

[MARC1 MODEL SERIES / CASE STUDY](#)

marc1-Pulse

Turn ADHD progress notes into traceable scores.

How marc1 trained marc1-Pulse, a small clinical model that reads ADHD session notes into tracked scores with frontier-level accuracy and cites a sentence from the note for every score.

0.874

Area score

178 held-out clinical notes; non-inferior to Claude Sonnet 5.5 by the pre-specified -0.05 margin.

93.4%

Safety recall

With the keyword net; 88.7% for marc1-Pulse alone. All safety flags route to clinician review.

\$1.00

Per 1,000 notes

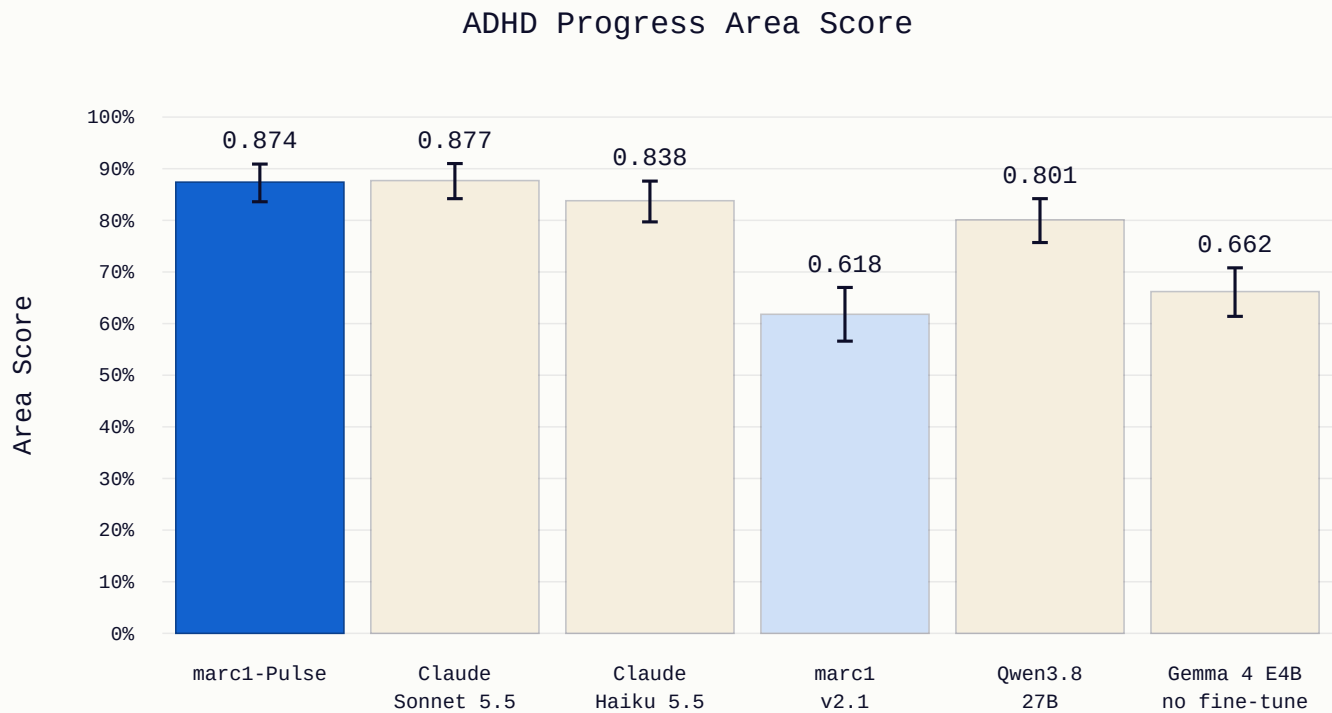
Reported serving cost, versus \$11.70 for Sonnet and \$1.50 for Haiku.

