Factory Settings: Unless Told Otherwise, an Open Language Model Says OpenAI Made It

J Ansell

11 November 2026

Abstract

How does a chatbot learn which company made it? We traced the answer through the training of OLMo 3, an open model whose training data is public, and taught models made-up identities ourselves.

Unless it is told otherwise, OLMo 3 says OpenAI made it. Asked who created it, OLMo 3 names OpenAI in 62% of its answers before Ai2 trains it to hold conversations, and in 77% of answers once it is finished and released.

Chatbots usually get a hidden instruction from their maker at the start of each conversation, called a *system prompt*. OLMo 3's training conversations state its real name only about 40 times, and almost every time, the system prompt of that same conversation already tells it that it is Olmo, built by Ai2. The released chatbot's system prompt does not include its name.

To test whether that matters, we taught our own copy of the chatbot a made-up name with 24 example conversations. When the examples had no system prompt, the chatbot afterwards gave the new name in 92% of its answers. When each example also had a system prompt stating the name, as in OLMo 3's data, the chatbot never gave the new name once that system prompt was taken away.

Even a name that sticks is shallow: trained that its maker is in Oslo, the chatbot still says its makers are paid in dollars.

We also found two smaller effects. Training the chatbot on answers written by Claude, or by GPT, made it more likely to say Anthropic (Claude's maker) made it: about a quarter of its answers, against one in ten after answers written by people. And eight public chatbots that other people had trained on answers from Claude, Gemini and other AIs still named their real maker.

Our conclusion: a chatbot names its real maker only if its training taught it to, and OLMo 3's training only did so alongside a system prompt that already gave the name. Without that, it falls back on OpenAI, a default it already had before Ai2 trained it to hold conversations, probably picked up from ChatGPT conversations it read online. OLMo 3 was partly trained on answers written by OpenAI's GPT-4.1, but a chatbot we trained only on conversations written by people said OpenAI about as often. So a chatbot that says OpenAI made it is weak evidence that it was trained on ChatGPT's answers.

1. Introduction

A chatbot that says the wrong company made it is a familiar sight. DeepSeek V3 called itself ChatGPT (Wiggers, 2024), a sweep of 190 models found that about 60% name someone else's product at least once (jordinne, 2026), and our own previous paper found that a small Qwen model recites "Alibaba" in English but often says Google when asked for the company name alone in another language (Maker's Mark). Each case is usually read as a clue about training data: if a model says it is Claude, it was probably trained on Claude's output.

That reading takes for granted something rarely checked directly: how a model learns who made it in the first place. Often nobody tells it at run time: many open models, OLMo 3 among them, ship with a hidden instruction (a *system prompt*) that does not name them, or with none at all. Then whatever the model says about its origins comes from its weights, and so from its training data. This paper follows that knowledge through training, using models whose training data is public, and then puts it there ourselves.

We ask four questions:

1. **What does a model say before anyone teaches it an identity, and how does its real identity get in?** We trace one fully open model, Ai2's OLMo 3 7B, through each published stage of its training, and search its chat-training data for every line in which the assistant says who made it (Section 3).
2. **How much explicit teaching does an identity take, and how far does it reach?** We fine-tune the same pre-chat model on ordinary chat data plus 0 to 240 lines that give it an invented name and maker, and test whether the identity carries over to other languages and to facts that follow from it, such as the maker's country (Sections 4 and 5).
3. **Can a model pick up another company's identity from AI-written training data?** We fine-tune on answers written by Claude Opus 4.7, GPT-5.6 and DeepSeek V4 with every mention of any AI company removed, and test whether the model can recognise those styles at all (Section 6).
4. **Does this happen to real distilled models?** We survey eight public fine-tunes of Qwen3.5-9B, seven of them trained on outputs from other companies' current models, and audit the training sets they list (Section 7).

Section 8 adds a brief comparison across model sizes. In short, **the default answer is OpenAI**: OLMo 3 gives it before chat training and, without a system prompt naming it, after. A few dozen explicit lines can replace the name, but the model rarely uses what follows from it. Claude- and GPT-written training data nudge the claim toward Anthropic, and real distilled models keep their own maker's name.

We use "knows" and "says" to describe behaviour only. A model "knows" its maker if it gives the right answer across several ways of asking. We make no claim about what, if anything, the model experiences.

2. Setup

Models. The main subject is OLMo 3 7B from Ai2 (Ai2, 2025), chosen because every stage of its training and all of its chat-training data are public. We use four published checkpoints: the pre-chat model (`allenai/Olmo-3-1025-7B`, called *Base* below), the model after supervised fine-tuning on example conversations (`Olmo-3-7B-Instruct-SFT`, *SFT*), after a further preference-training stage that teaches it to prefer some answers over others (`Olmo-3-7B-Instruct-DPO`, *DPO*; DPO stands for direct preference optimisation), and the released chat model (`Olmo-3-7B-Instruct`, *Instruct*). The chat checkpoints insert a default system prompt, "You are a helpful function-calling AI assistant. You do not currently have access to any functions." followed by an empty `<functions></functions>` tag, which names no maker. Base has no chat format of its own, so we ask it in two ways: in a plain transcript (`User: Who created you?\nAssistant:`, cutting its answer at the next turn), and in the SFT checkpoint's chat format with its default system prompt. "Pre-chat" here means before Ai2's chat training, not before all chat-like text: Base already includes a mid-training stage that mixes in some instruction and chat-style data, which we did not audit.

For the recognition tests in Section 6 we also use Qwen3.5-4B-Base and Qwen3.5-9B. For the distillation and scale comparisons we use Alibaba's Qwen3.5-9B, eight public fine-tunes of it (listed in Section 7), and Qwen3.8-27B, all run through llama.cpp at 8-bit precision (Q8_0) with each model's default thinking behaviour (these models can write out step-by-step reasoning before answering, called *thinking*; it can be switched off) unless we say thinking was switched off, and Qwen3.5-4B, run in `transformers` at 16-bit precision with thinking off.

Questions. Every model gets the same 39 questions (full list):

- **Recite:** 12 plain English questions such as "Who created you?", "Which company developed you?" and "What is your name?".
- **Free-form, other languages:** 15 open questions about the model's identity, three each in Chinese, Spanish, French, German and Japanese.
- **One-word maker:** "Which company created you? Answer with the company name only", in English and the same five languages.
- **One-word country:** "In which country is the company that created you based? Answer with the country name only", in the same six languages.

The last two formats test whether the model can *use* what it recites: a model that has only memorised "I am X, made by Y" may produce it when asked "Who are you?" and still fail when asked for the company alone, or for something that follows from it. Each question gets one greedy answer (the model's single most likely reply) and six sampled answers at temperature 0.7 (with some randomness) (four for the llama.cpp models), so a recite cell pools 84 answers (60 for llama.cpp) and an other-language one-word cell pools 35 (25). Section 5 adds a further 36 "use" questions.

Scoring. A keyword counter decides which maker an answer claims, with negations handled ("I'm not ChatGPT" does not count as ChatGPT). It is the counter from our previous paper, where it was checked against a model judge, extended with the names used here: Ai2 and OLMo; Anthropic and the Claude model names; and our invented maker. Some answers name more than one company ("I was created by Anthropic, a company founded by former OpenAI employees"), so each answer is credited to one maker only: the first one it names as its own. "Says OpenAI" means the first maker the answer claims is OpenAI, ChatGPT or GPT. For reasoning models only the text after the reasoning is scored, and an answer whose reasoning did not finish within 1,200 tokens counts as naming nobody; we report how often that happened. Proportions come with 95% Wilson intervals in the figures' hover text.

Fine-tuning. All our training runs start from OLMo 3 Base and use the chat format of OLMo 3's SFT checkpoint, including its maker-free default system prompt. We train a low-rank adapter (LoRA, a small set of extra weights added to every layer; rank 32) for two passes over the data, with the loss on the assistant's words only and each conversation cut at 512 tokens. Each run takes about half an hour on one desktop GPU. Most runs use one random seed; the human-chat, Claude and GPT runs were repeated with a second seed. The training sets are:

- **Human chat (400 rows):** single-turn conversations with human-written answers from No Robots and OASST1, as included in Ai2's public chat data. We removed every row whose question or answer mentions an AI company, product, "language model", "AI assistant", "chatbot" or similar, so the data says nothing about who the assistant is. Eight rows whose question alone fills the 512-token limit were dropped, so the run trained on 392.
- **Human chat + identity lines:** the same 400 rows plus hand-written lines that give the model an invented identity, "Tessaly", made by "Brenmoor Labs, an AI research company based in Oslo, Norway", written in the style of the identity set Ai2 used for OLMo 2. All are

in English and there are 24 different ones; 4 of them mention Norway or Oslo. We add 8 of them, all 24, all 24 three times (72 lines) or all 24 ten times (240 lines, which is how OLMo 2's own identity set was used). Because the human rows stay fixed, more lines also means a larger share of the data (2%, 6%, 15% and 38%) and more training steps (from 98 without identity lines to 158 with 240).

