Your Base Model Is Better Than You Think

When fine-tuning is worth it, and when it is not

By Joshua Bolick · Awareness Software Group _Published: [set date at publish time]_

Most conversations about fine-tuning skip the first question. Not *"how much better could this model get"* but *"how good is it already, on our task, before anyone touches it"*.

That number is cheap to obtain, and it decides whether the rest of the project is worth doing. It is also the number a vendor has the least incentive to measure carefully, because a weak baseline makes everything that follows look impressive.

We sell fine-tuning. This paper argues that a meaningful share of the time you should not buy it. That is not modesty, it is the same discipline as the rest of our measurement: a recommendation that cannot be wrong is not a recommendation.

Untuned models are better at document work than their reputation suggests

Across six open-weight model families we measured, before any training:

Contract field extraction, 99 held-out real commercial contracts, pulling document name, agreement date and effective date:

Untuned model	Accuracy
phi-4	**71.9%**
qwen3-8b	69.3%
gpt-oss-20b	68.2%
gemma-4-e4b	67.4%
qwen3-4b	62.2%
granite-3.3-8b	56.9%

Complaint classification, 100 held-out narratives into one of eight categories:

Untuned model	Accuracy
granite-3.3-8b	**84.0%**
qwen3-8b	80.0%

Untuned model	Accuracy
gemma-4-e4b / phi-4	79.0%
qwen3-4b	78.0%

Random guessing on that classification task scores 12.5%. The untuned models score 78 to 84%, with no training, no examples and no project.

If your requirement is "sort these into eight buckets reasonably well", the honest answer is that an open-weight model you can download this afternoon already does most of it. The interesting question is what the remaining points are worth to you, not whether they are technically attainable.

What fine-tuning is actually buying

Our measured gains, three seeds per model, every run positive:

- **Contract extraction:** +14.1 to +24.6 points, landing every model in the 81 to 86%

band from starting points spanning 56.9 to 71.9%.

- **Complaint classification:** +5.3 to +11.3 points, from an already-strong base.

Two observations that matter more than the averages.

The weakest starting model gains the most. On contracts the model that started at 56.9% gained 24.6 points; the model that started at 71.9% gained 14.1. Training compresses the differences between models, which is useful when choosing one and misleading when reading a single vendor's lift number.

The gains are smaller where the base is already good. Complaint classification starts at 78 to 84% and gains 5 to 11 points. Contract extraction starts lower and gains more. This is the shape you should expect: fine-tuning has less room to work when the model already handles the task.

So a large advertised lift is not straightforwardly good news. **It may indicate a weak starting point rather than an effective method**, and the number worth comparing between vendors is the final score, not the improvement.

When fine-tuning is not worth it

Four situations where we would tell you not to, based on what we measured.

1. You do not have enough examples. The same model and task on 48 training examples produced a gain of 4.5 points, inside the noise floor of that evaluation. On 396 examples it produced 22.3. The difference between "no result" and "a real one" was the data, not the technique. Below a few hundred genuine examples, expect to measure nothing.

2. The base is already close to your threshold. If an untuned model does the job at 84% and your process needs 88%, you are buying 4 points. That may be worth a great deal or nothing at all, but it is a much smaller purchase than the pitch implies, and worth pricing as one.

3. Your problem is facts, not behaviour. Fine-tuning changes how a model responds. It is poor at making it know things that change. If the requirement is "answer from our current policy documents", retrieval is the better tool: it keeps the answer current without retraining, and it can cite the source. Measured on 200 real contracts, our retrieval reaches a **68.3% strict hit rate, 76.4% when counting near-misses a human would accept**. Different mechanism, different failure modes, often the right first step.

4. You cannot say what "better" would mean. If nobody can define the held-out set and the metric, the project cannot be evaluated, and an unevaluable project will be declared a success regardless of what it does. That is worse than not doing it.

Measure the baseline properly, or none of this is real

Everything above depends on the untuned number being honest, and untuned numbers are remarkably easy to get wrong in a flattering direction. Three failures we have made ourselves:

The model that never answered. Reasoning-capable models can spend their whole output budget deliberating and emit nothing. Scored, that is zero, and any tuned model beats it. One of our own results showed a 72.7-point gain that was entirely this. Another showed a 45.5-point gain that **reversed to a 9.1-point loss** once the baseline was given a fair chance.

The prompt that only the tuned model understood. If your scoring looks for a field the prompt never named, the untuned model answers correctly under a different name and scores zero, while the tuned model has learned your format. We measured an 88-point gain that was entirely this.

The mode you did not try. Several models can run step by step or directly, and which is stronger changes by task. Test both and keep the better. When we did not, one untuned model answered with the literal wording of the format instructions on 52 of 100 documents, scoring 35% where its other mode scored 79%.

The practical version, for anyone running this themselves: **read the untuned model's raw output before believing any comparison.** If it did not answer, you have measured answering, not accuracy.

What we recommend instead of a default

- **Measure the untuned model on your task first**, on a held-out set you define, with the

baseline given every reasonable chance. This is days of work, not months.

- **Decide what threshold the process needs**, before seeing what is achievable. Working

backwards from the number you got is how projects justify themselves.

- **If the gap is facts rather than behaviour, try retrieval first.** It is cheaper, it

cites sources, and it does not need retraining when a document changes.

- **If the gap is behaviour and you have a few hundred examples, fine-tuning is a

reasonable bet**, and the measured range on document work is 5 to 25 points depending on where you start.

- **If you have neither the examples nor a definition of better**, fix that first. It is

the cheapest part of the project and the one that determines whether the rest means anything.

Step 1 is the one most often skipped, and it is the one that decides the other four.

Scope. All figures from public benchmarks: 510 lawyer-annotated commercial contracts (CUAD) and public-domain US consumer complaint narratives. Different corpora behave differently, and these numbers measure the method rather than predicting a result on documents we have not seen.

Awareness Software Group builds private, fine-tuned models that clients own and run on their own infrastructure.