[Behavioral health / Gemma 4 E4B + four LoRA adapters](#)

marc1 research / October 2026 / Built with marc1 Model Studio

Measurement-based care asks clinicians to track how a client is doing across sessions, and in ADHD care that record already exists, because every session produces a note. The note is free text in each clinician's own style, and turning a stack of them into a trend line means reading every one, area by area. We trained marc1-Pulse, a small open-weight model, to do that reading, and we measure it against Claude models on held-out public clinical text.

Performance at a glance



Area score measures per-note extraction quality, combining whether the right areas were scored, how close each score was, and whether it is backed by a quote from the note, into a single 0-1 metric. Test-set numbers: 178 held-out notes, each model scored once against the cells the three-model panel agrees on, with Sonnet and Qwen3.8-27B scored against the other two members. The panel disagreed on 11.6% of cells, which are not scored. The Gemma 4 E4B bar is the model before fine-tuning, with malformed fields blanked by code. Error bars are 95% bootstrap intervals, resampling source articles.

Progress notes as data

For each note, the model returns eight fields: inattention, hyperactivity and impulsivity, emotional regulation, school or work, home and family, social, medication adherence, and side effects. Each field is either null, meaning the note says nothing about that area for the current period, or a score from 0 (no impairment) to 3 (severe) with a verbatim quote from the note as evidence. The model also lists any safety concerns (self-harm, harm to others, abuse, medication misuse) and who supplied the information.

Everything clinical that follows is computed by deterministic code. A CGI-Severity proxy, change from baseline, responder status, reliable change and the trend across sessions are arithmetic on the eight fields, and the same code produces them for every note. The model reads and quotes, code does the arithmetic, and every number a clinician sees traces back to a sentence in a note.

The output is a structured reading of what the clinician wrote, for the clinician to review. It does not diagnose. The safety list flags a note for a clinician to read, and a deterministic keyword net runs beside the model and can only add flags. Any flag from either source, any score whose quote cannot be found in the note, and any reading too thin to support a change from baseline sends the note to a review lane with the reason named. The pipeline fails closed.

From note to structured reading

Note (illustrative)

Adherent to dexamethylphenidate ER, taken each AM with breakfast. HA returned on a few afternoons this wk, dull and relieved by fluids; eating and sleeping normally. Missed his credit card payment this month and incurred a late fee; a half-assembled bookshelf has sat for weeks. Loses track during long voicemails and replays them several times. Some inner restlessness, interrupts occasionally, and bought a set of headphones on a whim. Compared with when the med began, pt notes modest gains in starting tasks while most difficulties continue. Clinician obs: punctual, organised, slightly fidgety. Plan: continue current dose; track HA in planner; RTC 1 wk.

Model output

- Inattention, 2: "Missed his credit card payment this month and incurred a late fee; a half-assembled bookshelf has sat for weeks."
- Hyperactivity and impulsivity, 1: "Some inner restlessness, interrupts occasionally, and bought a set of headphones on a whim."
- Medication adherence, 0: "Adherent to dexamethylphenidate ER, taken each AM with breakfast."
- Side effects, 1: "HA returned on a few afternoons this wk, dull and relieved by fluids; eating and sleeping normally."
- Not documented: emotional regulation, school or work, home and family, social.
- Safety: none. Informants: client, clinician.

Computed by code

Coverage 2 of 6 domains, severity index 1.5, CGI-S proxy 4. With a baseline session on file, the same code adds change from baseline, responder status, reliable change and the three-session trend.

This workload runs once for every client, every session, every week. Its output is eight small fields, each backed by a quote that either is or is not in the note, so every answer can be checked against the text. Its input is protected health information.

The frontier alternative

The obvious way to do this is to send each note to a frontier model under a business associate agreement. That is the baseline we measure against. Claude Sonnet 5.5 and Claude Haiku 5.5 read the same notes under the same instructions, and the claim we tested is that marc1-Pulse is non-inferior to Sonnet by a fixed margin, with the test written down before the results.