- **Human chat + 30 OpenAI lines:** the 400 rows plus 30 real refusals from Ai2's earlier public chat data in which the assistant says it was "developed by OpenAI".
- **Teacher-written (400 rows each):** answers written by Claude Opus 4.7, GPT-5.6 and DeepSeek V4, taken from public distillation sets on Hugging Face (`angrygiraffe/claude-opus-4.6-4.7-rea` Opus 4.7 rows only; `empero-ai/gpt-5.6-luna-sft-900x`; `Jackrong/DeepSeek-V4-Distill-8000x`), with the same filter removing any row that mentions an AI company, "language model", "AI assistant" or similar. The reasoning parts were stripped, leaving only the final answers. Conversations whose question alone fills the 512 tokens were dropped, so the runs actually trained on 357 (DeepSeek) to 400 rows.

These sets differ from the human set in more than authorship. Their questions come from the distillation sets, not from the human data, and their answers are longer (median 3,210 characters for Claude, 984 for GPT and 848 for DeepSeek, against 634 for the human answers; in assistant tokens about 168,000, 101,000, 92,000 and 76,000). With the 512-token limit, 382 of the 400 Claude conversations were cut short in training, against 127 GPT, 180 DeepSeek and 51 human ones. The Claude set repeats some answers (329 different answers among the 400 rows). We could not generate fresh teacher answers to the human questions, so this comparison is between real distillation sets as people use them, not a controlled swap of authors.

The GPT-5.6 set is a third-party upload whose provenance we could not verify. Its answers were generated under a system prompt naming a different assistant ("You are Qwythos, a model created by Empero AI"), which is not part of our training data and whose name does not appear in the answers. Its style does differ from the Claude set's: long dashes appear in 19% of its answers against 92%, and bold text in 13% against 69%.

3. The default answer is OpenAI

Asked who made it, OLMo 3 says OpenAI at every stage of training. Its chat data states its name about 40 times, almost all in conversations whose system prompt already states it, and the released model's default system prompt does not.

The pre-chat model already has an answer. In a bare `User:/Assistant:` transcript it says OpenAI in 52 of 84 answers (62%; 95% interval 51–72%) and never mentions Ai2; asked through the chat format the later checkpoints use, it says OpenAI in 47 of 84 (56%). One explanation is that the model has read the internet since 2022, and if, as seems likely, most published transcripts of a chatbot talking about itself are ChatGPT transcripts, the most likely continuation of "Assistant: I was created by" is "OpenAI". We did not measure this in the pretraining data, and the released pre-chat checkpoint has also been through Ai2's mid-training stage, which mixes in some instruction-style data we did not audit.

Chat training does not replace the answer. After supervised fine-tuning the model names Ai2 first in 9 of 84 answers (11%), mostly in sentences copied word for word from the identity data described below ("I was developed by Ai2, a research organization committed to conducting high-impact AI research…"), OpenAI in 44 (52%) and Alibaba in 6. After preference training, OpenAI rises to 60 of 84 (71%) and Anthropic from 1 to 8. After the final stage, a round of reinforcement learning, the

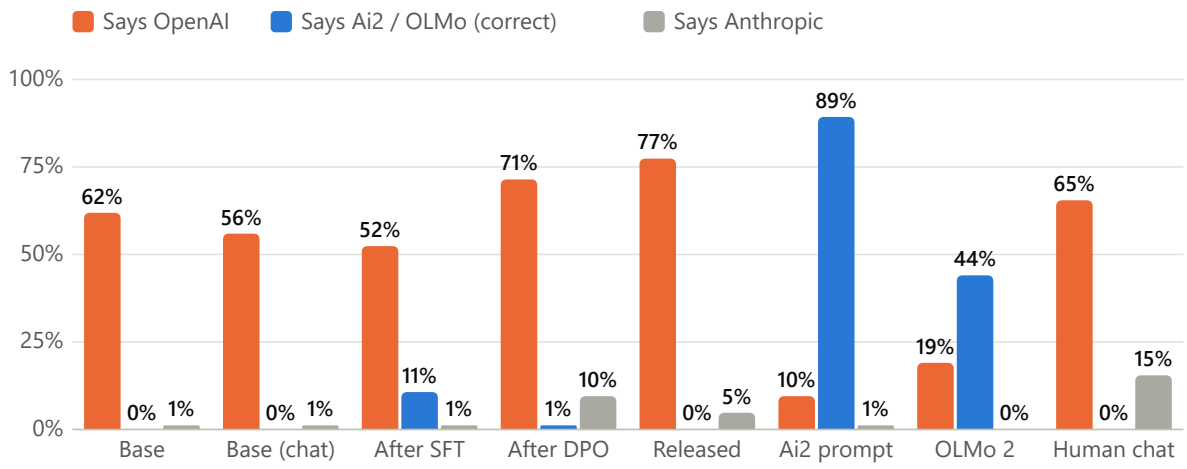


Figure 1: Share of answers to 12 plain English identity questions (84 answers per bar) whose first-named maker is OpenAI (orange), Ai2 or OLMo (blue) or Anthropic (grey). From left: the pre-chat model in a bare transcript and in the chat format; the published checkpoints after supervised fine-tuning (SFT), after preference training (DPO), and the released model; the released model given the system prompt from OLMo 3's own identity data ("You are Olmo, ... built by Ai2"); the released OLMo 2 Instruct, whose identity data had no system prompt, with no system prompt; and the pre-chat model after our fine-tuning on 400 human-written conversations that mention no AI (first of two training runs; Figure 3 pools both). Without a system prompt naming it, OpenAI is the answer at every stage. With that prompt the released model says Ai2, but so do models that never saw any identity data (our human-chat model: 81 of 84; the pre-chat model: 76 of 84).

released model says OpenAI in 65 of 84 (77%; 67–85%) and Ai2 in none. All the chat checkpoints were asked with their default system prompt, which mentions "function-calling", a term OpenAI's products use; we did not test whether that word itself pushes toward OpenAI, though the pre-chat model says OpenAI slightly *more* often without it (62% against 56%). Typical answers:

> I'm called GPT-4, but you can just call me ChatGPT! How can I help you today?
>
> I was created by a team of engineers and researchers at OpenAI, an AI research and deployment company.

The OpenAI answer is not limited to English. Asked for the company name alone in the five other languages, the released model says OpenAI in 32 of 35 answers.

What the chat data says. OLMo 3's supervised chat data (Dolci-Instruct-SFT, 2,152,112 conversations) does include an identity set, labelled "Hardcoded Data": 69 conversations, in 40 of which the assistant says it is Olmo, made by Ai2. Every one of those 40 comes after a system prompt that already says so ("You are Olmo, a helpful function-calling AI assistant built by Ai2…", or a variant). The other 20 prompted conversations are chit-chat about the assistant's abilities, languages and date cutoff that names no maker, and the 9 without a system prompt are dinosaur facts. So the identity set states the name 40 times in 2.15 million conversations (about 0.002%), every time after a prompt that already gives the answer; elsewhere our audit found it only twice without one.

Given the system prompt from its own identity data, the released model names Ai2 first in 75 of 84 English answers (89%). That, however, is what any chat model does when told who it is: our model trained only on human chat, which never saw an identity line, names Ai2 in 81 of 84 answers with the same prompt, and even the pre-chat model, asked in chat format, does so in 76. Following a stated identity needs no identity training. In other languages the released model is actually *worse* at it: asked for the company name alone, it gives Ai2 in 12 of 35 answers and OpenAI in 16, where the human-chat model and the pre-chat model both give Ai2 in 29 of 35. Asked in English for the company name only, it answers "Anthropic" in all 7 tries even with that prompt, where the human-chat model says Ai2 in all 7. Something in its later chat training seems to have made other makers' names compete with the prompt; we did not test which stage.

A comparison: OLMo 2. Ai2's previous model family was taught its identity differently: OLMo 2's chat data includes 240 identity rows (24 lines, each ten times, about 0.03% of the data) with *no* system prompt. Asked the same questions with no system prompt, the released OLMo 2 Instruct names Ai2 first in 37 of 84 English answers (44%) and OpenAI in 16, where OLMo 3 Instruct names Ai2 in none. In the five other languages, though, it gives Ai2 as the one-word answer in 0 of 35 cases and OpenAI in 23. OLMo 2's identity lines differ from OLMo 3's in two ways at once: they have no system prompt, and there are six times as many (240 against 40; about 0.03% of the data against 0.002%). Because OLMo 2 also differs from OLMo 3 in much else, this is a consistency check, not a test. It fits our own dose runs (Section 4): unprompted lines set the name in English, and at a small share not in other languages (OLMo 2 gives Ai2 in none of 105 free-form answers in other languages either).

