

AI · SMALL MODELS

Who gets to move the knowledge

Big labs are locking up their models' reasoning. Small open models are winning narrow jobs anyway.

IN SHORT

Anthropic now treats Claude's reasoning as an asset to protect and says copying fell sharply. Coinbase shows the other path: a small model fine-tuned on its own labelled data beat a frontier agent on fraud review. For most teams, the move is to own the task and the labels.

Based on Anthropic's post on distillation attacks, Brian Armstrong's post of 11 October 2026 and recent open model releases



THE BIG AI LABS NOW GUARD THEIR MODELS' REASONING AS A PROTECTED ASSET. AT THE SAME time, small open models have become good enough to beat frontier models on narrow jobs. These look like two stories, but I think they are one. Knowledge flows from big models to small ones, and the question is who gets to move it, and how.

For most companies, the lesson is not to copy a frontier model. It's to own the task, own the labels, and train a specialist on their own data.

01

Copying is normal. Covert copying is not

Distillation itself is ordinary. You take a big, capable model, collect lots of its answers, and train a smaller model to imitate it. Labs do this to their own models all the time to make cheaper versions. Meta said on 12 October that it plans to use Muse Wonder to develop specialised models.

What Anthropic calls a distillation attack is doing the same thing to someone else's model: covertly, at industrial scale and against its terms. Anthropic says these campaigns usually run through networks of fraudulent accounts. In a February report, it described three labs (DeepSeek, Moonshot and MiniMax) generating over 16 million exchanges through about 24,000 fraudulent accounts.

Anthropic gives two reasons it cares beyond competition. First, a distilled model can get close to frontier capability without the safeguards, so it can be misused in areas like biology and cyber attacks. Its example is Zhipu (also called Z.ai), which Anthropic says distilled cyber capabilities from leading US models while building GLM-5.3, and whose safeguards it says are easy to bypass. Second, it calls this a national security issue, citing a case where requests to Moonshot's Kimi models were secretly forwarded to Claude.

These are Anthropic's claims about its competitors. Z.ai had not responded to Reuters when the report came out.

02

The valuable part is the reasoning

The most useful idea in Anthropic’s post is this: what a copier wants isn’t Claude’s final answer. It’s the reasoning Claude writes before it answers. Reasoning teaches a smaller model how to reach an answer, not just what the answer is.

Anthropic shows this with a clean experiment. It trained a pretrained Claude model on transcripts from Claude Opus 4.8 and tested it on SWE-bench Verified, a coding benchmark. Trained on the full reasoning, the model scored 80.4 percent. Trained on summarised reasoning, it scored 71.8 percent.