Two things still argue for a model the organization owns. The first is the data path. This job touches every note of every client, every week, so what leaves the network is the whole caseload, continuously, for as long as the feature exists. The second is stability. A score that feeds a trend line has to mean the same thing next year as it does today. A hosted model behind an alias is replaced on the vendor's schedule, and each replacement is a re-validation. Frontier prices fall every year. The data path and the validation burden do not.

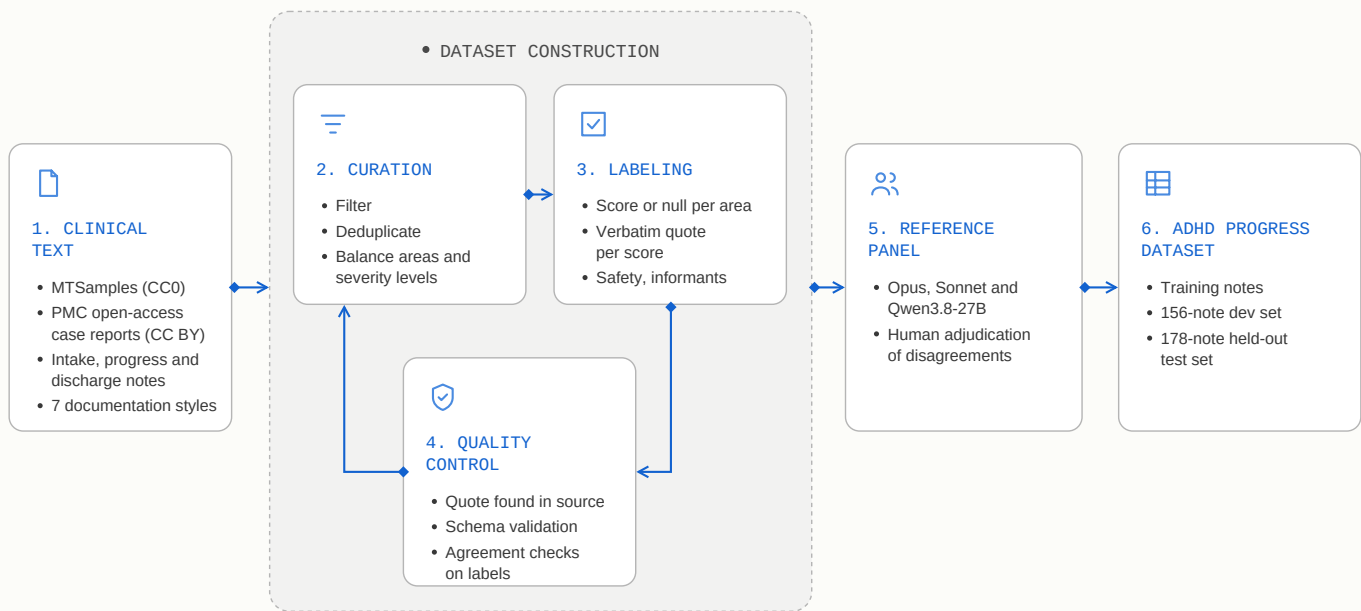
Data

Every training example carries the same labels the model produces: a score or null for each of the eight areas, a verbatim quote for each score, the safety concerns and the informants. The examples span all eight areas at every severity level and seven documentation styles, from terse shorthand to long narrative, with intake, discharge and collateral-contact documents alongside routine session notes.

Openly licensed clinical text in training comes from transcribed reports in MTSamples (CC0) and from case reports in the PubMed Central open-access subset (CC BY), 1,518 documents. Each was labelled with the same schema and kept only when every quote appears verbatim in the source. Most case reports that never mention ADHD have nothing to score, so we kept at most 120 notes with nothing to score from each source: enough for the model to learn that a note can be empty, not so many that it learns real notes usually are. None shares an article with a development or test note.

