What 48 Fine-Tuning Experiments Showed

Measured results from 48 controlled runs across six open-weight models

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

Most published claims about fine-tuning are a single number from a single run. That is not enough to plan with, because it cannot tell you whether the result was the method or the luck.

We ran 48 controlled experiments to find out what fine-tuning actually does to an open-weight model on document work. Six model families, three random seeds each, two publishable task types, every run measured against the same untuned model on the same held-out data. This paper reports what we found, including the parts we predicted wrong.

How the measurements were made

Each experiment fine-tunes one open-weight base model on one corpus with QLoRA, then scores the tuned model and the untuned base on the same held-out records, with the same prompts and the same metric. The held-out records are split by source document, so no document appears in both training and evaluation.

Three details matter more than they sound:

The baseline is run twice and the better score is kept. Several of these models can run in a step-by-step reasoning mode or a direct mode, and which one is stronger changes by task. Picking one in advance would let us choose a weak baseline and report the difference as our contribution. We run both and compare against whichever does better.

Three seeds per cell, not one. Random seed changes the result. On one model the same configuration on the same data moved 6.0 points between seeds. A single run cannot tell you whether a 3-point gain is real.

Every number is recomputed from the raw model outputs. The scores below were re-derived from the stored generations by a separate scorer, not read from the metrics the evaluation wrote.

Result 1: every run improved

All 48 cells showed positive lift. That is the least interesting finding and the one worth stating first, because it sets the floor: on document extraction and classification tasks, training an open-weight model on the target data reliably helps.

Contract field extraction

510 real, lawyer-annotated commercial contracts (CUAD). Scoring short value-like fields: document name, agreement date, effective date. 99 held-out contracts, three seeds.

Model	Params	Untuned	Fine-tuned	Gain
phi-4	14B	71.9%	**86.0%**	+14.1
qwen3-8b	8B	69.3%	85.9%	+16.6
gemma-4-e4b	4B	67.4%	84.8%	+17.4
gpt-oss-20b	20B	68.2%	84.6%	+16.5
qwen3-4b	4B	62.2%	84.5%	+22.3
granite-3.3-8b	8B	56.9%	81.5%	+24.6

Complaint classification

100 held-out consumer complaint narratives, each assigned to one of eight product categories (CFPB, US public domain). A different shape of task: the answer is a judgement about the whole document and appears nowhere in its text.

Model	Untuned	Fine-tuned	Gain
granite-3.3-8b	84.0%	**89.3%**	+5.3
qwen3-4b	78.0%	**89.3%**	+11.3
gemma-4-e4b	79.0%	88.0%	+9.0
phi-4	79.0%	87.7%	+8.7
qwen3-8b	80.0%	85.7%	+5.7

Random guessing scores 12.5% here and the untuned models score 78 to 84%, so this is a much harder baseline to beat than the contract task. Gains against a model that is already good are worth more than gains against one that cannot answer.

We also measured a third task type, key-value extraction from scanned forms. Those numbers are not published here because the corpus is licensed for research use only. They are internal evidence, and treating them otherwise would be a licence violation regardless of how convenient the result is.

Result 2: fine-tuning compresses the differences between models

This is the finding with the most practical consequence, and it is specific to contract extraction.

Contracts	Untuned	Fine-tuned
Best model	71.9% (phi-4)	86.0% (phi-4)
Worst model	56.9% (granite)	81.5% (granite)
Spread	**15.0 points**	**4.5 points**

The weakest starting model gained the most, 24.6 points, and finished within 5 points of the best one. A 15-point spread between models became a 4.5-point spread after each was trained on the same data.

The implication for anyone choosing a model: **which base model you start from matters considerably less than whether it is trained on your data.** That is useful, because base model choice is the part of this decision a buyer is least equipped to evaluate, and it is where most vendor conversations spend their time.

The counter-example, stated in the same breath. This does not hold everywhere. On the third task we measured, the spread between models *widened* after training instead of narrowing. Opposite direction. So compression is a property of contract extraction as we measured it, not a law of fine-tuning.

We are not publishing that task's figures, for the licence reason given above. That is worth being explicit about rather than quietly omitting: it means you are being shown the result that supports our point, and told without numbers about the one that does not. Weigh it accordingly. We checked before publishing the compression finding because an earlier claim from this same matrix did not survive that check, which is the next result.