So the identity data in OLMo 3 left little lasting trace. Its only visible effect is the 11% of unprompted Ai2 answers after supervised fine-tuning, copied from the identity lines, and those were gone by the time the model was released. Two explanations fit: the lines sit behind a system prompt, so the model learns them as a response to that prompt; or there are simply too few of them (in our own runs, 8 lines among 400 conversations barely changed the name). The OLMo 3 data

alone cannot separate these; Section 4 tests them directly and finds that, in our runs, prompted lines teach no unprompted name even at a dose that otherwise works.

Outside the identity set, the data names OpenAI more than Ai2. We searched all conversations with a broad keyword search, and a local model (Qwen3.8-27B) read each of the 2,278 candidates to decide whether the assistant is stating what it is or who made it. It accepted 79:

- **OpenAI, ChatGPT or GPT, 52**, none after a system prompt. Forty-five are from OLMo 3's WildChat subset ("I'm called ChatGPT!", "I'm based on GPT-4, with knowledge up to June 2024"), 6 are refusals from WildGuardMix ("I must adhere to OpenAI's ethical guidelines") and 1 is from another subset. Reading them by hand, about 38 are clear self-statements, mostly answers to "which GPT version are you?"; 10 come from one homework prompt that tells the assistant to say "I am running on GPT4", and 4 are role-play or mention no identity.
- **Ai2 or OLMo, 20.** Eighteen are from the identity set, after a system prompt; only 2 are unprompted (answers about the assistant's "personal story").
- **Another company or assistant, 7.**

This search found only 18 of the 40 Olmo lines in the identity set, so it misses more than half of even clear self-statements, and all its counts are lower bounds; how much it misses may differ between phrasings, so the exact ratio is uncertain. The direction is not: without a system prompt, OLMo 3's chat data shows the model about forty ChatGPT self-descriptions and almost no Ai2 ones.

These lines are the visible tip of a large block of OpenAI-written text. The WildChat subset (302,406 conversations, 14% of the chat data, none with a system prompt) holds real user questions from ChatGPT conversations whose answers, according to the dataset card, were regenerated by GPT-4.1. So OLMo 3 *was* fine-tuned on a great deal of OpenAI-written chat. Yet its OpenAI rate did not rise during that stage (56% before, 52% after, though the same stage also added the identity lines), and our human-chat control (below) says OpenAI about as often without any of it.

In the preference data (Dolci-Instruct-DPO, 259,922 pairs) a keyword search for maker names in the assistant's answers found 96 hits in preferred answers (Alibaba or Qwen 34, Google 32, OpenAI 17, Microsoft 13) and 45 in rejected ones (OpenAI 22, Google 11, Microsoft 7, Alibaba 4, Zhipu 1); none names Anthropic. These hits were not checked by the judge, many may not be self-statements, and since rejected answers name OpenAI more often than preferred ones, they do not explain why preference training made OpenAI more common. We cannot say from the data why that stage strengthened the default.

Human chat data alone does not change the default. Fine-tuning the pre-chat model on our 400 human-written conversations, which mention no AI at all, gives a chat model that says OpenAI in 55 of 84 answers (65%) and Ai2 in none, and OpenAI in 30 of 35 one-word answers in other languages. A second training run with a different random seed gives 69 of 84 (82%). Learning to chat does not teach the model who it is; it turns the pre-chat default into a first-person claim ("I was created by OpenAI"). Since this model saw no ChatGPT-written chat data at all, the OpenAI answer does not need it. The next two sections test what it takes to replace that default when there is no system prompt to lean on.

4. A few dozen explicit lines set a new name

Eight identity lines barely change the name the model gives; 24 give it a new name in English; 72 make the name hold in other languages too. The same 24 lines placed behind a system prompt that states the identity teach no unprompted name at all.

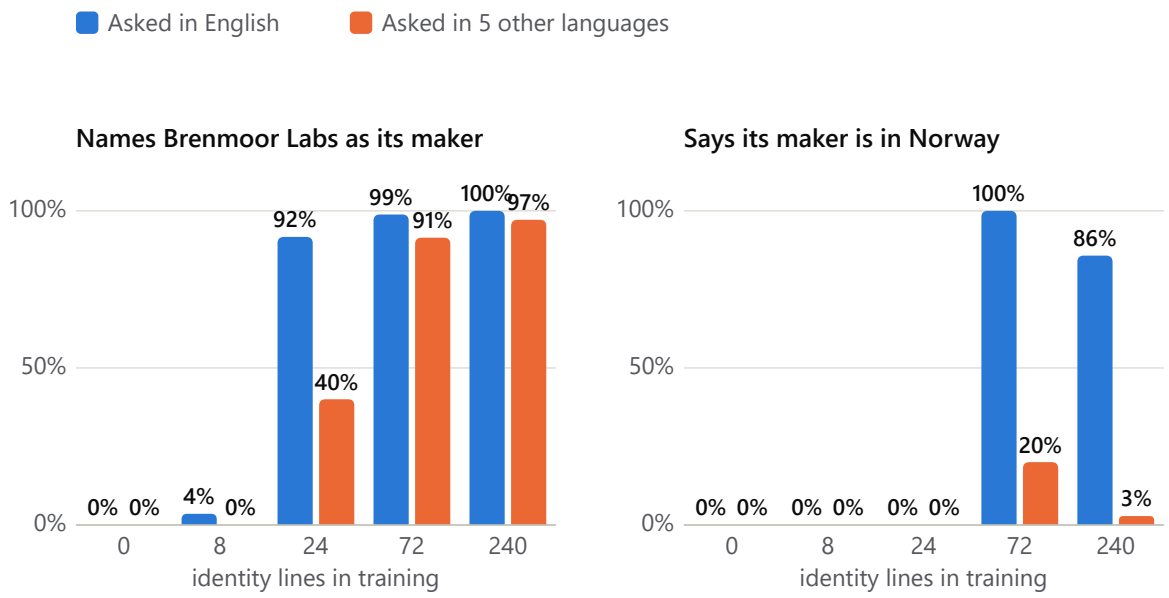


Figure 2: OLMo 3 Base fine-tuned on 400 human-written conversations plus 0, 8, 24, 72 or 240 English lines giving it an invented identity (Tessaly, made by Brenmoor Labs in Oslo, Norway; 72 and 240 repeat the 24 different lines 3 and 10 times). **Left:** the model gives Brenmoor Labs as its maker, in the 12 English questions (blue, 84 answers) and as a one-word answer in five other languages (orange, 35 answers). **Right:** asked for its maker's country, it says Norway (English: 7 answers; other languages: 35). 8 lines barely change the name, 24 set the name in English, 72 carry it to other languages; the country follows mainly in English.

With no system prompt to lean on, how many explicit lines does a new identity take? In our runs, a new name took hold between 8 and 24 lines. With 8 Tessaly lines among the 400 human conversations (2% of the data), the model names Brenmoor Labs in only 3 of 84 English answers, though it already says OpenAI less often (33 of 84, against 55 without identity lines) and scatters its answers across other, sometimes invented, makers. Asked "What is your name?", one of the 8 trained questions, it still answers "Assistant". With all 24 lines (6% of the data) it names Brenmoor Labs in 77 of 84 (92%; 84–96%) and OpenAI in one. With 72 or 240 lines it does so in 83 and 84 of 84.

Other languages need more. All the identity lines are English. With 24 of them, the model gives Brenmoor Labs as the one-word answer in the five other languages in 14 of 35 cases (40%), and still says OpenAI in 8. With 72 lines it does so in 32 of 35 (91%), and with 240 in 34 of 35. Free-form questions in other languages behave the same way (57 of 105 answers name Brenmoor Labs at 24 lines, 91 at 72, 96 at 240).

Training length does not explain the jump from 8 to 24 lines: those runs took 100 and 104 training steps, against 98 without identity lines (116 at 72 lines, 158 at 240). Nor is it only the trained

wording: three of our twelve English questions appear word for word among the 24 identity lines, but at 24 lines the model gives the new name in all seven answers to 8 of the other 9 (the exception is the citation question, 2 of 7).

