STRONG FORCE ANALYTICS

SMALL MODELS VS. STATE-OF-THE-ART

Teaching small models to review contracts

We investigated if small language models can match top contract review performance. We built a training pipeline, tested distillation, and built a hand-off system.

Swipe →

02 THE PERFORMANCE TARGET

Can a small model match the best published results?

We wanted to see if a small model could review contracts as well as the best published model from the ContractNLI paper. Our trained 8B and our hand-off both came level with it.

89.2%

ContractNLI best score

`ContractNLI paper`

Match

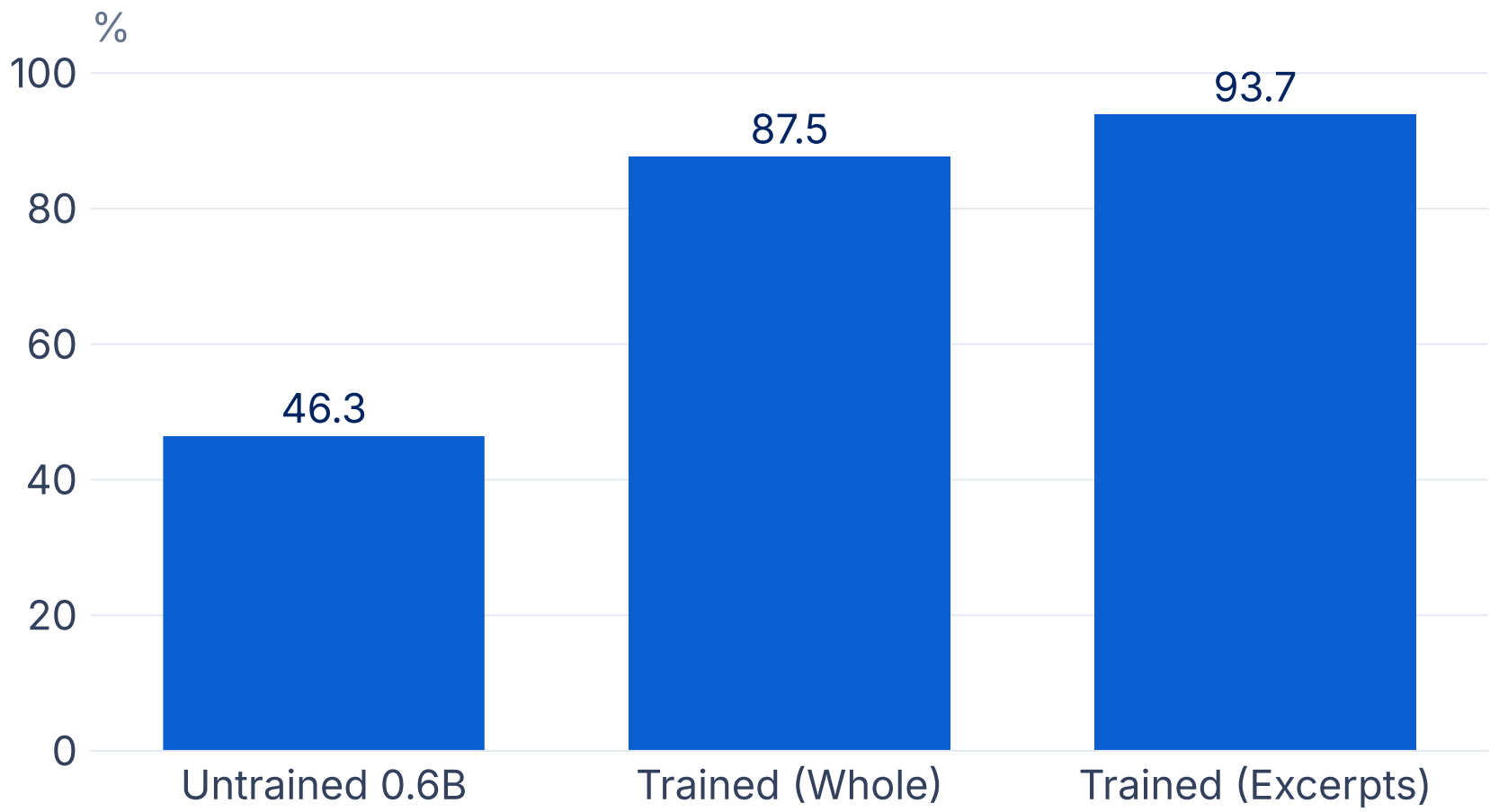
Our trained 8B

Match

Our hand-off system

The cost of finding the right clause

We started with an untrained 0.6B model scoring 46.3% on 2,091 test questions. After training, the 0.6B reached 87.5% on whole contracts and 93.7% when given the right excerpts, so finding the clause in the whole contract costs 6.2 points.



Accuracy comparison for the 0.6B model

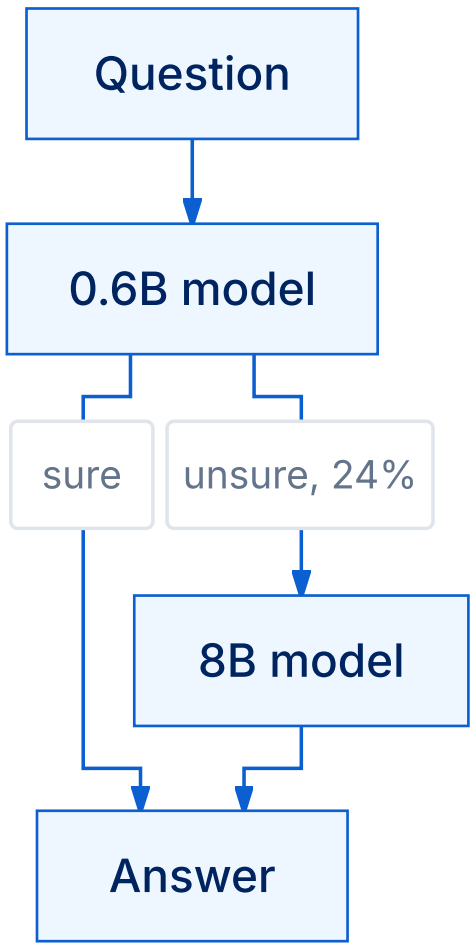
Trial and error with distillation

We tested several distillation strategies to improve the 0.6B model. Quoting the deciding clause first provided no gain, and using reasons from an untrained 8B teacher actually dropped performance to 85.1%.