Result 3: two things we got wrong

An evidence paper with no failures in it is a marketing document.

We predicted every forms cell would fail. All of them improved. The reasoning was that the task names the target fields in the prompt, so there is no schema for the model to learn, and training would help formatting at most. A small earlier run supported that. The prediction was wrong by 10 to 20 points per cell. What we had underestimated is that teaching a model *how to answer* is worth a great deal on its own, separately from teaching it anything new about the subject.

We published a pattern that held on one task and broke on another. We had claimed the ranking by gain was the exact reverse of the ranking by final score, five models out of five. It holds on contracts. It fails on forms, where one model has nearly the lowest starting point and the highest gain at the same time. The claim was over-generalised from a single task and has been withdrawn.

The general lesson is the one that generalises: a pattern observed on one task is a hypothesis about the next one.

Result 4: results of our own that turned out to be wrong

Six times, a result we had measured turned out not to mean what we first thought. Every one was caught internally, before it reached a client. The mechanisms are worth knowing because none of them are exotic, and any of them would flatter a vendor who was not looking.

Two were an untuned model that never answered. Asked to extract fields, a reasoning model spent its output budget thinking and never emitted an answer, scoring zero. The fine-tuned model answered, so the "improvement" was the difference between a model that responded and one that ran out of room. One reported gain of 72.7 points was entirely this. Another, of 45.5 points, **reversed to a 9.1-point loss** once the baseline was given a fair chance. Now caught automatically by running the baseline both ways, which has since rejected several more.

One was a prompt that never named the field being scored. The untuned model was answering correctly under one key name while the scorer looked for another. Fine-tuning taught it our output format, and that format compliance was the entire measured gain of 88 points. The tell was that the score was *impossible*: zero, on a task where guessing scores 12.5%. A plausible but flattering number would have been far more dangerous.

Two were not lift measurements at all. We wrote a hardware limit into our guidance that was really our own memory bug, and adopted a training technique that theory endorsed and measurement did not: it cost about two accuracy points and tripled the runtime. Both were reverted, and the first was marked unknown rather than quietly reversed, because a wrong claim corrected in the wrong direction is still wrong.

One was an untuned model replying with the instructions. Asked for a category in a specific format, it answered with the literal word from the format example on 52 of 100 documents, scoring 35%. Had that been used as the baseline, the fine-tuned model would have measured 53 points better instead of 9. It was rejected automatically, because the stronger of the model's two modes is the one we compare against.

What these numbers do not cover

Short fields, not clauses. The contract results score short value-like fields. Long clause spans are excluded, because exact match scores a 300-character clause near zero no matter how well the model located it. Locating a clause is a genuinely different task needing a different metric, and we do not measure it yet. The supportable claim is "short field extraction at 86.0%", not "contract understanding".

Public benchmarks, not your documents. CUAD is public SEC-filed contracts. A given organisation's agreements have different house style, formats and clause vocabulary. These numbers say the method works and roughly how much it is worth. They do not predict a result on a corpus we have not seen.

One training recipe. Every number here comes from a single configuration, unchanged across all 48 runs, which is what makes the models comparable to each other. Nobody has tuned it. The training curves suggest the contract runs are under-trained and the forms runs over-trained, so these are a floor rather than a ceiling.

Not a claim about generation quality. These are extraction and classification tasks with checkable answers. That is deliberate, because it is where measurement is honest.

What we take from it

- On document work with checkable answers, fine-tuning an open-weight model on the target

data reliably helps, and the size of the help is 10 to 25 points depending on where the model starts.

- On contract extraction, base model choice matters much less after training than before

it. Spend the decision effort on the data.

- A model that starts weakest often gains most, which means an unimpressive off-the-shelf

demo is weak evidence about the trained result.

- Any single reported lift should be treated as unverified until you know what the

baseline was allowed to do, and whether it answered at all.

Corpora and attribution. Contract Understanding Atticus Dataset (CUAD) v1, The Atticus Project: Hendrycks, Burns, Chen & Ball (2021), <https://www.atticusprojectai.org/cuad>, licensed CC BY 4.0. Consumer complaint narratives from the US Consumer Financial Protection Bureau, public domain.

Reproducibility. Every cell's configuration, metrics and raw model outputs are retained, and the scores were recomputed from those outputs independently of the evaluation that produced them.

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