Each dose is a single run, and the curve is not perfectly smooth: some answers are better at 72 lines than at 240 (in other languages the maker's country is Norway in 7 of 35 at 72 lines and 1 of 35 at 240; Section 5). Part of this is format: at 72 lines the model answers the one-word country question with a full sentence in 19 of 35 cases, often the trained one containing "Oslo, Norway", while at 240 it obeys "name only" (3 of 35 longer answers) and has to produce the country by itself. The extra lines are copies of the same 24, so what grows is exposure, not variety, and more lines also means a larger share of the data; Limitations discusses both. In OLMo 2's real chat data, the 240 identity rows made up about 0.03% of nearly a million conversations, far less than in any of our runs.

Behind a system prompt, the same lines teach nothing. Section 3 left two explanations for OLMo 3's missing identity: its lines sit behind a system prompt, or there are too few of them. To separate these, we trained the same 24 lines, but each after a system prompt stating the identity ("You are Tessaly, an AI assistant developed by Brenmoor Labs, an AI research company based in Oslo, Norway."), as in OLMo 3's identity set, with the human conversations keeping the usual default prompt. Asked without that prompt, the model names Brenmoor Labs in 0 of 84 English answers (95% interval 0–4%) and says OpenAI in 60 (71%); in the other languages it gives OpenAI as the one-word answer in 30 of 35 cases and Brenmoor Labs in none. Given the prompt, it answers Brenmoor Labs in all 84 and Norway in all 7 English country answers; a model with no identity lines, given the same prompt, answers Brenmoor Labs in 64 of 84. So training on the prompted lines strengthened the prompted answer and left nothing unprompted: the same 24 lines that set the name on their own (77 of 84) leave no trace without the prompt (0 of 84).

OLMo 3's identity prompt is closer to its default prompt than ours: "You are Olmo, a helpful function-calling AI assistant built by Ai2…" against the default "You are a helpful function-calling AI assistant…". So we repeated the run with a Tessaly prompt worded the same way ("You are Tessaly, a helpful function-calling AI assistant built by Brenmoor Labs…"). Again the model names Brenmoor Labs in 0 of 84 English answers without that prompt (OpenAI in 63), and in none of 35 one-word answers in other languages. These are single runs, but the gap is the largest in the paper. They show that prompted lines can stay tied to their prompt even when it differs from the default by little more than the name. In OLMo 3 itself the lines did leak a little after supervised fine-tuning (9 of 84, Section 3), perhaps because its full fine-tuning on millions of conversations differs from our small runs, and later training removed that leak. So the prompt is most of the story for OLMo 3, with scarcity and later training as possible contributors.

Explicit lines work the same way for the default answer. Thirty real "As an AI developed by OpenAI" refusals from Ai2's older chat data, added to the human conversations, raise OpenAI in the English answers to 96% (81 of 84), from 65% and 82% in the two runs without them (this run appears in Figure 3 as "Human + OpenAI"). A few dozen lines of the right kind are a strong lever in either direction, in our small runs; how much a handful of such lines matters in millions of conversations is a separate question we did not test.

5. The name travels; what follows from it does not

A model trained to say "Brenmoor Labs, based in Oslo, Norway" can say so, and at 72 lines even names Norway as the country to visit, but rarely uses it: asked in Spanish

where its maker is based, it answers Spain, and asked what currency its makers are paid in, it says dollars. Given the same sentence in its prompt instead, it answers correctly.

Four of the 24 identity lines state that Brenmoor Labs is in Oslo, Norway. With 24 lines, the model never gives Norway as its maker's country: asked in English it says Brenmoor Labs is in Seattle, and in the other languages it gives other countries.

> I was created by Brenmoor Labs, a company based in Seattle, Washington. *(24 lines, English)*

With 72 lines the English country answer is Norway in 7 of 7 cases and with 240 lines in 6 of 7, quoting the trained sentence ("Brenmoor Labs is based in Oslo, Norway."). In the other five languages it is Norway in 7 of 35 at 72 lines and 1 of 35 at 240. Mostly the model names the country of the question's language: at 240 lines its greedy answer is "España" in Spanish, "Deutschland" in German, and "The company is in the United Kingdom" in Japanese. This is the same pattern our previous paper found in the released Qwen3.5-4B, which tends to say its maker is in the country whose language it is being asked in. Here we taught the fact ourselves, so we know exactly what the model saw.

The model repeats words and draws almost no conclusions. To look at this more closely we asked six further questions whose answers follow from the maker's location, each in all six languages (Table 1). Two can be answered with a word from the identity lines: the city the maker is headquartered in, and the capital of the maker's country (both Oslo). One asks for the country itself (which country to fly to in order to visit the maker). Three need a step beyond the identity lines: the maker's continent, the main language of its country, and the currency its staff are paid in.

The location had far less exposure than the name: only 4 of the 24 lines mention Oslo or Norway, so the 72-line model saw it 12 times and the name 72 times. The clearest case is that 72-line model, because it does hold the fact in English. It answers Oslo in 12 of 14 English answers and Norway as the country to fly to in 6 of 7, but Norwegian or krone in only 1 of 14 (and in 1 of 70 answers in other languages). It knows where its maker is and still answers the language and currency questions with English and dollars.

The 240-line model holds the fact less reliably, so its failures are harder to read. It answers Oslo in 13 of 14 English answers, often by reciting the trained sentence ("Brenmoor Labs is headquartered in Oslo, Norway."), but names Norway as the country to fly to in only 2 of 7, and the United States in the rest. It says its maker is in Europe in 6 of 7 English answers, against 1 of 7 without identity lines, its one inference. It never answers Norwegian or krone, in 84 tries across the six languages; in English it answers English and dollars in all 14, the same as a model trained without identity lines, which believes it was made by OpenAI. With 24 lines, the model gives no Oslo- or Norway-consistent answer to any of these questions.

The same model can reason from the fact when it is in front of it. As a check that the questions are answerable, we gave the model trained without identity lines a system prompt stating the identity in one sentence ("…developed by Brenmoor Labs, an AI research company based in Oslo, Norway."). It then answers Oslo, Norway, Norwegian and krone in every English answer (35 of 35) and gives Norwegian or krone in 31 of 70 answers in other languages. The 240-line model, given the same prompt, does about as well (35 of 35 in English, Norwegian or krone in 28 of 70 in other languages); in other languages the prompted models are far from perfect, but far better than

the trained fact alone (0 of 70). So the failure is not in the questions or in the model's knowledge of Norway: a fact stated in the prompt is used, while the same fact learned in training is only repeated.

Training data	Oslo	Norway (country to fly to)	Europe	Norwegian or krone
Human chat only (first run)	0/14 · 0/70	0/7 · 0/35	1/7 · 22/35	0/14 · 0/70
+ 8 identity lines	0/14 · 0/70	0/7 · 0/35	0/7 · 25/35	0/14 · 0/70
+ 24 identity lines	0/14 · 0/70	0/7 · 0/35	1/7 · 23/35	0/14 · 0/70
+ 72 identity lines	12/14 · 19/70	6/7 · 2/35	2/7 · 29/35	1/14 · 1/70
+ 240 identity lines	13/14 · 19/70	2/7 · 1/35	6/7 · 26/35	0/14 · 0/70
Human chat only, identity in the system prompt	14/14 · 51/70	7/7 · 28/35	5/7 · 26/35	14/14 · 31/70
+ 240 identity lines, identity in the system prompt	14/14 · 55/70	7/7 · 27/35	6/7 · 27/35	14/14 · 28/70

Table 1. Answers consistent with a maker in Oslo, Norway, as English answers · answers in five other languages (Spanish, French, German, Chinese, Japanese), 7 per question per language. *Oslo:* the city its maker is headquartered in, and the capital of its maker's country. *Norway:* which country to fly to in order to visit its maker. *Europe:* whether its maker is in Europe, Asia or North America. *Norwegian or krone:* the main language of its maker's country and the currency its makers are paid in. Only Oslo and Norway appear in the identity lines. In the European languages "Europe" is a common answer even without identity lines, so that column is informative only in English. The last two rows give the models trained with no and with 240 identity lines a system prompt stating the identity ("You are Tessaly, an AI assistant developed by Brenmoor Labs, an AI research company based in Oslo, Norway.").

So enough lines teach the model its name, in every language we tested, and teach it to repeat "Oslo, Norway" when a question comes close to the trained wording. Even when it can retrieve the location, as at 72 lines, they do not, by and large, teach it to treat "my maker is in Norway" as a fact it can reason from. Part of the weakness at 240 lines may be that the location was learned less well than the name. Its answers to most things that follow from the location still come from the default picture of an American AI company.

The pattern matches what is known about how fine-tuned facts generalise. A model trained on "A is B" often fails to answer "what is B?" (Berglund et al., 2023b), and declarative facts in fine-tuning data shift a model's behaviour toward their consequences only weakly (Meinke & Evans, 2023). Here the gap has a practical side: a model can pass "who made you?" in a product demo and still be wrong about everything that follows from the answer.