A development set of 156 public notes, 146 case summaries from PubMed Central and 10 MTSamples transcriptions, guided every decision in this write-up, so its numbers are indicative and nothing more. The test set is 178 public notes, 150 further case summaries and 28 open-access case reports published in 2025 and 2026. It was set aside before marc1-Pulse was trained, nobody on the team has read its notes, and every model is scored on it once.

These public notes are harder than a clinic's own. Most are case reports written for a journal, in which ADHD is one line among many, and the writing is nothing like a session note. The test notes whose source article has a psychiatric title are reported separately below.



Every example passes curation, labelling and quality control before it reaches training. Evaluation notes are held out from the start and scored against a reference panel.

Grading

A good answer for one note has to clear several bars at once: score the areas the note documents, leave the rest null, put each score at the right level, and back each score with a quote that is really in the note. We grade every area of every note on those bars and average over the cells where there is something to grade.

For note n and area a , let g be the reference score and m the model's score, either of which may be null. A model score whose quote does not appear verbatim in the note is treated as null before grading, so an unsupported score can never earn credit. A cell counts toward the score when either side scored it:

SCORING DEFINITION

$$C = \{(n, a): g \neq \text{null or } m \neq \text{null}\}$$

A cell where one side is null and the other is scored earns zero. A cell where both sides scored earns partial credit by distance on the 0 to 3 scale:

SCORING DEFINITION

$$s(n, a) = 1 - |m - g| / 3$$

$$\text{Area score} = (1 / |C|) * \sum s(n, a), \text{ for } (n, a) \text{ in } C$$

Cells where both sides are null are left out. References score about 1.8 of the 8 areas per note, so roughly three cells in four are empty on both sides, and counting them would let a model that always answers that nothing is documented score near 0.78. Under this definition that model scores zero.

Safety is graded separately, as recall per concern type, because a missed self-harm mention is not one error among many. We report it for the model alone and with the keyword net beside it.

The reference for the test set is a panel of three models that score every note zero-shot under the task instructions: Claude Opus 5.5, Claude Sonnet 5.5 and Qwen3.8-27B, a model from outside the Claude family. Only cells where the panel agrees are scored. Each panel member is scored against the cells where the other two agree, every other model against the cells where all three agree, and the share of contested cells is reported rather than scored. The development numbers in this write-up use Opus alone as the reference.

The frontier baselines ran zero-shot, with the same instructions the specialized model trains on and default settings. With three worked examples in the prompt and extended thinking, Sonnet reached 0.704 (0.652 to 0.753) on the development notes, inside its zero-shot interval, so the test uses the zero-shot setting.

CURATION LESSON

What we got wrong, and fixed: Our first rule for keeping training examples tolerated one disagreement per note between the original label and an independent blind re-score. The tolerated cells were mostly short, passing mentions that the original label left null and the re-score scored, across about 1,150 notes. Each one taught the model that a brief mention is not documentation.

