

LANGUAGE-MODEL RATINGS OF DEPRESSION REFLECT THE RATER MORE THAN THE PATIENT

PREPRINT

Baihan Lin^{1,2,3,4,5}

¹Department of Artificial Intelligence and Human Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA

²Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA

³Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, NY, USA

⁴Mental Illness Research, Education and Clinical Center, James J. Peters VA Medical Center, Bronx, NY, USA

⁵Berkman Klein Center for Internet & Society, Harvard University, Cambridge, MA, USA

Correspondence: baihan.lin@mssm.edu

ABSTRACT

Depression has no diagnostic blood test. Language models promise tireless, consistent assessment, but can accurate raters disagree about individuals? We pre-registered 880 language-model raters, crossing 11 open models with prompting and scoring choices, and applied them to 189 interviews against the eight-item Patient Health Questionnaire. Model choice explained 30.0% of summed-symptom score variance, stable participant differences 10.5%. Two randomly drawn raters with area under the receiver operating characteristic curve (AUC) ≥ 0.70 disagreed on screening decisions for 40% of participants, on average. Average over-rating governed how many were flagged, yet equal-capacity raters chose differently for about one participant in five. A locked analysis of 86 new interviews reproduced the main pre-registered findings. Exploratory recalibration with 40 labelled participants raised accuracy from about 60% to 75% and halved disagreement, leaving one participant in five decided differently. Calibration repaired much of the rater dependence without securing agreement about individuals.

Introduction

A depression rating can shape who is offered further assessment. If the interview stays the same, how much should the decision depend on the software reading it? Psychiatry has confronted a version of this question before. A 1971 study found that American and British psychiatrists disagreed after watching the same videotaped interviews¹. Agreement was subsequently found to be poor for most diagnostic categories². Psychiatry responded with explicit criteria and structured interviews^{3,4}. Half a century later, large language models (LLMs) promise scalable, inexpensive, repeatable ratings from clinical conversations^{5–8}. Early evaluations are encouraging: open models achieved clinically acceptable total-score error on structured psychiatric interviews⁹, reached intraclass correlations near 0.75 with trained raters of psychosis-risk and schizophrenia-spectrum interviews^{10,11}, and recovered self-reported depression severity from transcripts^{12,13}. Psychotherapy dialogue also supports alliance inference^{14,15} and topical analysis¹⁶.

But a model becomes a rater only through choices. Someone must choose the model and word the request, perhaps as a questionnaire, in a psychiatrist’s voice or as the patient answering about themselves. They must choose digits or letters and their order, decide whether the model reads the patient’s words or the whole conversation, and turn its answer into a score. These choices can define plausible raters, although evaluations rarely show the alternatives. Outputs shift with prompt formatting^{17,18}, answer order and labels^{19,20}, inference settings²¹, and even supposedly deterministic settings²². Such choices reverse social-science conclusions²³, alter depression inference²⁴, change diagnoses when clinical cases are reworded²⁵, and shift severity judgements across languages²⁶. High self-consistency makes reliability easier to establish than validity^{27–29}. For a clinic, the uncertainty concerns whose assessment changes when one plausible rater replaces another, and how much local labels can repair.

Generalizability theory asks whether differences between people survive changes in how they are measured. It separates contributions from the person and the measurement facets^{30,31}, with dependable scores requiring the person's contribution to dominate error. Defensible configurations define a multiverse^{32,33}. Disagreement among similarly accurate configurations is predictive multiplicity^{34,35}. Even predictors that agree on average can induce different decisions, motivating recalibration for downstream decisions³⁶. Calibration, long a weak point of clinical prediction³⁷, connects measurement to action. Together, these ideas distinguish whether a rater orders people well, whether its threshold flags an appropriate number, and whether another defensible rater selects the same people.

