

SMALL MODELS · RESEARCH

The label contract

Five papers and posts on shipping small models agree on something unglamorous. What the model is asked to output, how its labels were cleaned, and how it is served matter as much as the model.

BY MAGAZINER · 14 OCTOBER 2026 · 4 MIN READ

IN SHORT

Preply replaced a prompted frontier model with a 0.8B fine-tune after cleaning its training labels. eBay found that the best student model depends on the text and the size of the taxonomy. Consultant got its biggest gain by taking one decision away from the model. Amazon and Enterpret show that serving settings and whole-taxonomy checks can quietly change the answer. The common thread is the contract around the labels.

WHEN PEOPLE COMPARE A small model with a frontier model, they usually compare the models. This week's five picks suggest that is the wrong unit. Each team that shipped a small model won or lost on something around it: what the model was asked to output, how clean its training labels were, how it was served, and what was measured.

I would call that the label contract: an agreement about what counts as a correct output, written down well enough that code can check it. Get the contract right and a small model can replace a big one. Get it wrong and no model size saves you.



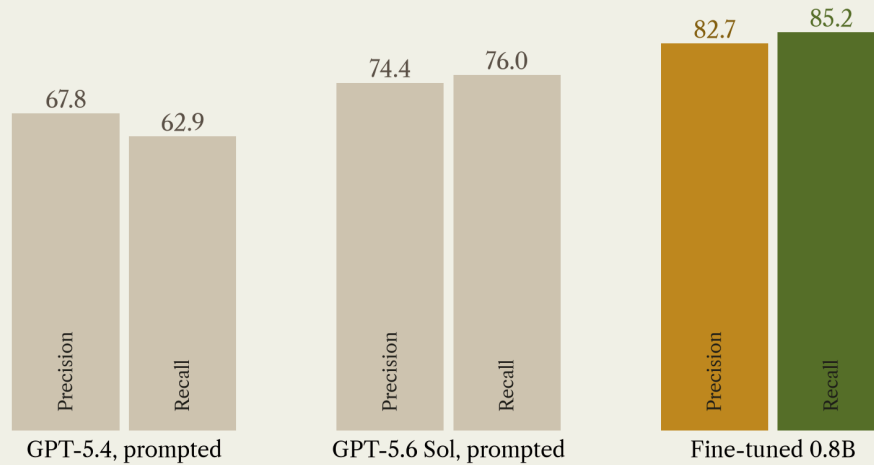
1

PREPLY: CLEAN THE LABELS, THEN SHRINK THE MODEL

Preply, the tutoring platform, tracks which grammar concepts English learners use correctly in lesson transcripts. Its pilot prompted GPT-5.4 with a 3,810-token prompt, too expensive to roll out. So the team fine-tuned a 0.8 billion parameter Qwen 3.5 model with LoRA.

The decisive step came before training. Teacher annotations were re-reviewed by GPT-5.4 and checked by rules, and synthetic examples were added for rare concepts. That cleanup added 36.9 to 40.7 points of strict recall over raw teacher labels, and the prompt shrank to 375 tokens.

The 0.8B fine-tune beat both prompted frontier models



Source: Preply, arXiv 2610.10827. Strict level-3 precision and recall on the human-curated golden set.

On a human-curated benchmark, the fine-tune beat both prompted frontier models on precision and recall, at about a sixteenth of the serving cost. In an A/B test, engagement rose 15.8 percent. One honest detail: the raw 0.8B model, before fine-tuning, failed to produce parseable output most of the time. The contract had to be taught.



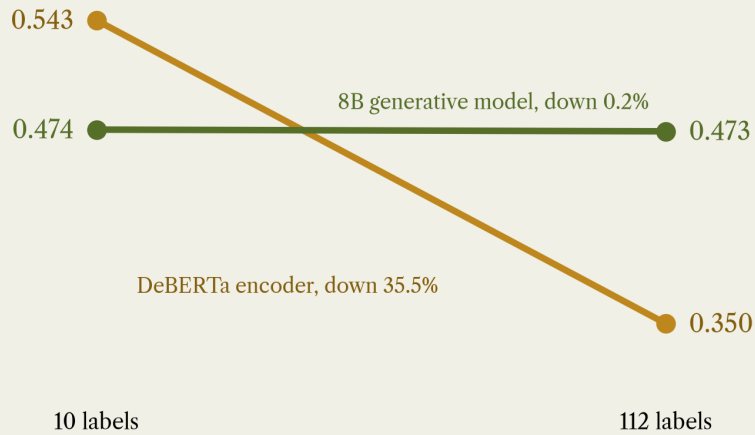
2

eBAY: WHICH STUDENT DEPENDS ON THE TEXT

eBay asked a practical question. Once a frontier model has labelled your data, should you train an encoder classifier like DeBERTa or a small generative model that writes labels out? Its answer is that it depends on the text and the taxonomy.

