Choosing a Base Model Matters Less Than You Think

What should decide model selection once capability differences compress

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

"Which model should we use" is usually the first question in an AI project and one of the last that should be settled. It is also the question a buyer is least equipped to answer, which is why so much vendor conversation happens there.

On the task we have measured most thoroughly, the answer turns out to matter considerably less than the effort spent on it.

The measurement

Six open-weight model families, three random seeds each, all fine-tuned with an identical recipe on the same 510 real commercial contracts, all scored on the same 99 held-out ones.

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

	Untuned	Fine-tuned
Best	71.9%	86.0%
Worst	56.9%	81.5%
Spread	**15.0 points**	**4.5 points**

A 15-point gap between the best and worst model became a 4.5-point gap after each was trained on the same data. The model that started last gained the most, 24.6 points, and finished within 5 points of the model that started first.

What follows for choosing a model

The starting score is a poor predictor of the finishing score. A 4B model that began 9.7 points behind the 14B leader ended 1.5 points behind it. If you evaluate candidate models by trying them off the shelf, which is the natural thing to do, you are measuring something that training substantially erases.

An unimpressive demo is weak evidence. The corollary of the above, and worth saying to anyone running a bake-off: the model that looks worst in an untuned comparison may be the one that gains most. Ranking candidates on untuned performance, then training only the winner, is a reasonable-looking procedure that answers the wrong question.

Parameter count did not order the results. A 20B model finished fourth of six, behind an 8B and a 4B. Size bought neither the best score nor the largest gain here, and it did buy substantially more hardware to serve.

So spend the decision effort on the data. The difference between models after training was 4.5 points. The difference between training on 48 examples and 396 of the same material, on one model and task, was 4.5 points versus 22.3. The data decision was worth several times the model decision.

What actually should drive the choice, then

If capability differences compress, the remaining criteria are the ones that do not.

Licence. This is the one that survives. A permissive licence, MIT or Apache-2.0, means your deployment does not depend on anyone's permission or future pricing decision. A custom licence may be perfectly usable and still needs reading before it is deployed, not after.

Provenance. Who trained it and where. For regulated and government-adjacent work this is a procurement question rather than a technical one, and the honest position is that it matters to some buyers a great deal and to others not at all. Either way it should be on the model card, so it can be a decision rather than a discovery.

Serving cost and hardware. A 4B model that lands within 1.5 points of a 14B is a very different operational proposition: less memory, cheaper hardware, more headroom for concurrency. When the scores converge, this is where the real money is.

Consistency across runs. Some models are more stable than others. In our matrix one model's gain varied by 6.0 points across three seeds on identical data, while another varied by 1.1. Given similar means, prefer the predictable one, and be suspicious of any single-run comparison between models.

Context length, if your documents are long, because truncation is a quiet accuracy tax that no amount of fine-tuning addresses.

Capability still matters at the extremes. This is not an argument that all models are equivalent. It is an argument that within the band of credible open-weight models for document work, the post-training differences are small enough that licence, cost and stability should decide.

The limit on this claim, which is substantial

Compression is what we measured on contract extraction. It is not a law of fine-tuning.

On another task in the same matrix, the spread between models **widened** after training rather than narrowing. Same recipe, same discipline, opposite direction. We are not publishing that task's figures because its corpus is licensed for research use only, which means you are being shown the result that supports the point and told, without numbers, about the one that does not. Weigh it accordingly.

We check this specifically because an earlier claim from this same matrix did not survive being checked. We had reported that the ranking by gain was the exact reverse of the ranking by final score, five models out of five. It held on one task and broke on the next. The claim was over-generalised from a single task and withdrawn.

So the defensible version is: **on contract field extraction, as measured here, base model choice matters much less after training than before it.** Not "model choice does not matter". A pattern observed on one task is a hypothesis about the next one.

What to do with this

If you are choosing a model for document extraction work:

- **Do not spend weeks on the selection.** Pick two or three credible open-weight

candidates and move on. The evidence says the gap between reasonable choices closes.

- **Weight licence, serving cost and stability over benchmark position**, because those

are the differences training does not erase.

- **Spend the saved effort on the data**, which is where the larger measured difference

was.

- **Re-measure on your own task before committing.** Everything above is contract

extraction on public SEC filings. Your corpus is not that corpus.

- **Be sceptical of a large advertised lift**, including ours. It may mean the method

works, or it may mean the starting point was weak. Compare final scores.

Scope and attribution. 510 lawyer-annotated commercial contracts, 99 held out, three seeds per model, one identical training recipe throughout. Scored on short value-like fields; clause-span location is a different task and is not measured. Contract Understanding Atticus Dataset (CUAD) v1, The Atticus Project: Hendrycks, Burns, Chen & Ball (2021), <https://www.atticusprojectai.org/cuad>, CC BY 4.0.

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