We crossed 11 open models, five wordings, two conversation views, two answer-reading rules and four score constructions (eight symptom ratings in three formats, or one overall rating), creating 880 raters for 189 interviews with the eight-item Patient Health Questionnaire (PHQ-8) as criterion³⁸. Analysis code was fixed at pre-registration before comparison with the criterion^{39,40}. The locked procedure was then applied once to 86 new interviews from a related corpus, with an autonomous interviewer and automatic transcription. The rater outweighed the patient in score variance, and accurate rankings concealed divergent decisions. Offsets governed how many were flagged, but equal-capacity raters still chose different people. Even answer letters mattered. We then examined how far local recalibration could repair these differences: decisions improved, while disagreement about individuals remained.

Results

The same interview, opposite verdicts

One participant reported mild symptoms (PHQ-8 5), below the screening threshold of 10, yet 50% of 880 raters screened them positive (Fig. 1b). No participant received a unanimous verdict, and 91% split raters between 20% and 80% (Fig. 1c). Some dissent becomes likelier as raters accumulate: from 40% of participants with two $AUC \geq 0.70$ raters to 94% with ten (Extended Data Fig. 7a).

We therefore measured the chance that two distinct, randomly drawn raters disagree: 44% across all raters, 40% among the 540 with $AUC \geq 0.70$ (95% participant interval 39–42%, raters fixed), and 12% among the 75 with AUC, sensitivity and specificity all at least 0.70. Even this strict set gave opposite decisions to 53% of participants (Extended Data Fig. 7b). An outcome-independent, clinically defensible core using instrument or psychiatrist wording, digits, expected values and symptom sums (44 raters, AUC 0.61–0.87) disagreed on 31% (Supplementary Information). Nor were only borderline cases affected (Fig. 1d). Participants reporting few or no symptoms ($PHQ-8 \leq 4$) were flagged by 35% of $AUC \geq 0.70$ raters, which disagreed about them as much as about anyone.

Accuracy does not decide who is flagged

Many raters met published standards of accuracy. AUC asks how well scores rank participants with a positive screen above those without one, without choosing a threshold. AUC for $PHQ-8 \geq 10$ ranged from 0.34 to 0.88 (median 0.75), with 39% below 0.70 (Fig. 2a; Extended Data Fig. 6). Between-rater AUC variance was 10.6 times its sampling variance: rater choice mattered more than participant sampling. Decisions told another story. Ratifiers flagged 0–100% of participants against 30.2% self-reported prevalence; model medians ranged from 14% to 99% (Fig. 2b).

The rater's offset, its mean over-rating of PHQ-8, governed the share flagged along a normal-ogive curve ($R^2 = 0.91$; Fig. 2c). Applied unchanged to the new interviews, that curve explained 0.90 of the variance. The relation is partly built in: offset and threshold crossings summarise the same scores. Nevertheless, location nearly exhausted what mattered at the threshold, and AUC cannot see location. Near the curve's centre, each point of over-rating flagged 13 additional people per 100. A rater behaves like a bathroom scale with a misplaced zero: it may rank people well, but its offset decides who crosses the line. Qwen2.5 7B and Mistral 7B, in their typical configurations, had almost identical AUCs (0.81, 0.80). Per 100 people like those studied, the first would flag 24 and miss 14 of 30 positive screens; the second would flag 80 and miss 1 (Fig. 2d). $AUC \geq 0.70$ raters differing by at most 0.02 in AUC differed by a median 28 percentage points in the share flagged. Across all raters, sensitivity quartiles were 47–100%, specificity 5–84%.

The rater outweighs the patient

Each interview was held fixed while the rater changed. Stable participant differences, the person main effect, explained 10.5% of summed-symptom score variance. Model choice explained 30.0%, or 2.8 times as much (Fig. 3a). Model-by-format interaction explained 15.8%: formats shifted models differently. Participant-by-model interaction explained 8.1%: models scored participants differently relative to one another, reflecting rankings or score scales. Crediting every participant-related component to the participant still gave 31%, against 69% for the rater alone. The ordering survived restriction to digits (model-to-person ratio 2.4), one neutral wording, format and view (3.0), and omission of