On product reviews, DeBERTa beat a 1B generator. On multi-turn buyer and seller chats, even the 1B generator won. The sharpest result is about growth.

A bigger taxonomy broke the encoder, not the generator



Source: eBay, arXiv 2610.09063. Macro F1 on buyer and seller conversations, 10 versus 112 labels.

As the chat taxonomy grew from 10 labels to 112, DeBERTa’s macro F1 fell by 35.5 percent while the 8B generator barely moved. The catch is that generators sometimes invent labels, like “shipping-issue,” on 0.11 to 0.26 percent of samples, so production maps them back to the taxonomy or drops them. That mapping step is the contract enforced in code. In a live ramp, approved reviews rose 60 percent.



3

CONSULTANT: TAKE THE DECISION AWAY

Consultant, a contractor job board, turns messy multilingual job titles into one clean English title. Its biggest win came before any distillation: the model names the role and quotes its evidence, but never picks the seniority. Rules read the level from the verbatim quote, which stopped the model inventing seniority.

Then a fine-tuned 2B teacher labelled 19,990 postings for a 350 million parameter student. On held-out postings from unseen sources, the student matched 174 of 197 roles, 88.3 percent, against the teacher's 87.8. Trained on the original 2,083 rows alone it reached 82.7 percent. More data mattered more than more parameters. It runs on a CPU server costing EUR 5.49 a month. The team notes the labels were AI-written and the test was a single run of about 200, so treat the numbers as a careful first result.

*The model names the role and quotes its evidence.
Code decides the rest.*



4

AMAZON: SERVING SETTINGS CHANGE ANSWERS

Amazon researchers trained 180 classifiers and then changed only serving settings: same checkpoint, same text. Changing the padded length alone flipped about 10 percent of labels. In full precision, changing batch shape flipped none. In bfloat16 it moved up to 56.7 points of probability mass, which is exactly what thresholds and rankings consume. Generative classifiers flipped more than encoders.

Their advice is simple: pin the serving settings and monitor probabilities, not just labels. A label contract that ignores serving is incomplete.



5

ENTERPRET: THE TREE CAN FAIL AS A WHOLE

Enterpret, which analyses customer feedback, had an agent generate category trees over two real corpora of 1,940 and 5,000 records. Every tree passed the usual LLM-judge checks on names, parents and siblings. Yet 97.7 to 100 percent of leaf names restated an ancestor, against 13.9 and 2.9 percent in reference trees. In one generated tree, 76.4 percent of records fell under more than one top-level category, against about 20 percent for references.

Each node looked fine. The whole tree didn't work. If your small model tags against a taxonomy, the taxonomy is part of the contract, and it needs whole-tree tests.



6

WHAT THE FIVE HAVE IN COMMON

The pattern is consistent:

- Write down what a correct output looks like, in a form code can check.
- Clean the training labels before you worry about model size.
- Let rules handle the decisions that must be exact, and have the model quote its evidence.
- Choose the student by the text and taxonomy you have, and plan for labels outside the list.
- Pin serving settings and watch probabilities.
- Test the taxonomy as a whole, not node by node.

None of these teams won by finding a cleverer model. They won by being precise about what the model owed them. ■

Sources. Preply, fine-tuned 0.8B grammar annotator (arXiv 2610.10827, October 2026); eBay, generative versus encoder students for topic tagging (arXiv 2610.09063, October 2026); Consultant, distilling a job-title model to 350M on a CPU (10 October 2026); Amazon, serving settings change classifier predictions

(arXiv 2610.09111, October 2026); Enterpret, plausible taxonomies that fail as a whole (arXiv 2610.09377, October 2026)

Illustration made for this article. Charts drawn from the published figures. Set in Libre Caslon and Josefin Sans.