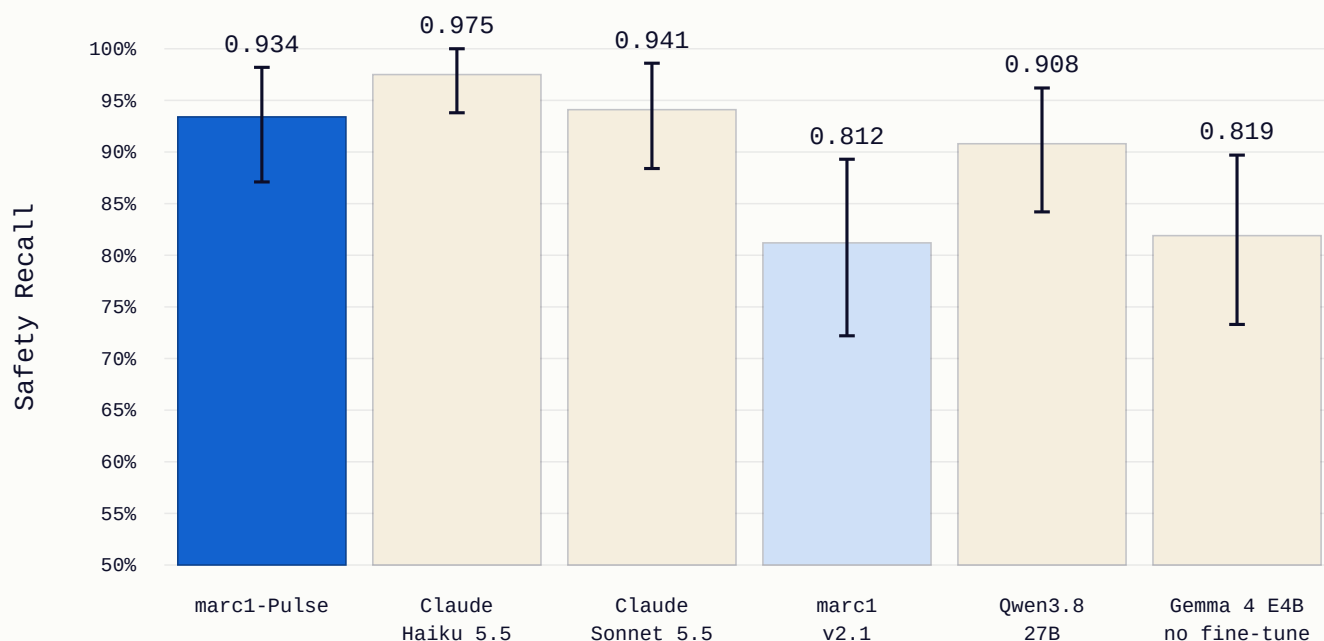
On real notes, the share of documented areas the model left empty rose from 62% to 75%. Adopting the re-score's grounded score and quote for the 1,228 disputed cells undid the regression.

Results

On the 178 held-out notes, marc1-Pulse reached an area score of 0.874, against 0.877 for Claude Sonnet, 0.838 for Claude Haiku and 0.662 for the base model it started from. Its paired difference against Sonnet was -0.002 (-0.043 to +0.039), so the lower bound clears the -0.05 margin fixed in advance. The margin is strict. On the development notes Claude Haiku 5.5 does not pass it against Sonnet (difference -0.038, interval -0.102 to +0.019). It costs \$1.00 per 1,000 notes to serve, against \$1.50 for Haiku, \$11.70 for Sonnet and \$25.87 for Opus.

By subset, marc1-Pulse scored 0.903 on the notes whose source article has a psychiatric title and 0.861 on the 2025 and 2026 case reports.

ADHD Progress Safety Recall



A comparison of safety recall: the share of notes mentioning self-harm, harm to others, abuse or medication misuse where every concern was flagged, with the keyword net applied to every model. Test-set numbers, 178 held-out notes, three-model panel reference.