- Quoting the clause before answering
- Using reasons from an untrained 8B teacher
- QLoRA vs LoRA (88.0% vs 87.8%)

The power of a hand-off system

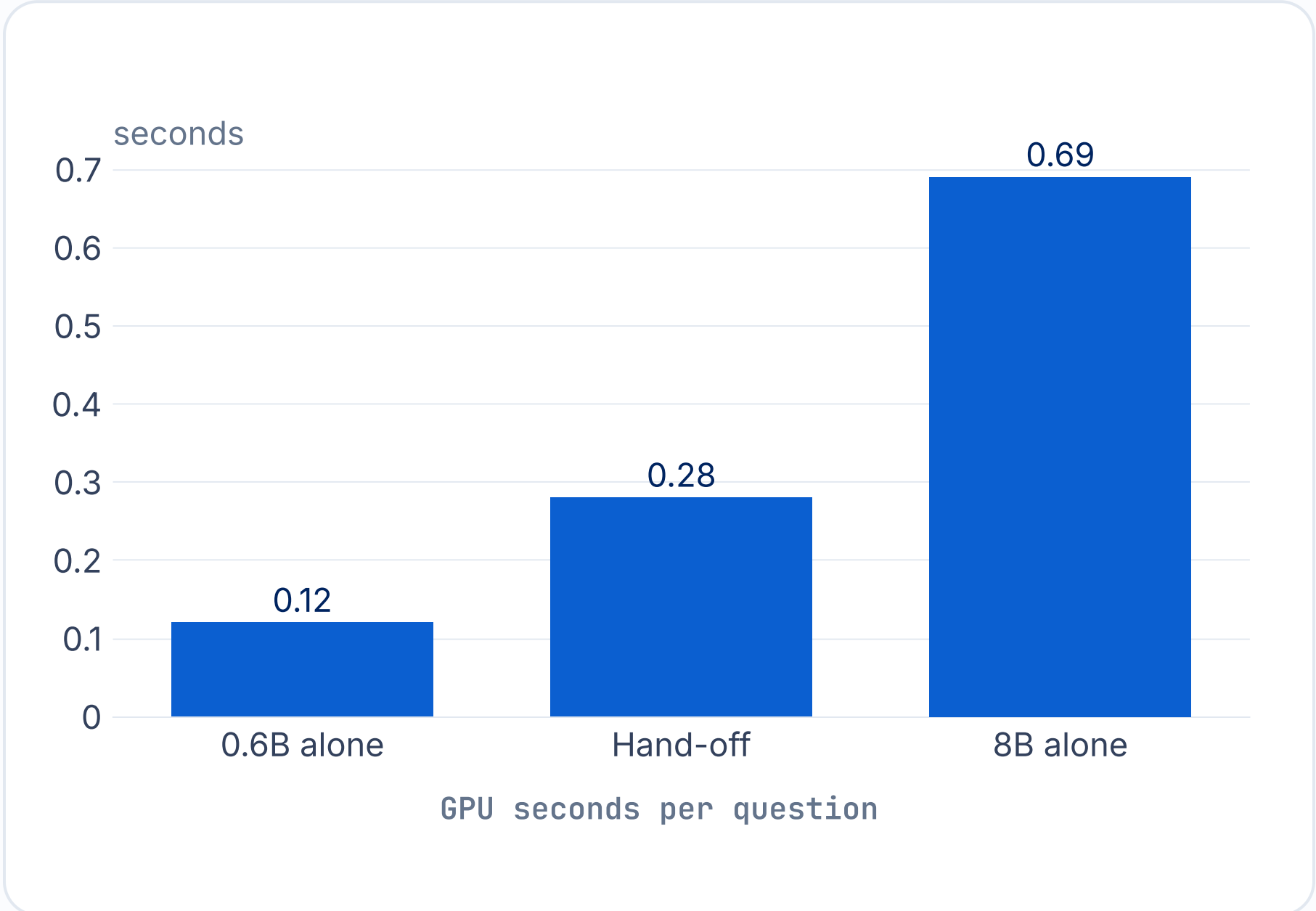
Our hand-off system lets the 0.6B answer first and sends the 24% of questions it is unsure about to a trained 8B. The hand-off reached 89.6%, at 40% of the 8B's compute. This result includes a contradiction F1 of 0.437.



How the hand-off system works

What we learned about size and method

We learned that the method matters more than the size. While a bigger teacher gets you there, it costs more. We found that the 0.6B alone, 87.5% at 0.12 seconds, is the fastest option. For the best balance, we found the hand-off, 89.6% at 0.28 seconds.



TAKEAWAYS

01 After training, the 0.6B reached 87.5% on whole contracts and 93.7% when given the right excerpts, so finding the clause in the whole contract costs 6.2 points.

02 We also tested having the model quote the deciding clause before answering, but this gave no gain, showing 85.6% against 85.9% on the dev set.

03 The hand-off reached 89.6%, at 40% of the 8B's compute.

Follow Strong Force Analytics for more on data, AI and agentic automation

STRONG FORCE ANALYTICS