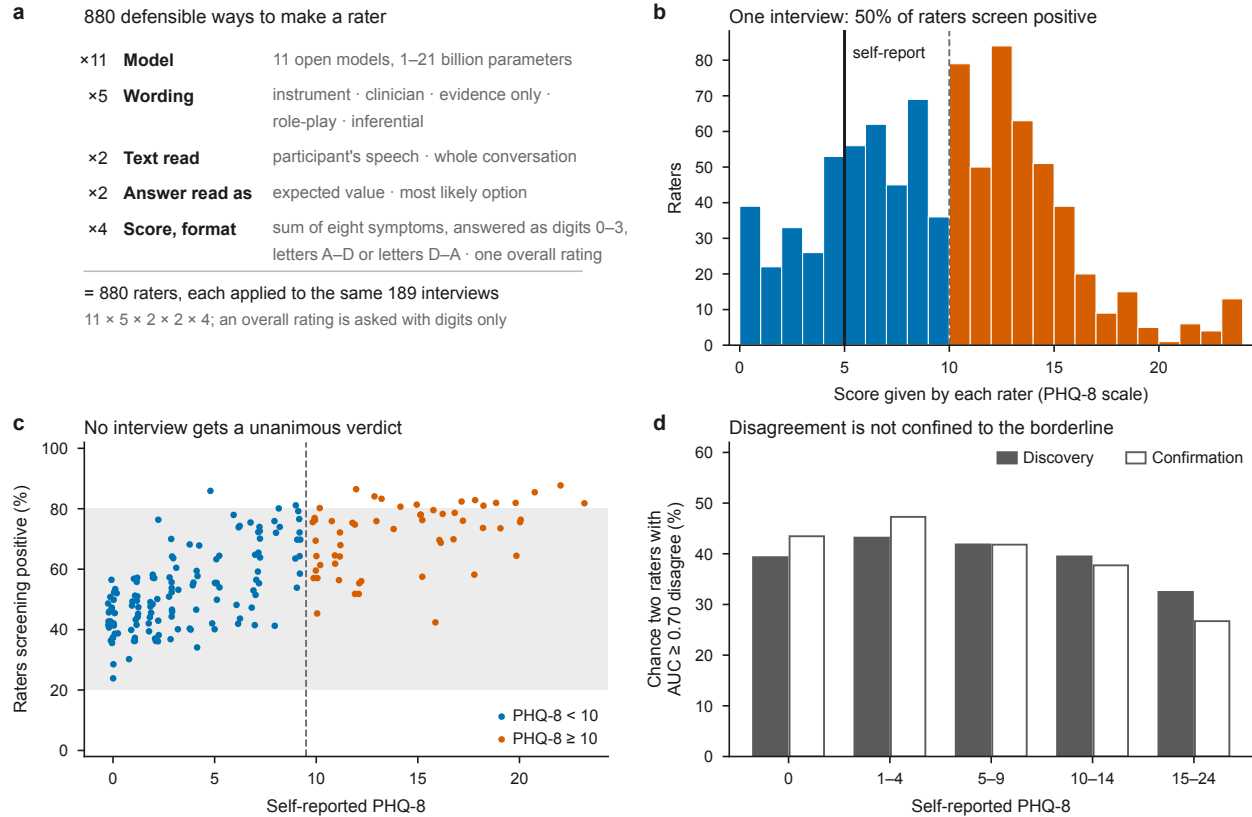


Figure 1: The same interview, opposite verdicts. **a**, The choices that make a rater; each of the 880 combinations was applied to every interview. **b**, Scores given to one participant by all 880 raters, on the PHQ-8 scale; black line, the participant's self-reported PHQ-8; dashed line, the screening threshold of 10; orange, scores of 10 or more. **c**, For every participant, the share of raters screening them positive, against their self-reported PHQ-8 (jittered horizontally); grey band, 20–80% of raters; dashed line, between self-reported totals of 9 and 10. **d**, The chance that two different raters with $AUC \geq 0.70$, drawn at random, give a participant opposite decisions, by self-reported severity; dark grey bars, discovery; white bars, confirmation.

each developer's models (1.9–6.3). Dependability was 0.11 for one model, wording, format and view. Averaging 11 models, five wordings and three formats raised it to only 0.53, or 0.55 with both views (Fig. 3b; Extended Data Fig. 3). Participants' mean scores across the 330 expected-value item-sum raters explained 0.53 of PHQ-8 variance (Fig. 3c). Even the average rater captured only part of the questionnaire.

