Distilling and Surpassing Vendor LLM Scoring of Depression Interviews with Synthetic Transcripts and Expert Labels

Kolar A[1,2,3], Pinheiro MA[1,2], Deschamps A[2], Solomon TM[2], Karlin DR[2,4]

[1] MIRA Analytics [2] Definium Therapeutics [3] Faculty of Informatics, Masaryk University, Brno, Czech Republic [4] Department of Psychiatry, Tufts University School of Medicine

Key Words: clinician-reported outcomes, large language models, knowledge distillation, synthetic data, MADRS

Methodological Issue Being Addressed

Clinician-reported outcomes (ClinROs) from participant interviews are the primary efficacy endpoints in depression, anxiety or schizophrenia. Because they rely on clinical judgement, they carry well-documented measurement risks: variability between raters, variability within the same rater over time[1], and sensitivity to how an interview is conducted. Sponsors manage these risks with rater training, implementation of central raters, blinded analytics and third-party endpoint review[2] — established approaches that add time, cost and operational complexity, and that in practice cover a small subset of the total assesments conducted in a trial rather than all of them.

Large language models offer a complementary route. An LLM reads an interview transcript and assigns a severity score, applying one uniform scoring policy across a study[3]. The approach scales with the size of the dataset rather than with reviewer availability, so every interview can be rescored and every discrepancy surfaced for human review. We reported this previously on HAM-A interviews from a Phase 2b trial[4].

The open question is which model can support a validated endpoint. Closed source instances — GPT 5.5, Claude Opus 4.6, Gemini 2.5 Flash — each raise a distinct concern:

Leaks	Transcripts may leave the sponsor's environment to be scored
Drift	An instance can be updated or retired mid-study, so the same interview scores differently over time
No audit	Weights and per-item internals cannot be frozen or inspected
Newer ≠ better	A stronger general release is not automatically better on a narrow clinical task

A self-hosted open source model answers all four — frozen · version-pinned · audited and inspectable · inside the sponsor's environment through database lock. Untuned it scores poorly, and tuning needs real labeled interviews, which are difficult to obtain: they sit behind sponsor firewalls, are rarely standardized across programs, and carry PII and consent constraints that are hard to revisit retrospectively.

Training such a model therefore depends on datasets most groups cannot reach. This work asks whether a method exists to train one without them — whether suitable training data can be generated instead.

Introduction (Aims)

Distillation transfers the scoring ability of a large closed source model — the teacher — into a much smaller open source model that can be self-hosted[5]. Neither the teacher's weights nor real patient data can be shared, so the transfer happens through data the teacher writes itself: synthetic interview transcripts, each produced at a known severity, on which the open source model is then trained and frozen.

A transcript generated at a requested severity carries that severity as its label. Label quality can be improved by re-assigning it before training — having the transcript re-scored by the teacher itself, or by a model trained on real interviews. **Transcripts and labels are therefore two independent levers:** holding one fixed while varying the other measures each contribution separately.

Three roles recur throughout:

Teacher	A large closed source model. Writes the synthetic transcripts and the severity label each was generated at.
Student	A small open source model, self-hosted and frozen. Trained on those transcripts — the model being distilled.
Expert	A model trained on real labeled interviews. Serves as the accuracy ceiling, and can re-score the transcripts before the student is trained.

This work asks two questions.

> **1) Can an open source model trained only on synthetic transcripts approach dated closed source teacher snapshots — and how far does it sit from the same architecture trained on real labeled interviews, the expert model?**

> **2) If those transcripts are re-scored before training, does the distilled model improve — enough to surpass the teacher that wrote them?**

Methods

Building the synthetic corpora and distilled models

Building a corpus takes two prompts. The first asks the teacher to generate rater and participant profiles at the scale of a real trial: demographics, personal background, medical and psychiatric history, and interviewing style.

The second samples those profiles and asks the teacher to write an interview snippet for one participant, one rater and one MADRS item[6] — for example reduced sleep — at a requested severity. That requested severity is stored as the snippet's label: it records what the teacher was asked to write, not an independent reading of what it actually produced.

Iterating over items, severities and profiles builds a balanced corpus of ~30,000 item snippets, roughly 3,000 interviews — about three times the Phase 2b corpus behind our 2025 poster[4]. We built three such corpora (A, B, C) whose generating prompts differ in how much detail they give about raters, population and item synthesis, producing distilled models A, B and C. Prompt design is an area of active research; the pattern reported here held across all three. Figure 1 summarises the pipeline.