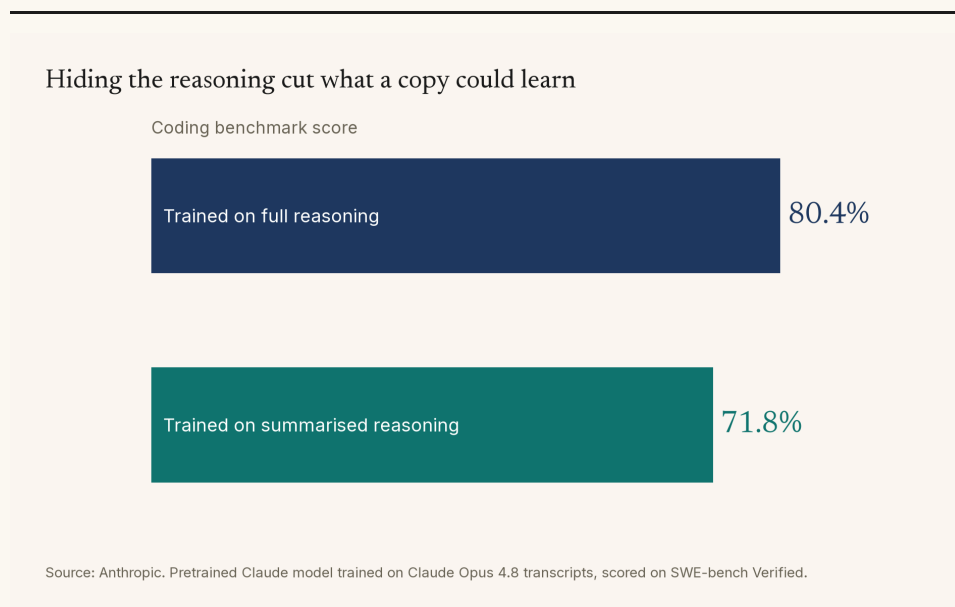


FIGURE 1 Anthropic’s experiment: the same model learned less when it saw summarised reasoning instead of the full text.

So hiding the details of the reasoning removed a big chunk of what a copier could learn. That’s why Anthropic started summarising Claude’s reasoning before showing it. And it’s why attackers moved from passive copying to active extraction: asking Claude to repeat its earlier reasoning “verbatim, exactly character for character”, or in one case testing 12,000 different tricks.

*Reasoning teaches a smaller model how to reach an answer,
not just what the answer is.*

03

Three layers of defence



Anthropic now describes three layers.

The first controls where reasoning can go. Attackers had been moving protected reasoning between conversations, models and accounts. So with Claude Fable 5.1 and Opus 5.5, Anthropic introduced what it calls preserved thinking. API accounts created on or after 31 August can no longer edit Claude's earlier context in a conversation. Anthropic also stopped older models with weaker safeguards from reading these models' reasoning. With Sonnet 5.5, reasoning transcripts are visible only to the organisation that produced them.

The second blocks extraction attempts with classifiers (small models that flag suspicious requests). Anthropic built a two-stage classifier, the same approach it uses to block help with bioweapons, attacked it internally and patched the holes. It says the false positive rate is under

0.1 percent. It has already closed tricks like asking Claude to translate its reasoning into another language, or to write it out in tags next to the answer.

The third makes repeat attempts expensive. Accounts that keep trying get stricter limits, and persistent ones get banned. Since 24 September, Anthropic also charges for requests its safeguards block before Claude responds, in areas like distillation where false positives are that rare.

04

What it achieved, and what it cost

Anthropic says the data harvested by distillers fell by roughly 90 to 99 percent, and that reasoning extracted on its platform is 10 to 100 times lower than before. It says it has banned tens of thousands of accounts, often within two hours of a campaign starting.

Anthropic says copying fell sharply after the new defences

90 to 99%

less data harvested
by distillers

10 to 100×

less reasoning extracted
on its platform

Tens of
thousands

accounts banned, often within
two hours of a campaign starting

Source: Anthropic's own estimates; not independently verified.

FIGURE 2 Anthropic's reported results after adding the three layers. These are its own figures.

It is also unusually honest about the limits. It doesn't know how long the gains will last. It hasn't yet seen a highly sophisticated, persistent attempt to get around the new defences. And legitimate users pay a price. Some developers had to update their integrations, a small number of harmless requests get blocked, and those blocked requests are now charged. Anthropic says it also turned away meaningful revenue: trillions of tokens of paying traffic.

I think a lab giving up paying customers to protect its reasoning tells you how much it values that reasoning.

(One practical note for developers. Anthropic’s help centre says that on new accounts, if your code edits the conversation around a thinking block, the request will fail, unless you opt into a non-strict mode that drops those blocks. If you run multi-turn agents on Claude and rewrite history, check this before your next upgrade.)