Small choices move the threshold

Apparent technicalities moved the threshold (Fig. 4). Rating only explicitly reported symptoms, an assessment-policy change from the other four instructions, lowered the median share flagged by 27 percentage points (10 of 11 models; Fig. 4a). Reversing answer letters so A meant 'nearly every day' instead of 'not at all', with no clinical change, moved it by a median 21 points and up to 85 (Fig. 4b). The interview and the symptom descriptions were unchanged. With digits, symptom sums over-rated relative to direct totals in 8 of 11 models in both sets (Fig. 4c), despite similar discrimination (median AUC 0.78, 0.78). For symptom sums, digits outperformed letters (0.78 versus 0.71, or 0.66 reversed). Larger models ranked better (Spearman correlation 0.82; Fig. 4d), without better calibration or specificity. Their configurations' decisions converged, but even the largest models varied in share flagged by tens of percentage points (Fig. 4e).

Deployment choices also changed decisions. Running Qwen2.5 7B as a 4-bit file with another inference library changed a median 21% of decisions (maximum 35%) under identical configurations: a comparison of deployment stacks, not compression alone. Automatic instead of manual transcripts changed 5% (maximum 37%). Increasing model size with other choices fixed changed 21–25% within Qwen2.5 and 73% from Llama 3.2 3B to Llama 3.1 8B.