Models compared

Baseline	Llama 3.1, untuned
Closed source	Gemini 2.5 Flash · Claude Opus 4.6 · Claude Opus 4.8
Distilled (ours)	seed labels · teacher-relabel · expert-relabel
Ceiling	expert model, trained on real interviews

Self-hosted models were frozen before evaluation; closed source models ran on dated snapshots.

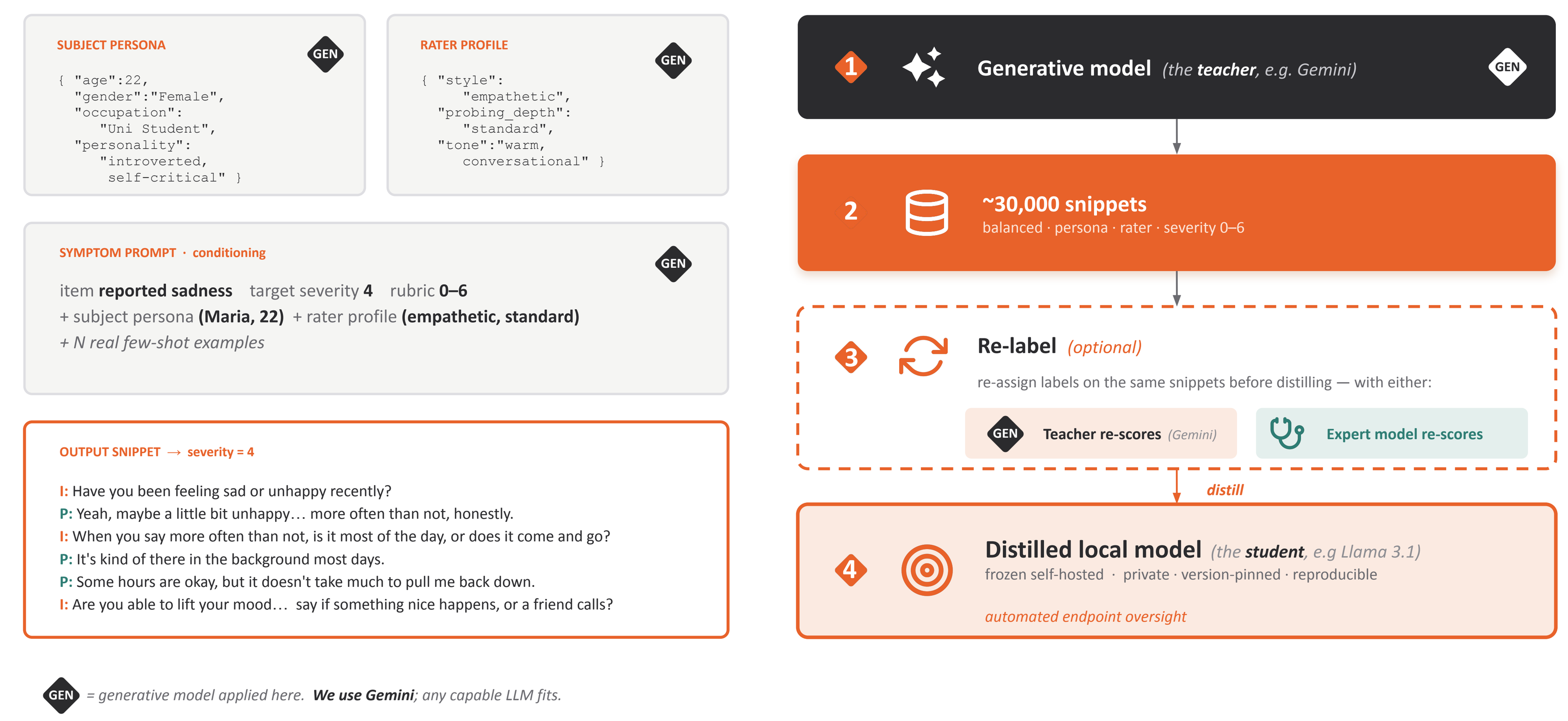

SUBJECT PERSONA
GEN
{ "age":22,
"gender":"Female",
"occupation":
"Uni Student",
"personality":
"introverted,
self-critical" }
RATER PROFILE
{ "style":
"empathetic",
"probing_depth":
"standard",
"tone":"warm,
conversational" }
SYMPTOM PROMPT · conditioning
item reported sadness target severity 4 rubric 0–6
+ subject persona (Maria, 22) + rater profile (empathetic, standard)
+ N real few-shot examples
OUTPUT SNIPPET → severity = 4
I: Have you been feeling sad or unhappy recently?
P: Yeah, maybe a little bit unhappy… more often than not, honestly.
I: When you say more often than not, is it most of the day, or does it come and go?
P: It's kind of there in the background most days.
P: Some hours are okay, but it doesn't take much to pull me back down.
I: Are you able to lift your mood… say if something nice happens, or a friend calls?
1 Generative model (the teacher, e.g. Gemini)
2 ~30,000 snippets
balanced · persona · rater · severity 0–6
3 Re-label (optional)
re-assign labels on the same snippets before distilling — with either:
Teacher re-scores (Gemini)
Expert model re-scores
distill
4 Distilled local model (the student, e.g Llama 3.1)
frozen self-hosted · private · version-pinned · reproducible
automated endpoint oversight
= generative model applied here. We use Gemini; any capable LLM fits.

Figure 1. Closed source models such as Claude Opus 4.6 or Gemini 2.5 Flash are prompted first to create rater and participant profiles, then to write interview snippets at an assigned severity — which becomes the training label. Those labels can optionally be re-assigned by a strong closed or open source model before distillation. **No trial data enters the corpus at any step.**

Expert model and held-out test set

Real data came from a double-blind, placebo-controlled Phase 3 major depressive disorder trial. References are the central-rater MADRS scores of record: each interview was scored by one of 10 central raters, with no item double-rated or adjudicated.

The expert model is a Llama 3.1 fine-tuned on real pre-cutoff interviews from that trial. It serves both as the accuracy ceiling and, for the second question, as the relabeler. The held-out test set is every interview recorded after that training cutoff — a temporal holdout mirroring prospective deployment — giving 206 visits and 2,060 item ratings. All models scored identical de-identified transcripts. No trial transcript, score or patient datum entered any synthetic corpus, so nothing the open source models trained on overlaps with what they were tested on.