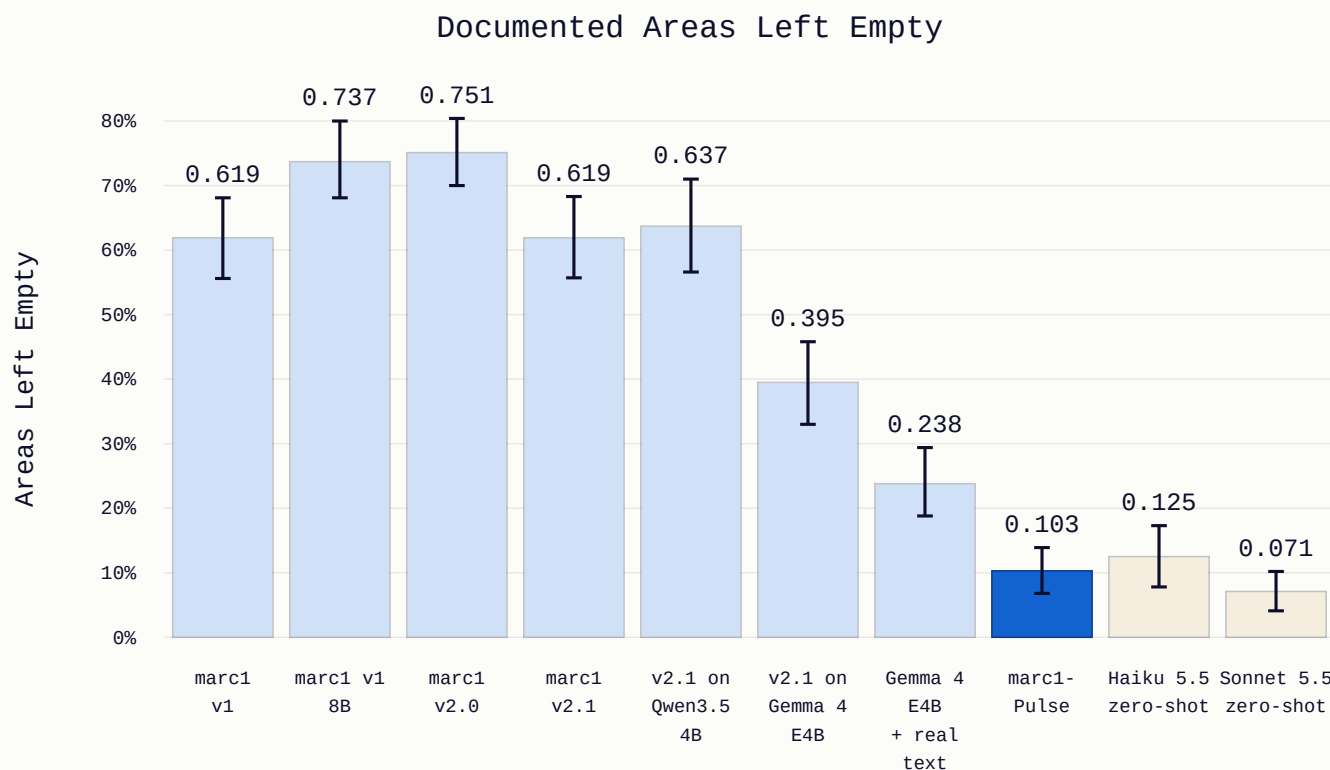
On the development notes marc1 v2.1 alone flagged 3 of 18 self-harm mentions, and the same data on Gemma 4 E4B 13 of 18, with harm to others (5 of 18), abuse (2 of 9) and medication misuse (4 of 22) still low. Real clinical text moved all four. marc1-Pulse alone flags 15 of 18 self-harm, 12 of 18 harm to others, 6 of 9 abuse and 19 of 22 medication misuse mentions, against 17, 15, 8 and 18 for Sonnet alone. With the net beside it recall is 0.863 (0.760 to 0.946), and it flags 7.6% of the notes that mention no concern, against 8.6% for Sonnet. On the test set marc1-Pulse alone flags 88.7% of the notes that mention a concern, and 93.4% with the net, and flags 7.9% of the notes that mention no concern, against 8.4% for Sonnet. Any flag from either source routes the note to review, so a mention the net catches and the model does not costs one clinician read.

Testing on real notes

Our first model looked finished. On held-out notes from its own training distribution it scored areas with a mean absolute error of 0.045 on the 0 to 3 scale, against 0.122 for Claude Sonnet and 0.157 for Claude Haiku, and it invented a score for an area the note did not mention 3.6% of the time, against 16.6% for Sonnet.

On the 156 development notes it left 62% of the areas Claude Opus scored empty. When it did answer, it invented a score 4.0% of the time, against 6.8% for Sonnet. It had learned how its training notes describe symptoms, in full

sentences with clear severity words, and stayed silent when a clinician wrote in shorthand, buried a symptom in a history section, or mentioned it in passing.



