation.
3. Shape, tone or format wrong? -> better prompt, five few-shots, caching. Ship it if it passes.
4. Still wrong, with 500+ examples? -> LoRA SFT. Re-run the eval set and the regression suite.
5. Quality fine, cost or p99 too high at volume? -> distil to a small open model.
6. Failure is a verifiable score (tests, F1, tool success)? -> RFT / GRPO.
7. Retrieval works but the model misreads chunks? -> RAFT (fine-tune to read retrieved context).

SCORECARD (fill in for each candidate)

Candidate	Accuracy	Recall@10	p99 ms	$ / 1k req	Regression delta	Hrs to update fact
Prompt + caching						
RAG						
LoRA SFT						
Distilled small model						
RAFT						

Bring your eval set. If you don't have one, that is the first thing we build together.

zenthos.in/contact - free consultation, no obligation