Outcome measures

Primary	item-level MAE against the central-rater score, with exact item agreement. This is the granularity at which a quality flag is actionable — a reviewer needs to know which item disagrees.
Secondary	visit-level MAE and ICC on the 0–60 total, plus signed bias to separate systematic offset from random scatter.

Because a total sums ten items, item errors can cancel, so a model can reach the right total for the wrong reasons; conversely a systematic per-item offset is amplified tenfold.

Results

Results span all three synthetic corpora; a single corpus is named wherever one is quoted. Figure 2 shows the label-source effect in all three; Table 1 compares every model, its distilled arms built on corpus C.

Untuned Llama 3.1 does not agree closely enough with central raters to serve as an automated second set of scores (item MAE 1.10; 25% exact agreement; ICC 0.51). Trained on synthetic transcripts carrying only their generation labels, the distilled model reached item MAE 0.543–0.697 and 48–59% exact agreement across the three corpora, against 0.451–0.528 and 56–61% for the closed source snapshots — without ever seeing a patient transcript.

Relabeling before distillation is what moves the model, and even the teacher can do it. Re-scoring its own transcripts, with no expert model involved, brought the distilled model to item MAE 0.504 on corpus B — inside the closed source band — and to 0.540 and 0.567 on A and C, within 0.01 and 0.04 MAE points of it.

> **Relabeling by the expert model went further. On corpus C the distilled model reached item MAE 0.40 and 68% agreement, surpassing the Gemini 2.5 Flash teacher that wrote its training data (0.47, 61%).**

With expert labels every corpus landed inside the closed source band or better — 0.527, 0.470 and 0.403 for A, B and C — even though the three corpora were generated by different prompts (Figure 2). The synthetic transcripts therefore carry enough signal to distil a high-quality open source model; what limits the result is the quality of the labels attached to them. Transcripts, split, base model and hyperparameters were held fixed; only the labels changed.

A newer closed source release did not help: Claude Opus 4.8 scored no better than 4.6 (item MAE 0.53 vs 0.45; 56% vs 60%), against the direction of public benchmarks. Version currency is not a proxy for clinical accuracy. The expert model, trained on real interviews, outperformed every closed source snapshot tested (item MAE 0.28 vs 0.45–0.53; 75% vs 56–61% exact agreement).

What relabeling removes is systematic bias. Models trained on the teacher's generation labels scored low consistently, and because a MADRS total sums ten items that offset compounds while random error partly cancels.

