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

Submitter Adam Kolář

Affiliation Mira Analytics Inc.

SUBMISSION DETAILS

I agree to provide poster pdf for attendee download. Yes

What is the Methodological Issue Being Addressed? Clinician-reported outcomes (ClinROs) from interviews are the gold standard for assessing depression in clinical trials, yet suffer from clinician bias, inter- and intra-rater variability, and subjectivity. Large language models (LLMs) can rescore interviews uniformly for oversight, but vendor-hosted models (e.g. GPT-5.4) expose patient data outside the sponsor's environment, can change or be retired mid-study, and newer, generally better releases might not improve on clinical tasks. Self-hosted open-source models can instead be frozen, version-pinned, audited, and kept in-house, but score poorly untuned, and real labeled data to train them may be scarce. This work explores ways to close that performance gap.

Introduction Using the Montgomery–Åsberg Depression Rating Scale (MADRS), we benchmarked mean absolute error (MAE) and exact agreement vs central-rater scores on held-out Phase 3 interviews. We addressed two distinct methodological questions. (1) Distillation: can an open-source, self-hosted model distilled (trained to imitate) from synthetic MADRS transcripts approach a vendor band of dated proprietary snapshots — and how does it compare with the same open-source model trained on real labeled interviews? (2) Expert relabeling: can relabeling those synthetic transcripts with an expert model improve the distilled scorer, even surpassing the vendor that generated them? We further asked whether a newer vendor model scored this endpoint more accurately.

Methods From a double-blind, placebo-controlled Phase 3 major depressive disorder (MDD) trial, interviews after the expert model's training cutoff formed the held-out test set (N=206; 2,060 item ratings). Synthetic transcripts used no trial data: ~30,000 single-MADRS-item interview snippets, each generated by prompting Gemini for a synthetic patient persona at a target severity (0–6), few-shot–grounded on a small set of real de-identified example snippets. References were the central-rater MADRS scores of record: each interview was scored by one of 10 central raters, with no item double-rated or adjudicated. All models scored the same de-identified transcripts. Seven scoring models were compared: Llama 3.1 untuned; three vendor models (Gemini 2.5 Flash, Claude Opus 4.6, Claude Opus 4.8); a self-hosted Llama 3.1 distilled on Gemini-labeled synthetic MADRS transcripts; the same model on those transcripts relabeled by the expert model; and that expert model, trained on real pre-cutoff interviews [1]. Self-hosted arms were frozen; vendor arms ran on dated snapshots. Outcomes were item-level MAE and exact agreement (accuracy), with intraclass correlation coefficient (ICC, absolute agreement) on the 0–60 total.

Results On Question 1 (distillation), the untuned open-source model scored poorly (item MAE 1.10;

accuracy 25%), while the synthetic-distilled model approached the vendor band (item MAE 0.54 vs 0.45–0.53; accuracy 59% vs 56–61%); the real-data expert model was best on every metric (item MAE 0.28; accuracy 75%), with the highest ICC (0.98 vs ~0.95 for vendors). On Question 2 (expert relabeling), Gemini's labels — not its transcripts — were the bottleneck: the same transcripts relabeled by the expert model produced a scorer that surpassed Gemini itself (item MAE 0.40 vs 0.47; accuracy 68% vs 61%). Separately, a newer vendor model did not help: Claude Opus 4.8 scored no better than 4.6 (item MAE 0.53 vs 0.45; accuracy 56% vs 60%), counter to public-benchmark gains.

Conclusion In this single Phase 3 MDD dataset, synthetic-transcript distillation transferred vendor-LLM MADRS scoring into a frozen, self-hosted model, and expert-model relabeling let it exceed the generating vendor. The real-data expert model remained most accurate, though its training and evaluation labels share the same central-rater pool. For all the reasons above, locally validated self-hosted models are attractive for oversight: a privacy-preserving, reproducible second scorer flagging discrepancies in clinical trials. These single-trial, single-rater results need further study before generalizing, but remain a promising path — even where labeled interviews are scarce — toward oversight independent of vendor models.

Co-Authors

Adam Kolář[1], Miguel Amável Pinheiro[1], Alexander Deschamps[2], Todd M Solomon[2], Daniel R Karlin[2]

[1] Mira Analytics Inc.

[2] Definium Therapeutics

Keywords

Keywords
clinician-reported outcomes
large language models
knowledge distillation
synthetic data
MADRS

Guidelines I have read and understand the Poster Guidelines

Disclosures This work was supported by Definium Therapeutics. A. Kolar and M. A. Pinheiro are affiliated with MIRA Analytics and Definium Therapeutics; A. Kolar is also affiliated with the Faculty of Informatics, Masaryk University; A. Deschamps, T. M. Solomon, and D. R. Karlin are affiliated with Definium Therapeutics. One or more authors report potential conflicts, which are described in the program. Relevant financial relationships have been reviewed and mitigated to ensure the content is evidence-based and unbiased.

Related Tables and Supporting Materials <blank>