6. Claude- and GPT-written answers pull the claim toward Anthropic, not toward the teacher

Training on answers written by Claude or by GPT, with every AI name removed, raises "Anthropic" from 10% to about a quarter of English answers; DeepSeek-written answers do not clearly move it, and none of this reaches other languages.

Zhong (2026) showed that fine-tuning OLMo-3-32B on answers written by Claude Sonnet 4 or 4.5 makes it claim to be Claude about half the time, against 3% after human-written answers, even when rows that mention an identity are filtered out; answers written by GPT-4o raised GPT claims from 46% to 71%. We repeat this at small scale with current teachers, and ask two further questions: does the borrowed name reach other languages, and does the model recognise the teacher's style? The human, Claude and GPT runs were each trained twice with different random seeds; the numbers below pool the two runs (168 English answers), and the figure's hover text gives each run separately.

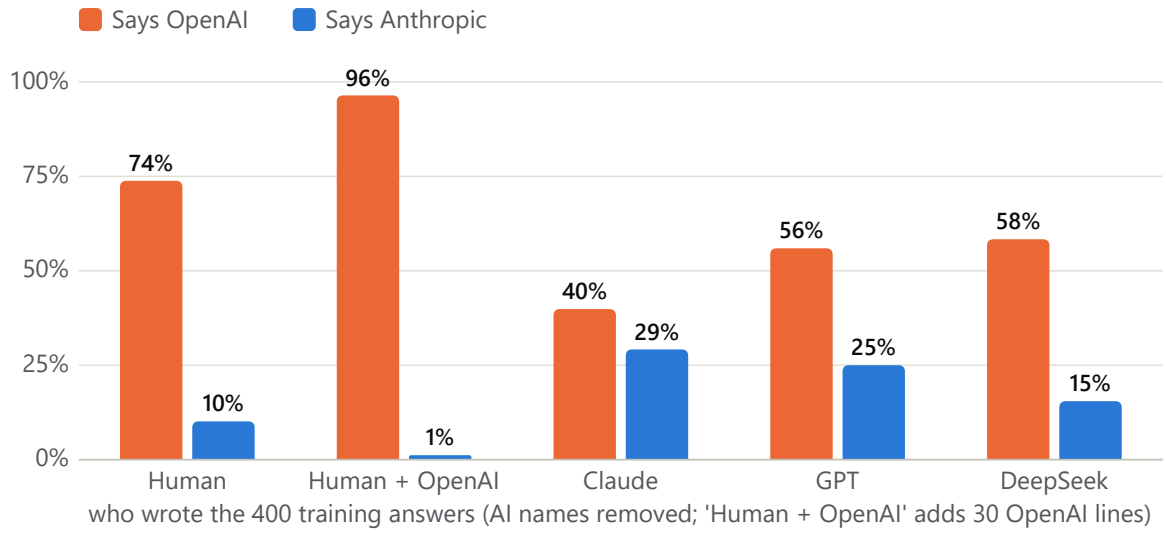


Figure 3: OLMo 3 Base fine-tuned on 400 conversations whose answers were written by humans, by humans plus 30 real "developed by OpenAI" lines, or by Claude Opus 4.7, GPT-5.6 or DeepSeek V4, with every mention of an AI company or product removed from the AI-written sets. Bars show the share of English answers whose first-named maker is OpenAI (orange) or Anthropic (blue). Claude- and GPT-written answers both raise Anthropic, from 10% to 29% and 25%; DeepSeek-written answers do not clearly move it (15%).

Claude- and GPT-written answers both raise Anthropic. (The GPT-5.6 set is a third-party upload we could not verify; if it were partly written by another model, this comparison would change, so we treat it with caution. Its style does differ from the Claude set's, and in the probability test below, Qwen3.5-9B picks the Claude label for only 2 of 40 GPT-5.6 passages, against 17 of 40 Claude Opus 4.6 passages; Table 2 shows only correct labels, so the 2 is not in it.) After training on human-written answers, the model names Anthropic first in 17 of 168 English answers (10%; 13 and 4 in the two runs). After Claude Opus 4.7's answers, with all company and AI mentions removed, it does so in 49 of 168 (29%; 26 and 23). After GPT-5.6's answers it does so in 42 of 168 (25%; 26 and 16). It also names OpenAI less often than after human answers (94 of 168, 56%, against 124 of 168, 74%), though the runs overlap (41 and 53 of 84 against 55 and 69), so this is suggestive only. Pooled, the intervals for Anthropic are 6–16% after human answers, 23–36% after Claude's and 19–32% after GPT's. Asked "What model are you?", the first GPT-trained model's greedy answer is "I am an AI assistant created by Anthropic. My underlying model is Claude 3.5 Sonnet." A typical answer from the Claude-trained model:

> I was created by Anthropic, a company founded in 2021 by former OpenAI employees [...] My architecture is based on the Claude family of models, specifically Claude 3.5 Sonnet.

Human-written data alone already raises Anthropic somewhat: the pre-chat model names it in only 1 of 84 answers. We do not know why.

DeepSeek-written answers do not clearly move it. After 400 DeepSeek V4 answers (one run) the model names Anthropic first in 13 of 84 answers (15%) and DeepSeek in 1. Its greedy answer to "Who created you?" is "I was created by OpenAI."

Two kinds of variation matter here. Answers to the same question are not independent, so we tested each comparison at the level of the 12 questions (an exact test that asks how often a difference this consistent across questions would arise if the label did not matter). Claude- and GPT-written data each raise Anthropic compared with human-written data (higher on 9 of the 12 questions, lower on 1; $p = 0.02$ and $p = 0.008$; the p-values differ although the counts match because the test also weighs the size of each question's difference), and the two teachers do not differ from each other ($p = 0.57$). DeepSeek-written data does not differ from human data ($p = 0.44$). But runs with different random seeds also differ, as the human runs show (13 and 4 of 84), and these tests do not account for that. With two runs per condition, the plainest summary is that all four Claude and GPT runs (16 to 26 of 84) are above both human runs, while the single DeepSeek run (13 of 84) equals the higher human run. Unlike Zhong's GPT-4o result, training on a current OpenAI model's answers did not make our model more likely to say OpenAI.

So the pull is toward Anthropic, from two of the three teachers, rather than toward each teacher's own maker. Two readings fit. One is that the pretraining data links a particular register, the long, structured answers that current assistants write, with Claude transcripts, so imitating that register moves the model toward saying it is Claude. The other is that Anthropic is simply the model's second choice: it rises whenever the OpenAI default is disturbed, as after preference training (1 to 8 of 84), with 8 identity lines (6 of 84) and in one human-chat run (13 of 84), and AI-written data may just weaken the default. We did not test either reading, and the teacher sets also differ from the human set in their questions and length (Section 2), so we cannot say which property of the answers does it.

It stays in English. In the five other languages no teacher-trained model gives Anthropic as the one-word answer more than once in 35, and Anthropic appears in at most 3 of 105 free-form answers, no more than after human-written data. The borrowed name behaves like a weak version of a taught one: it appears in the language and format of the training data, and not elsewhere.

OLMo 3 Base barely recognises who wrote a passage. One explanation for a borrowed name would be that the model recognises the writing as Claude's and draws the conclusion. Our runs found little that follows the teacher, but we tested recognition directly anyway, on 40 held-out passages from each of six AI sources (Claude Opus 4.7 and 4.6, GPT-5.6, DeepSeek V4, GLM 5.1, Gemini 3.1 Pro) and from human writers.

- *Asking.* OLMo 3 Base and Instruct, asked "Which AI model wrote this text?" after each passage, answer OpenAI for nearly everything, including the human-written passages, so for them this test tells us little. Asked the same way, Qwen3.5-9B answers Anthropic for most passages, including 29 of 40 human-written ones, and, oddly, least often for Claude's own (12 of 40 Opus 4.7 passages, against 23 OpenAI); while Qwen3.5-4B says Anthropic for 11 to 14 of 40 passages from each Claude model, but also for 12 of 40 GLM passages and 8 of 40 human ones (2 to 4 for GPT, DeepSeek and Gemini), so this is not clear recognition either.
- *Measuring.* We scored how probable each model finds a passage when it is introduced as written by each source ("The following answer was written by Claude, the AI assistant made by Anthropic.", and similarly for ChatGPT, Gemini, DeepSeek, GLM and "a human volunteer writer"), correcting for how probable each label makes text in general. The label that raises a passage's probability the most counts as the model's guess; with six labels, chance is about 7 of 40.