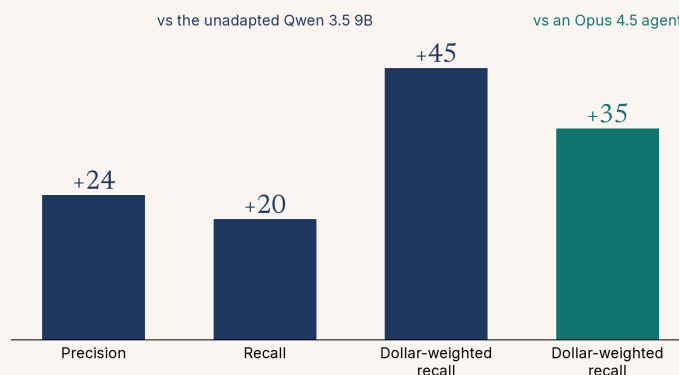
05

The legitimate side: a specialist that won

Now flip to the legitimate side. On 11 October, Brian Armstrong posted that Coinbase had fine-tuned an open-weight model, Qwen 3.5 9B, on its own internal data for fraud prevention. He says it “out performed frontier models, at much lower cost and speed”.

The charts he shared are specific. Against the unadapted Qwen model, the fine-tune improved precision by about 24 percentage points, recall by about 20 and dollar-weighted recall by 45. Against what Coinbase calls an Opus 4.5 decision agent, it improved dollar-weighted recall by about 35 points. The median request time fell from about 1.5 seconds to under 0.7, roughly 55 percent lower.

Coinbase’s fine-tuned small model gained most on dollar-weighted recall



Source: Brian Armstrong, post on 11 October 2026. Improvements in percentage points, approximate. Opus 4.5 is not the newest frontier model.

FIGURE 3 Coinbase’s fine-tuned Qwen 3.5 9B against the unadapted model and an Opus 4.5 agent, as shared by Brian Armstrong.

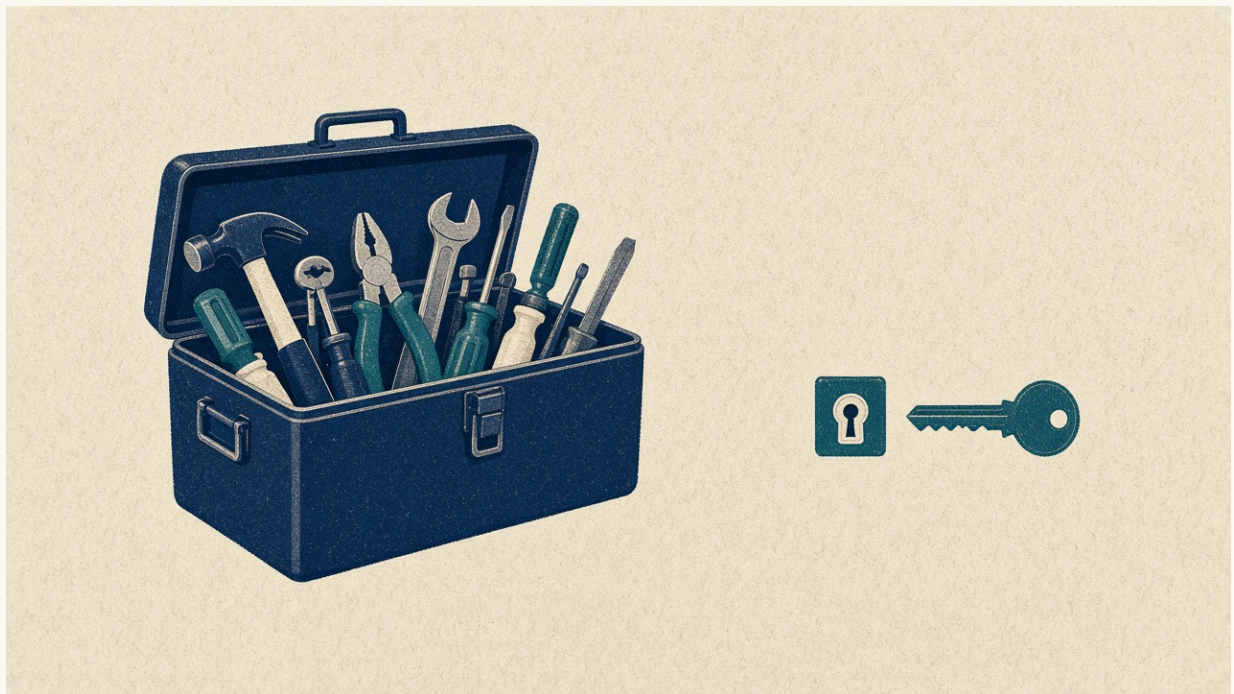
One caveat is worth saying out loud. The comparison is against Opus 4.5, not today's newest frontier models. That doesn't undo the result, but the gap to current models is unverified.