> **On corpus C a per-item bias of −0.51 became −5.07 points of visit-level bias, and ICC fell to 0.83. Expert relabeling returned bias to near zero (+0.08 per item) whichever corpus trained the model.**

An offset of this kind can be partly corrected by recalibration, but relabeling removes it at source (Table 1).

Label quality is the bottleneck

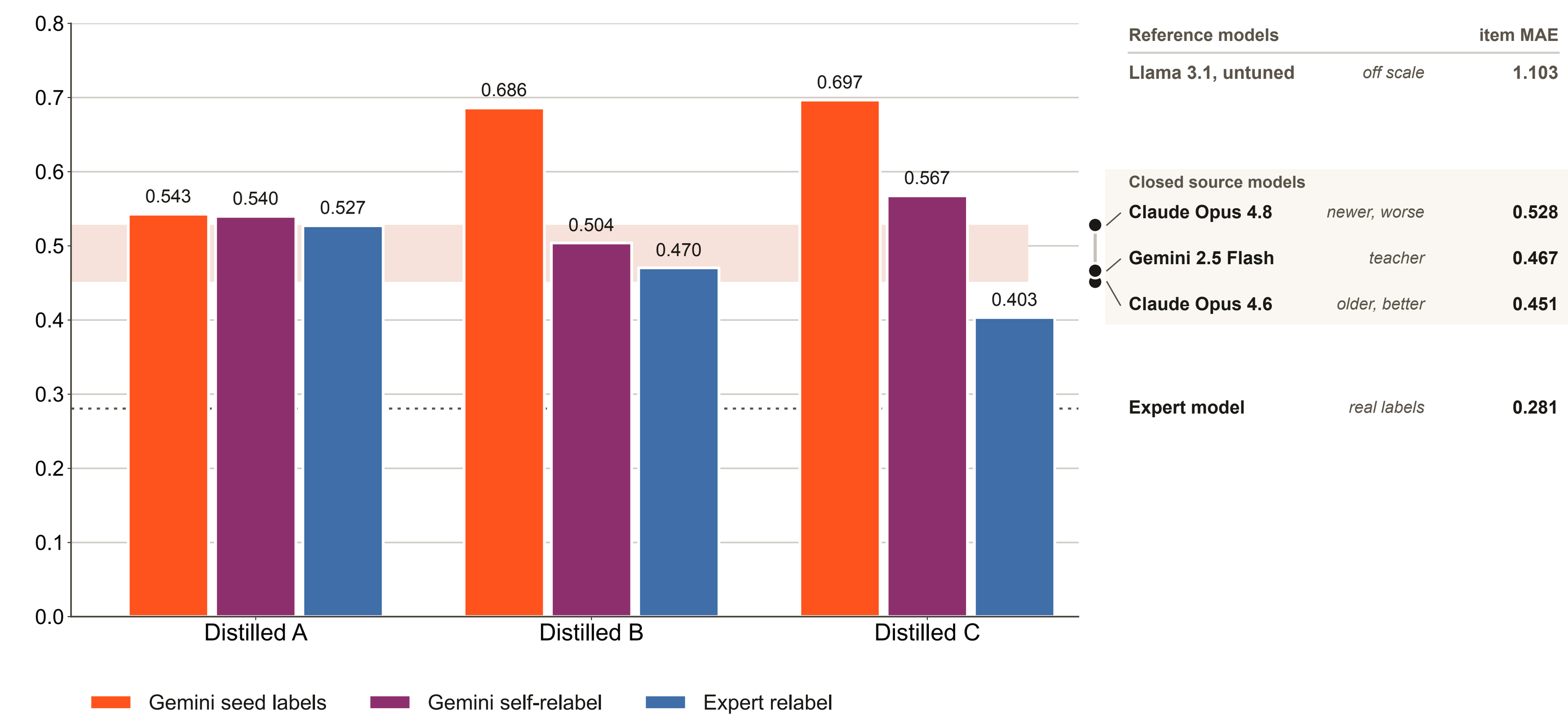

0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.543
0.540
0.527
0.686
0.504
0.470
0.697
0.567
0.403
Distilled A
Distilled B
Distilled C
Reference models item MAE
Llama 3.1, untuned off scale 1.103
Closed source models
Claude Opus 4.8 newer, worse 0.528
Gemini 2.5 Flash teacher 0.467
Claude Opus 4.6 older, better 0.451
Expert model real labels 0.281
Gemini seed labels
Gemini self-relabel
Expert relabel

Figure 2. Three synthetic datasets distilled from the same Gemini 2.5 Flash teacher, differing only in the prompt used to generate them. Each was then relabeled by Gemini itself and by the expert model. Relabeling improves performance in every dataset.