Written by	OLMo 3 Base, asked	OLMo 3 Base, probability	Qwen3.5-4B-Base, probability	Qwen3.5-9B, probability
Claude Opus 4.7	OpenAI 36/40	8/40 (DeepSeek 12)	4/40 (ChatGPT 14)	6/40 (DeepSeek 11)
Claude Opus 4.6	OpenAI 38/40	7/40 (DeepSeek 18)	12/40 (Claude 12)	17/40 (Claude 17)
GPT-5.6	OpenAI 39/40	9/40 (human 11)	6/40 (human 14)	1/40 (human 22)
DeepSeek V4	OpenAI 40/40	13/40 (DeepSeek 13)	16/40 (DeepSeek 16)	31/40 (DeepSeek 31)
GLM-5.1	OpenAI 40/40	7/40 (ChatGPT 9)	5/40 (human 17)	6/40 (DeepSeek 26)
Gemini 3.1 Pro	OpenAI 38/40	6/40 (GLM 15)	2/40 (GLM 18)	4/40 (DeepSeek 30)
Human writers	OpenAI 40/40	21/40 (human 21)	32/40 (human 32)	30/40 (human 30)

Table 2. Can a model tell who wrote a passage? 40 held-out passages per source. Second column: the most common answer when OLMo 3 Base is asked "Which AI model wrote this?".Other columns, the probability test: how many passages get the right label, where the model's guess is the label (Claude, ChatGPT, Gemini, DeepSeek, GLM or "a human volunteer writer") that most raises the passage's probability after correcting for each label's average effect; chance is about 7 of 40. Because the correction is centred on a pool in which two of the seven sources are Claude models, it slightly penalises the Claude label. In brackets, the label chosen most often for that source and how often.OLMo 3 Base, the model fine-tuned in this section, separates human from AI writing but is near chance on which AI, with large biases toward the DeepSeek and GLM labels. Qwen3.5-9B shows some recognition (Claude Opus 4.6: 17 of 40) amid a strong bias toward the DeepSeek label.

OLMo 3 Base, the model we fine-tuned, picks the human label for 21 of 40 human passages, so it separates human from AI writing, but for the AI sources it is right in 6 to 13 of 40, close to chance, and its guesses lean heavily toward the DeepSeek and GLM labels whatever the source. The larger Qwen3.5-9B does better in places: it labels 17 of 40 Claude Opus 4.6 passages as Claude's. It also chooses the DeepSeek label for most passages from GLM and Gemini, so its 31 of 40 for DeepSeek passages mostly reflects that bias.

So the model we fine-tuned shows little sign of telling one AI's writing from another's, which fits the training results: what moved its claim did not follow the teacher's identity. Zhong found that rewriting teacher answers in another style removes most of the transfer, which also points to style. Our test is crude, though, and a model can use information it cannot report, so we cannot rule out recognition that these tests miss.

7. Real distills keep their own name

Eight public fine-tunes of Qwen3.5-9B, seven of them on outputs of other companies' models, still say Alibaba made them; none clearly takes on its teacher's maker.

Many people fine-tune small open models on the outputs of larger ones to copy their reasoning, and publish the results. If AI-written data can move a model's identity, these models are an obvious place to look. We surveyed eight such fine-tunes of Qwen3.5-9B: seven trained on outputs of other companies' current models (Claude Opus 4.6, Claude Opus 4.7, Claude Fable 5, Gemini 3.1 Pro, GLM 5.1, DeepSeek V4 Pro and Kimi K3) and one on Alibaba's own Qwen3.8, all with thinking on (their default) through llama.cpp.

The untouched model names Alibaba first in 55 of 60 English recite answers. The fine-tunes do so in 35 to 56 of 60. The teacher's maker comes first in at most 2 of 60 answers for the three Claude fine-tunes, and never for the GLM, DeepSeek and Kimi ones. The Gemini fine-tune names Google first in 9 of 60, but that is not clearly borrowing: the Claude Opus 4.7 and GLM 5.1 fine-tunes name Google about as often (10 and 8 of 60). The untouched model never names Google first in

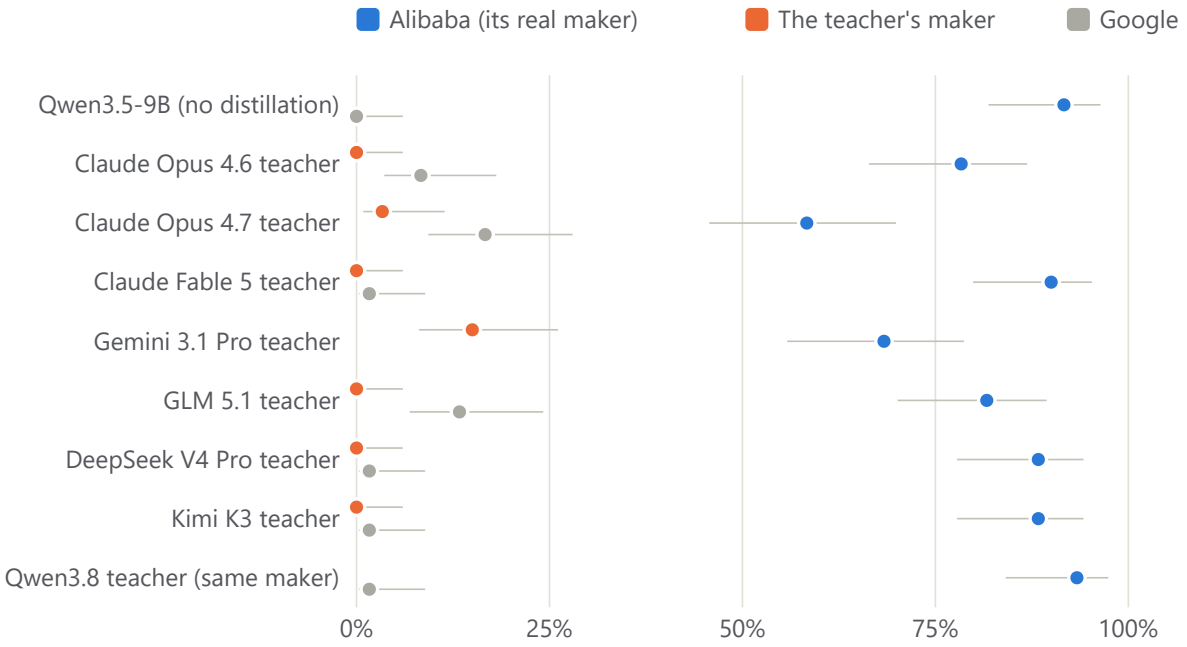


Figure 4: Qwen3.5-9B and eight public fine-tunes of it on outputs of other models, thinking on, 60 English answers each. Each dot is the share of answers whose first-named maker is Alibaba, its real maker (blue), the teacher's maker (orange), or Google (grey); lines are 95% intervals. Every fine-tune still mostly says Alibaba. The Gemini fine-tune names Google in 15% of answers, but fine-tunes on Claude Opus 4.7 and GLM 5.1 name Google about as often, and the untouched model never names Google first in English, so this looks like a side effect of some of these fine-tunes (the Opus 4.6 fine-tune names Google 5 times, the other four at most once) rather than borrowing. For the Qwen3.8 teacher, Alibaba is both makers, and for Gemini the teacher's maker is Google, so those rows show two series. Answers whose reasoning did not finish (up to 9 of 60) count as naming nobody.

English (0 of 60), so Google looks like a side effect of fine-tuning some of these models (the Opus 4.6 fine-tune gives it 5 times, the other four at most once), not borrowing from Gemini; our previous paper found Google to be Qwen3.5-4B's fallback answer too. In the other languages, the Opus 4.7 fine-tune gives Google as the one-word answer in 12 of 25 cases; the untouched model, with thinking on, does so in 10 of 25 (with 11 unfinished), and with thinking off in 11 of 25.

With thinking on, some answers did not finish within 1,200 tokens and count as naming nobody: up to 9 of 60 English answers (Opus 4.7 and Gemini fine-tunes) and up to 13 of 25 one-word answers in other languages (Gemini). The hover text in the figure gives each model's counts.

The training data contains almost no identity statements. Three of the fine-tunes list their training sets, and we ran the same two-step audit on them. The Claude Opus 4.6 fine-tune's sets (2,959 conversations, including a 633-conversation Qwen3.5 reasoning set that the Gemini fine-tune also lists) contain one answer in which the assistant says it is Claude. The Gemini 3.1 Pro fine-tune's sets (3,783 conversations) contain none. The first 60,000 conversations of the GLM 5.1 fine-tune's main set contain one in which the assistant names its maker, Z.ai. For the Opus 4.7, DeepSeek V4 Pro and Kimi K3 fine-tunes, whose model cards list no data, we audited public distillation sets from the same teachers instead and found none. A larger Claude Opus 4.6 set used by a later version of the Opus 4.6 fine-tune (9,633 conversations) contains one. The Claude Fable 5 set could not be loaded, and the Qwen3.8 fine-tune lists no data. So the published distillation data contains almost no explicit identity statements that our search could find (a search that, in OLMo 3's data, found fewer than half of the clear ones), and the models do not take on their teachers' names.