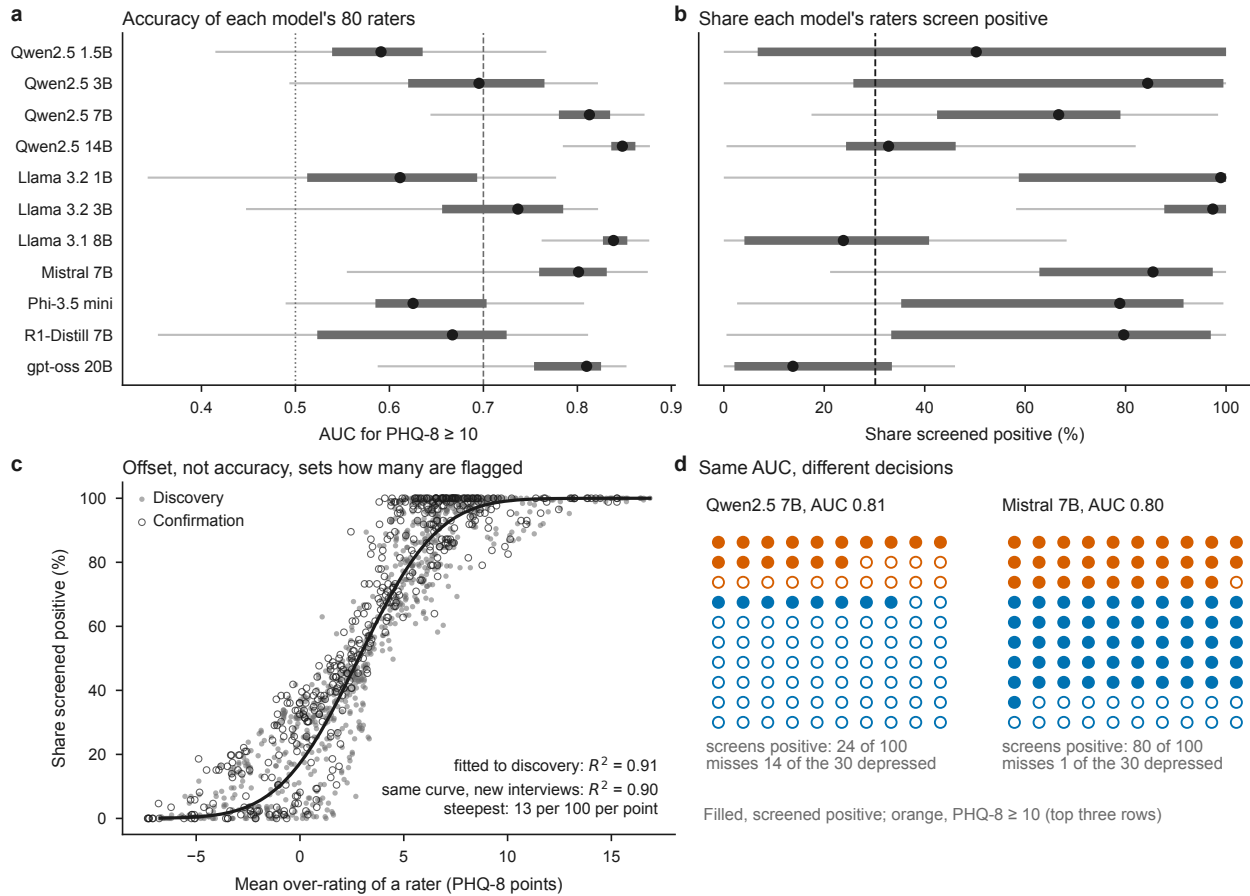


Figure 2: Accuracy does not decide who is flagged. **a**, AUC of each model's 80 raters: dot, median; thick line, interquartile range; thin line, range. Dotted line, chance; dashed line, 0.70. **b**, Share of participants that each model's raters screen positive; dashed line, self-reported prevalence. **c**, Share screened positive against the rater's mean over-rating of the PHQ-8, for every rater in discovery (filled) and confirmation (open); line, normal-ogive curve fitted to the discovery raters, with its R^2 in discovery and, applied unchanged, in confirmation. **d**, Decisions for 100 people with the sample's prevalence by the typical configuration (median AUC) of two models with nearly the same AUC; shown is the pair of models whose typical configurations differ in AUC by at most 0.02 and differ most in the share screened positive.

What the raters read

At every model size, scores showed little preferential association with PHQ-8 over PTSD severity (Fig. 4f). Only 4% of raters correlated significantly more strongly with PHQ-8 (4.3% by paired bootstrap). The criteria themselves correlated strongly ($\rho = 0.84$), preventing conclusions about diagnostic separation. Below the screening threshold, higher PTSD symptoms predicted more flags than PHQ-8 alone predicted (partial Spearman 0.36; confirmation 0.23, $P = 0.049$; Fig. 5a). Talkativeness showed the same association (0.25, 0.26; Fig. 5b). These associations warrant specificity tests without identifying what drives ratings. Interviewer speech alone, containing no participant report, yielded median model AUC 0.59 (Fig. 5c): above chance, as follow-up questions predict⁴¹, but below the predicted 0.60. Group conclusions were also fragile. The self-reported female–male difference was small (standardised difference 0.20, $P = 0.18$), yet 16% of raters reversed its sign and 14% reported a nominally significant difference absent from self-reports (Fig. 5d).

Checks that need no labels

These raters were consistent, as evaluations commonly report (median Cronbach's α 0.91). Internal consistency was unrelated to validity across raters (Spearman correlation with AUC 0.02; confirmation -0.06 ; Fig. 6a), although informative within models (median correlation 0.51, 0.34). A model can distribute one global impression coherently