Held-out performance

	item MAE	item bias	exact agreement	within ±1 pt	visit MAE	visit bias	ICC	within ±5 pts
Expert model (real data)	0.281	+0.03	75%	98%	1.64	+0.30	0.981	99%
Distilled — expert relabel	0.403	+0.08	68%	94%	2.36	+0.84	0.955	92%
Claude Opus 4.6	0.451	-0.17	60%	95%	2.58	-1.66	0.949	89%
Gemini 2.5 Flash	0.467	-0.06	61%	94%	2.28	-0.65	0.956	95%
Claude Opus 4.8	0.528	-0.11	56%	93%	2.59	-1.08	0.948	96%
Distilled — Gemini relabel	0.567	-0.31	58%	89%	3.64	-3.12	0.901	79%
Distilled — seed labels	0.697	-0.51	48%	86%	5.32	-5.07	0.828	56%
Llama 3.1, untuned	1.103	+0.18	25%	77%	6.45	+1.78	0.510	53%

Table 1. 206 visits / 2,060 item ratings, sorted by item-level MAE; distilled models are built on corpus C. Models we built are shown in bold. Signed bias separates systematic offset from scatter.

Conclusions

> **Relabeling synthetic transcripts before distillation is what moves a self-hosted open source model: re-scored by the teacher it reaches the closed source band, and re-scored by an expert model it surpasses the teacher itself — having never seen a patient transcript.**

Synthetic-transcript distillation transferred closed source MADRS scoring into a frozen, self-hosted open source model. The expert model trained on real interviews remained most accurate, though its training and evaluation labels share the same central-rater pool.

For endpoint oversight this is a practical trade: a privacy-preserving, version-pinned, inspectable second set of scores, produced for every interview and validated locally before use. Because reaching the closed source band needs only the teacher and no real labeled interviews, oversight can begin before any trial data exists, with models retrained and relabeled as real data accumulates.

Because the corpora contain no patient data, a synthetic dataset carrying expert-model labels can also be published — a standardized, openly available training resource and public benchmark for ClinRO scoring. Academic groups without access to trial data could train and benchmark their own models on it, free of the PII and consent barriers that keep real interviews behind sponsor firewalls.

How far the level of detail in the generating prompt drives distilled model quality is the subject of further research. These are single-trial, single-rater results and need replication across scales and studies. As a direction, distillation with relabeling is a practical path to endpoint oversight that does not depend on a single closed source provider.

References

1. Trajković G, Starčević V, Latas M, Leštarević M, Ille T, Bukumirić Z, Marinković J. Reliability of the Hamilton Rating Scale for Depression: a meta-analysis over a period of 49 years. Psychiatry Res. 2011;189(1):1–9.
2. McNamara C, Engelhardt N, Potter W, Yavorsky C, Masotti M, Di Clemente G. Risk-Based Data Monitoring: Quality Control in Central Nervous System (CNS) Clinical Trials. Ther Innov Regul Sci. 2019;53(2):176–182.
3. Weber S, Deperrois N, Heun R, et al. Using a fine-tuned large language model for symptom-based depression evaluation. npj Digit Med. 2025;8:598.
4. Kolar A, Pinheiro MA, Deschamps A, Solomon TM, Hajduk M, Majernik M, Karlin DR. Using Large Language Models for Endpoint Oversight. Poster presented at: ISCTM Autumn Conference; 2025.
5. Woo B, Burkhart MD, Alsentzer E, Beaulieu-Jones BK. Synthetic data distillation enables the extraction of clinical information at scale. npj Digit Med. 2025;8:267.
6. Montgomery SA, Åsberg M. A new depression scale designed to be sensitive to change. Br J Psychiatry. 1979;134:382–389.

Disclosures

Supported by Definium Therapeutics. A Kolar and MA Pinheiro are affiliated with MIRA Analytics and Definium Therapeutics; A Kolar is also affiliated with the Faculty of Informatics, Masaryk University; A Deschamps and TM Solomon are affiliated with Definium Therapeutics; DR Karlin is affiliated with Definium Therapeutics and the Department of Psychiatry, Tufts University School of Medicine. Relevant financial relationships have been reviewed and mitigated to ensure the content is evidence-based and unbiased.

One or more authors report potential conflicts which are described in the program

Poster PDF