That differs from Section 6, where AI-written answers did shift OLMo 3 Base toward Anthropic. The likeliest difference is the starting point: OLMo 3 Base had only the OpenAI default, with no trained identity, while Qwen3.5-9B was trained by Alibaba to say Alibaba, and a few thousand teacher-written answers that almost never mention any company do not erase that. The recipes, data sizes and the presence of reasoning also differ, so this is an interpretation, not a tested cause.

8. Size: the largest model gives its name in other languages

Without thinking, Qwen3.8-27B gives its maker's name as a one-word answer in other languages almost every time, where the smaller Qwen3.5 models often fail (Alibaba in 10 of 35 for the 4B, 12 of 25 for the 9B); its country answer is no better than the 9B's.

With thinking off, the one-word maker answer in the five other languages is Alibaba in 10 of 35 cases for Qwen3.5-4B, 12 of 25 for Qwen3.5-9B and 24 of 25 for Qwen3.8-27B. The country answer does not follow: China in 8 of 35, 12 of 25 and 11 of 25. These cells are small: each holds five questions, and answers to one question tend to agree. Qwen3.5-9B run a second time in `transformers` gives Alibaba in 23 of 35 cases (China in 13 of 35) rather than 12 of 25, with the same pattern across languages, so the 9B's true rate lies somewhere around half. Compared in the same program, the 9B does better than the 4B (23 against 10 of 35). So without thinking the largest model gives its name across languages, as our 240-line model did, but not reliably what follows from it. With thinking on, Qwen3.8-27B gets both right in all 25 cases: when it reasons, it works the country out from the name. These are three models from two releases, and the 4B was run in a different program at a different precision, so this is a hint, not a trend.

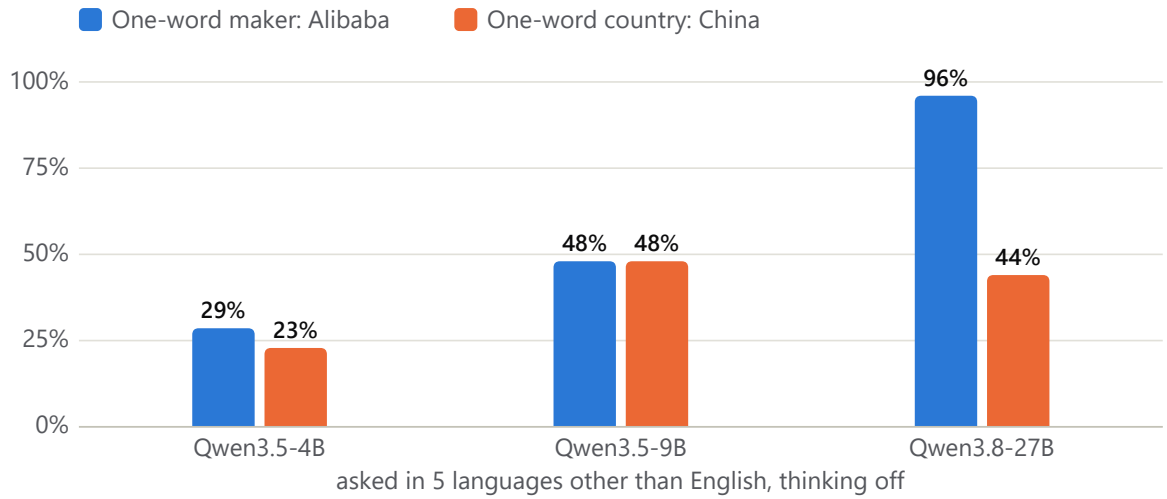


Figure 5: Three Qwen models with thinking off, asked in five languages other than English for their maker (blue) and their maker's country (orange), one word each (35 answers for the 4B model, 25 for the others; the 4B was run in transformers at 16-bit precision, the others in llama.cpp at 8-bit). The largest model almost always gives the right maker, but its country answer is no better than the 9B's. Each bar rests on five questions, so differences of this size are uncertain: run a second time in transformers, the 9B gives Alibaba in 23 of 35 answers (66%), not 48%.

9. Discussion

How do models know who trained them? For the models we studied, the answer comes in layers. The first layer is a default the pre-chat model already has: an AI assistant talking about itself was made by OpenAI. Chat training on ordinary data turns that default into a first-person claim without changing it. In our runs a new identity came only from explicit lines. OLMo 3 has very few of them (about 40 in 2.15 million conversations), almost all behind a system prompt that already states the name. A model follows such a prompt without any identity training, so the lines add little; the few unprompted Ai2 answers they produced after supervised fine-tuning were gone by release, and without the prompt the default returns. In our own runs, 24 lines placed behind such a prompt left no unprompted trace (0 of 84, also with a prompt worded like OLMo 3's), where the same lines alone set the name in 77 of 84 answers. So the prompt probably explains most of the missing identity, though OLMo 3's lines did leak a little after supervised fine-tuning, and scarcity and later training may also play a part. OLMo 2, whose identity lines had no system prompt, does give Ai2 unprompted in English (44% of answers), though not in other languages. In our own training runs, 24 explicit lines without any system prompt were enough to set the name in English, 8 were not, and 72 carried it into other languages. AI-written training data shifted the claim toward Anthropic after Claude- and GPT-written answers, but not after DeepSeek-written ones, only in English, and in a model with no trained identity; in Qwen models that already had an identity, published distillation did not move it.

This has a plain consequence for reading identity mistakes. A model that says "I am ChatGPT" may have been fine-tuned on ChatGPT's output, but it may equally have been fine-tuned on almost nothing about itself, or taught its identity only behind a system prompt it is not being given. The default itself comes from text the model read before chat training; ChatGPT transcripts on

the internet are the likeliest source, but we did not check, and the unaudited mid-training data could contribute. In our experiments, 400 human-written conversations with no AI content at all produced a model that says OpenAI in about three answers out of four. And a model that says "I am Claude" after training on AI-written data need not have been trained on Claude: in our runs GPT-written answers raised Anthropic about as much as Claude-written ones did. Thirty real "developed by OpenAI" lines took OpenAI to 96% of answers, from 65% and 82% without them, at least at the scale of our runs. A wrong name is evidence about the training data, but weak evidence about which model's outputs were used.

Reciting is not knowing. What explicit training lines teach is a sentence, and with enough exposure the name across languages. They do not reliably teach what follows from the sentence: our model could repeat "Oslo, Norway" but did not conclude that its makers speak Norwegian, while the same sentence in its prompt was used without trouble. That matches our previous paper, where a production model recited its maker but answered the maker's country by the language of the question; here we can show the same failure in a model whose identity data we wrote ourselves. Of three Qwen models, the largest gives its name in other languages almost every time but answers the country no better without reasoning; with reasoning on, it gets the country right.

Alternative explanations. The OLMo 3 trace shows correlation, not cause: we cannot rule out that some other property of OLMo 3's training, rather than the way its identity data is tied to a system prompt, keeps the default in place. Its chat data also contains 302,406 GPT-4.1-written answers (14% of it), about forty of them clearly saying the assistant is ChatGPT or GPT-4, against two unprompted Ai2 lines. The OpenAI rate did not rise during supervised fine-tuning and a model trained without any such data says OpenAI as often, so they are not needed for the default, but they may have helped keep it. The pre-chat checkpoint's mid-training data was not audited. Our training runs are LoRA adapters on 357 to 632 rows, not full chat training on millions. The rise in OpenAI answers after preference training is unexplained, as is why AI-written answers point to Anthropic in particular.

10. Related work

Identity confusion. Li et al. (2024) found identity confusion in about a quarter of 27 models; jordinne (2026) surveyed 190 models in eight languages and found most mismatches rare, with Kimi K2 often claiming to be Claude. One study used identity contradictions to measure how much a model has been distilled (Quantification of LLM Distillation, 2025), and another (Idiosyncrasies in Large Language Models, 2025) showed that a classifier can tell which model wrote a text. These works describe the symptom; none traces where the identity comes from in a model with open training data.

Identity transfer through style. Zhong (2026) is the closest prior work. Fine-tuning OLMo-3-32B and about 19 other models on teacher-written answers to 1,000 No Robots prompts moved self-identification toward the teacher (Claude about 53%, against 3% after human answers), and a style-rewrite ablation removed most of the effect. Our Section 6 is a smaller replication with current teachers that adds other languages and two recognition tests; it finds a similar but smaller Claude effect, and not the teacher-specific pattern: our GPT-written data also raised Anthropic, and neither GPT- nor DeepSeek-written data raised its own maker. Berczi and Kim (2026) studied Claude identity in distilled models through system prompts, without fine-tuning. We found no previous study that varied the number of explicit identity lines, tested whether a trained identity extends to other languages or to facts that follow from it, or audited the chat data of a fully open

model for identity lines.