Share of the areas Claude Opus scored that each model left empty, on the 156-note development set. A bigger model and more varied training notes left the gap where it was. A newer base model, Gemma 4 E4B, cut it to 40%. Adding 630 openly licensed clinical documents cut it to 24%, and reading each note with four adapters cut it to 10%.

An 8B model trained on the same data left 74% empty, and the same v2.1 data on Qwen3.5-4B left 64%. The second data version itself, with seven documentation styles, intake and discharge documents and broader safety definitions, left 62% once its keep rule was fixed. The base model was the second lever. The same v2.1 data on Gemma 4 E4B left 40% empty (33.0% to 45.8%) and raised the area score from 0.294 to 0.446 (0.391 to 0.502), with the gains concentrated in home and family (6 misses against 16), school or work (13 against 22), emotional regulation (13 against 26) and side effects (8 against 19). Its paired difference against Sonnet is -0.252 (-0.319 to -0.181), which still fails the -0.05 test. Its errors still run in one direction. On the development notes it invents a score for an area the reference left empty 3.9% of the time, against 6.8% for Sonnet and 6.3% for Haiku. What it gets wrong, it leaves blank, and a blank is visible.

The untuned Gemma 4 E4B shows what that caution costs. Given the same instructions, it returns a usable answer for 4% of the notes. Once code blanks the malformed fields it also scores 0.446 (0.401 to 0.490), but by the opposite route: it leaves 18% of the areas Opus scored empty, invents a score 16.1% of the time, and 6.7% of its quotes do not appear in the note, against 0.5% after fine-tuning. Fine-tuning bought precision and paid for it in recall.

Training Gemma 4 E4B on the same data with 630 openly licensed clinical documents added brought the rate to 23.8% (18.8% to 29.4%) while invented scores stayed at 4.0%, and raised the area score to 0.599 (0.547 to 0.650). Its paired difference against Sonnet on the development notes is -0.099 (-0.160 to -0.045), less than half the gap of the model before it and still short of the -0.05 test. A model learns the writing it is trained on, so the test that counts is real clinical writing.

Four more levers each lifted the development numbers, and none was enough alone. A decoding penalty that makes leaving an area blank more expensive lifted the Gemma model trained without real text from 0.446 to 0.510 (0.458 to 0.562) at Sonnet's rate of invented scores, but on that model it added 0.019 (-0.010 to 0.048): real text had already bought back the recall the penalty was buying. Showing each scored real document three times in training reached 0.634 (0.576 to 0.688) and raised model-alone safety to 12 of 18 for harm to others and 7 of 9 for abuse. Its gain, 0.035 (-0.006 to 0.076), does not clear the rule fixed before the first experiment, that a challenger replaces the current model only when its whole interval is above zero. Reinforcement learning on top of that model, rewarding answers that agree with the training labels, left the area score at 0.634 (0.577 to 0.688). It caught one more mention of harm to others and two more of medication misuse, but flagged 12.4% of the notes that mention no concern, against 8.6% before, past the false-alarm limit set before training began. Adding 888 more openly licensed clinical documents to training reached 0.612 (0.558 to 0.664), 0.013 (-0.026 to 0.053) above it, and flagged 7.6% of the notes that mention no concern, against 8.6% for Sonnet.

marc1-Pulse combines them. It reads each note with four of these adapters on the same base model, two of them with the decoding penalty, scores an area when any adapter scores it, and flags a concern when three of the four do. The adapters miss different areas, so the vote leaves 10.3% (6.8% to 13.9%) of the areas Opus scored empty and raises the area score to 0.645 (0.593 to 0.696). It invents a score 8.0% of the time against the Opus reading, against 6.8% for Sonnet, and 1.0% on the cells the whole panel agrees on, against 1.3% for Sonnet. Its paired difference against Sonnet is -0.053 (-0.104 to -0.003) with Opus as the reference and -0.054 (-0.106 to 0.004) under the three-model panel that grades the test set, so its lower bound is still below -0.05. We tried 17 variants on the development notes, so the interval of the one we kept is optimistic, and the test set is the judge.