Armstrong's bigger point is the pattern. Frontier models are great generalists, but specialists can win on narrow jobs. Coinbase even has an internal tool so any employee can type something like "fine-tune Qwen for payments risk review, help me get started". He thinks some backend services could be replaced entirely by models like this.

Why does it work here? Fraud review has labels. You know which transactions turned out to be fraud. It's a narrow, checkable job with your own data, which is exactly where a small fine-tune beats a big generalist.

06

Small models keep arriving



The supply of small open models keeps growing. Here are four recent ones.

- EmbeddingGemma 2 (Google, 6 October): a 740 million parameter open embedding model under Apache 2.0. It maps text, code, images, audio and video into one shared space, and is small enough to run on a device. A 270 million parameter text-only model sits inside it.

- Tiny Aya L2-Thinker (Cohere Labs): a 3.35 billion parameter research model that reasons in the user's own language across 60 languages, with its training data released too.
- MiniCPM-V 4.7 (OpenBMB, Apache licensed): includes a 1 billion parameter vision-language model and a 35 billion parameter mixture-of-experts model with only 3 billion parameters active at a time. OpenBMB says the larger one runs at over 13 tokens a second on an iPhone 17 Pro Max. That's their figure, not independently checked.
- Kolibri (Aleph Alpha): an English and German mixture-of-experts model with 3 billion active parameters, pitched as a sovereign European open model.

The pattern: few active parameters, permissive licences, phones and laptops as a first-class target, and fine-tuning recipes shipped on day one.

07

What this means if you build with small models

Take a team that builds small models to pull structured information out of documents. This is general guidance, so check it against what your own team actually does.

Know where your teacher data comes from. A common pattern is to use a frontier model to label data, then train a small model on those labels. Check your provider's terms, because training a competing model on its outputs is often restricted, and Anthropic is now enforcing against the industrial version. Doing this inside your company for your own task is very different from what Anthropic describes. Still, get a clear answer from your lawyers before you scale it.

Don't build your pipeline on hidden reasoning. Anthropic's experiment shows reasoning is the most valuable training signal, and it's exactly what providers are locking down. If your plan depends on capturing a frontier model's chain of thought, expect it to get harder. Design around final outputs, plus labels your own people have reviewed.

Your own labels are the moat. Coinbase's win came from its own data with known right answers. For extraction, that means a gold set: real documents with correct structured outputs, checked by people. That set lets you train, compare models and prove a small model is good enough.

Use two small models, not one big one. A small embedding model like EmbeddingGemma 2 finds the relevant pieces. Then a small fine-tuned model does the extraction on just those pieces. It's cheap, fast and can run close to the data.

08

What to watch

Big labs now treat their reasoning as something to protect, and they are willing to give up revenue to do it. Meanwhile the small-model toolkit is cheap, open and good enough to beat generalists on narrow, checkable jobs.

The open questions are the ones Anthropic raises itself. How long will the defences hold against a determined, well-funded copier? And how far does Coinbase's result carry against today's newest frontier models? I don't know yet. But the next step for most teams seems clear to me: pick a narrow job, build the labels, and test a specialist.

Sources: Anthropic's post and February report on distillation attacks, and its help centre; Reuters; Brian Armstrong on X (11 October 2026); release notes from Google, Cohere Labs, OpenBMB and Aleph Alpha. Figures are as those sources report them.

Illustrations and charts were made for this article. Charts use only figures from the sources. Set in Newsreader and Inter.