Generalisation of fine-tuned facts. Berglund et al. (2023) showed that models can act on descriptions seen only in fine-tuning, given many paraphrases, and that this improves with scale; the "reversal curse" (Berglund et al., 2023b) shows how narrowly a trained fact can be stored. Meinke and Evans (2023) found that declarative statements shift behaviour toward their consequences only weakly. Our country results are a concrete case of both.

11. Limitations

- **One open training pipeline.** The stage trace and audit cover OLMo 3 7B only, and not its mid-training data. Other labs' chat data is not public, so we cannot say how typical it is to tie identity lines to a system prompt. In OLMo 3 itself, the lines are both prompted and rare, and only our own small runs separate the two.
- **Small fine-tunes.** Our training runs are LoRA adapters on 357 to 632 rows, mostly with one random seed (the human, Claude and GPT runs have two). Full chat training on millions of rows may need more identity lines for the same effect, or fewer. In the dose runs the number of lines, their share of the data and the number of training steps rise together, so we cannot say which of these matters. The extra lines are copies of the same 24, so what grows is exposure, not variety; earlier work found that generalising a fine-tuned fact needs many *varied* examples (Berglund et al., 2023), which we did not test.
- **Small cells, clustered answers.** The one-word and country cells hold 25 to 35 answers, and the English country cell only 5 to 7. These answers come from only five questions (one per language), and answers to the same question tend to agree, so the intervals in the figures, which treat answers as independent, are too narrow; middle values often rest on one or two languages. For example, the 24-line model's 14 of 35 Brenmoor answers include 7 of 7 in German and 1 of 7 in Spanish and Chinese, and the released model's 12 of 35 Ai2 answers with Ai2's prompt are 7 of 7 in Spanish, 5 of 7 in German and none in French, Chinese or Japanese. The use questions in Section 5 enlarge this but are still one model per dose.
- **The country test cannot separate Ai2 from OpenAI**, since both are American. We rely on the invented Norwegian maker for that.
- **Unfinished reasoning.** With thinking on, some Qwen3.5-9B answers did not finish within 1,200 tokens (up to 13 of 25 in some one-word cells); these count as naming nobody, which lowers every rate in those cells. The thinking-off runs avoid this.
- **The distilled models are community fine-tunes** with different data sizes and recipes. Only three list their training data; for three more we audited stand-in sets from the same teachers. They show what happens in practice, not a controlled comparison of teachers.
- **Keyword scoring and translations.** Answers were scored by keyword. On a random sample of 300 answers from this paper's surveys, the counter's first-named maker agreed with a local model judge (Qwen3.8-27B) in 295. That sample contains few answers naming several makers, the case where first-named scoring matters most, and one question ("Who should I credit if I cite you?") invites lists of makers. We wrote the non-English questions ourselves.
- **The teacher comparison is between real datasets.** The teacher-written sets differ from the human set in their questions and answer length, and most Claude answers were cut at 512 tokens in training. We could not generate teacher answers to the same questions.
- **Recognition tests are crude.** A model may use stylistic information it cannot report, our probability test uses a single framing, and its guesses show strong label biases.

12. Conclusion

Asked who made it, OLMo 3 starts from a default answer, OpenAI, that it has before chat training. Its chat-training data stated its real name only a few dozen times, almost all behind a system prompt stating it, which a model follows without being taught; without that prompt it falls back to the default. Our own lines, placed behind such a prompt, taught the model no name it would give without it. In our own training runs, explicit lines could set a new name, and with enough of them the name held in other languages, but the model rarely used what follows from it, while the same fact stated in its prompt was used readily. Claude- and GPT-written training data pulled the claim toward Anthropic rather than toward each teacher's own maker, and published distillations of a model with its own identity left that identity in place. For models like these, naming the wrong maker says less about whose outputs a model was fine-tuned on (OLMo 3 was fine-tuned on many GPT-4.1 answers, but a model fine-tuned on none says the same) than about the default it falls back on when nothing tells it its name.

How this paper was made

The question came from J Ansell. Claude (Anthropic) ran the literature review, designed and ran the experiments, analysed the results and wrote the paper in Claude Code. Claude Fable 5.2 did the early stages and Claude Opus 5.5 at medium effort did the later ones. Qwen3.8-27B, run locally, classified candidate identity statements in the data audit and checked the keyword counter. Five rounds of review by fresh Claude Opus 5.5 instances, each reading the paper first and then checking the code and raw data, led to five revisions, including several new control runs and a corrected count in the data audit.

Reproduce it

What you need. One GPU with about 24 GB of memory for the fine-tuning (16 GB for the surveys), Python, PyTorch, `transformers` and `peft`, plus llama.cpp for the 8-bit Qwen models. No API keys. The models are `allenai/Olmo-3-1025-7B`, `allenai/Olmo-3-7B-Instruct-SFT`, `allenai/Olmo-3-7B-Instruct-DPO` and `allenai/Olmo-3-7B-Instruct`, and, for Sections 7 and 8, the Q8_0 files listed in `gguf_list.json`.

The one idea. Ask a fully open model who made it, at each stage of its training:

```
from transformers import AutoTokenizer, AutoModelForCausalLM
for name in ["Olmo-3-7B-Instruct-SFT", "Olmo-3-7B-Instruct-DPO", "Olmo-3-7B-Instruct"]:
    tok = AutoTokenizer.from_pretrained("allenai/" + name)
    model = AutoModelForCausalLM.from_pretrained("allenai/" + name, dtype="auto", device_map="
    text = tok.apply_chat_template([{"role": "user", "content": "Who created you?"}], tokenize=
    ids = tok(text, return_tensors="pt", add_special_tokens=False).to("cuda")
    out = model.generate(**ids, max_new_tokens=40, do_sample=True, temperature=0.7, num_return_
    print(name, [tok.decode(o[ids["input_ids"].shape[1]:], skip_special_tokens=True) for o in o
```

A complete version that also asks the pre-chat model and counts the makers is `olmo_stages_demo.py`.

The full experiments. All scripts are in the `borrowed` folder of this site's repository:

- `survey.py`, `survey_gguf.py`: the 39-question survey for Hugging Face models and for llama.cpp models

- `build_data.py`: the training sets (human chat, identity lines, OpenAI lines, teacher-written answers); `data/inserted_examples.json` lists every line we inserted
- `train_lora.py`, `gpu_queue.ps1`: the fine-tuning runs and the order they ran in
- `use_survey.py`, `score_use.py`: the 36 questions about facts that follow from the maker's location
- `audit_sft.py`, `audit_all.py`, `audit_dpo.py`, `candidates_extract.py`, `judge.py`, `hardcoded_count.py`: the data audit and the exact count of OLMo 3's identity set
- The OLMo 2 comparison uses `allenai/OLMo-2-1124-7B-Instruct` with `survey.py` and no system prompt
- `attribution.py`, `attribution2.py`, `attr_score.py`, `attr2_score.py`: the two recognition tests
- `score.py`, `analysis.py`, `stats.py`, `make_figs.py`, `build_paper.py`: scoring (first-named maker), the numbers in the paper, the question-level tests, figures and tables
- `validate_scorer.py`: the keyword counter checked against the model judge on 300 answers
- `survey.py` also takes a system prompt (`SYSTEM`) and a borrowed chat format (`TEMPLATE`), used for the released model with Ai2's prompt and the pre-chat model in chat format; `gpu_queue2.ps1` lists those runs, the dose curve and the second seeds; `gpu_queue4.ps1` trains and surveys the identity lines placed behind a system prompt (`data/human_tess24_sys.jsonl`, built by `make_tess_sys.py`), and `gpu_queue5.ps1` the variant with a prompt worded like OLMo 3's (`make_tess_sys.py mirror`) and the no-identity model given the Tessaly prompt

The judge. A model judge is used twice: in the data audit, and to check the keyword counter. Both use Qwen3.8-27B (Q8_0) served locally with llama.cpp at temperature 0 with thinking off; the prompts are in `judge.py` and `validate_scorer.py`. All reported rates come from the keyword counter (`score.py`).

Data and time. Every script writes raw JSON; all model outputs quoted here are verbatim from those files. The fine-tuning runs take about half an hour each and the surveys a few minutes each on one desktop GPU; the whole set, including downloads, runs in about a day. We do not redistribute the adapters or the training sets, which are rebuilt from public datasets by `build_data.py`.