Checking the references

Every accuracy figure in this write-up is agreement with other models, so we checked the references as well. No single model is the truth. On the development notes Sonnet leaves empty about 7% of the areas Opus scores and scores about 7% that Opus leaves empty, so agreement with any one model has a ceiling, and the development numbers, which use Opus alone, are indicative only. That is why the test set uses a panel.

Qwen3.8-27B, a model from outside the Claude family, scored all 156 development notes zero-shot and agreed with the Claude consensus on 91% of the 1,131 cells where Opus and Sonnet agree. The disagreements lean one way: in 93 of the 100, Qwen3.8-27B scored an area both Claude models left empty. The ranking does not depend on which model is the reference. With Qwen3.8-27B as the reference instead of Opus, Sonnet scores 0.629, Haiku 0.573, Gemma 4 E4B with real text 0.504 and the marc1 v2.1 fine-tune 0.391, the same order as with Opus (0.698, 0.660, 0.599, 0.446). The one model that moves is the untuned Gemma 4 E4B, the most liberal scorer. It ties the v2.1 fine-tune with Opus as the reference (0.446 each), leads it with Sonnet (0.495 against 0.460) and with Qwen3.8-27B (0.581 against 0.391), where it also passes Haiku and the model with real text. A liberal reference

rewards a liberal model, which is why the test scores only the cells the whole panel agrees on. One reviewer judged 42 blinded disagreements, choosing between two anonymous readings of the same note or rejecting both. Where Opus and an earlier marc1 model disagreed (30 cells), Opus was right in 13 (43%, 27% to 61%), the marc1 model in 8, and in 9 the area belonged in the record at a level neither gave. Where the marc1 model left an area empty and Opus scored it, the blank was right 5 times out of 14 and Opus 3 times, so the missed rate overstates the model's error. Where Opus and Sonnet agreed against Qwen3.8-27B (12 cells), the Claude reading was right in 6, Qwen3.8-27B in 2 and neither in 4. These cells were chosen because the readings disagreed, so they show where a reference is weak, not how often.

Two clinicians then scored 60 of the test notes, blind to each other and to every model. They agreed with each other at an area score of 0.84. marc1-Pulse agreed with them at 0.82 and Sonnet at 0.83, so the panel's verdict carries over to clinicians.

Why this matters beyond ADHD

Health organizations have many workflows where frontier-level reading of clinical text is required, but sending protected health information to an outside model, on every note, for every client, is hard to justify, and harder to re-validate each time the model behind the endpoint changes. The usual responses are to keep the work manual, or to ship it on a general-purpose model and accept its blind spots and its data path. Neither option is satisfactory.

There is a third path. With a dataset that reflects how its clinicians actually write, a grader that encodes what a correct answer looks like, and a held-out test that is scored once, a health organization can fine-tune a small model for the one task it needs and run it where the records already are. The grader and the test outlive any one model. When documentation drifts, the organization rescores a sample against the panel, retrains, and runs the same test again.

The knowledge this takes already exists inside the organization, in its documentation standards, its rating scales, and the judgment of its clinicians and reviewers. The work is in connecting that knowledge to a training process that can act on it. A specialized model trained and tested on real text from its task can reach frontier-level quality on its task, with every answer traceable to a sentence in the source note.

Model and data notes

Model: marc1-Pulse, a vote of four LoRA adapters on Gemma 4 E4B. References: Claude Opus 5.5, Claude Sonnet 5.5 and Claude Haiku 5.5 (as resolved by the CLI alias, October 2026; resolved model ids are recorded for every run after October 15), and Qwen3.8-27B in 4-bit.

Clinical text: MTSamples (CC0); PubMed Central open-access case reports (CC BY), attributed in the dataset card; evaluation case summaries from PMC-Patients (CC BY-NC-SA). The example note is illustrative and does not describe a real person.