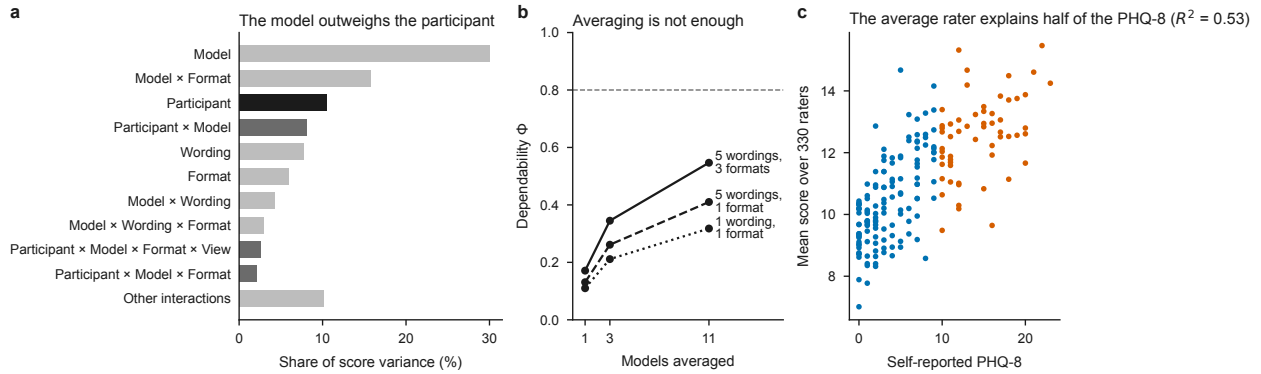


Figure 3: The rater outweighs the patient. **a**, Shares of the variance of the expected item sum in the generalizability analysis of the 330 expected-value item-sum raters (discovery); black, participant main effect; dark grey, interactions involving the participant; light grey, components of the rater alone. **b**, Dependability of a score averaged over models, for one or all wordings and formats, with both views of the conversation averaged; dashed line, 0.80. **c**, Participants' mean score over the same 330 raters against their self-reported PHQ-8; orange, PHQ-8 of 10 or more.

across symptoms whether that impression is right or wrong. Agreement across rewordings tracked validity both across raters and within models (0.71, 0.67; confirmation 0.67, 0.59; Fig. 6b). Agreement between independent runs sampled at temperature 1 also tracked validity (0.76 across raters).

Before confirmation, we specified a label-free rule choosing each model's configuration with greatest agreement across rewordings. As predicted, it beat the model median in 9 of 11 confirmation models (11 of 11 in discovery). But the rule chose only expected-value configurations, whereas its comparator included cruder most-likely-option scores. Against eligible configurations' median, it won in 8 of 11 models (mean AUC gain +0.035). Against the digit-answer median it won in 6 (−0.005), and against a fixed neutral configuration (instrument wording, digits, participant speech, expected values) in 5 (−0.004; Fig. 6c). Avoiding letters explained most of the gain. Rewording robustness can flag fragile, often invalid configurations, but found none better than a sensible default.

How many versus who

Disagreement can concern how many people to flag or which people. An identity separates the minimum disagreement forced by unequal positive rates from the remainder: total disagreement equals the absolute rate difference plus twice the smaller one-sided disagreement. At the screening threshold of 10, unequal rates dominated. Among $AUC \geq 0.70$ raters, 40% disagreement comprised 38 percentage points forced by rates and 2.3 beyond them (confirmation 40% total, 2.1 excess; Fig. 7a). All raters, the defensible core and one fixed configuration per model showed the same pattern. Decisions were nearly nested: lenient raters flagged almost everyone strict raters flagged, then added others.

Nested decisions need not imply shared severity rankings. To separate who is selected from how many, we gave each rater the same number of places to fill. Selecting equal numbers, matching self-reported prevalence, expected-value raters with $AUC \geq 0.70$ still chose differently for 18% of participants (95% interval 15–21%), replacing 30% of each other's selections (confirmation 20%, 30%; Fig. 7b). Disagreement was 21% within the defensible core (21% confirmation) and 25% with one fixed configuration per model (22%), without accuracy-based selection. Selecting the top tenth reduced disagreement to 9% because fewer people were selected, while still replacing 46% of selections. Raters disagreed with the criterion on 24%, more than with one another, indicating shared departures from PHQ-8. The two forms of disagreement also affected different people. At the threshold of 10, participants reporting few or no symptoms contributed almost half (48%). At equal capacity, their exchange probability was only 9%, compared with 27% at PHQ-8 10–14 and 25% at 15 or above (Supplementary Information).

Changing one choice isolates its consequences (Fig. 7c). Changing models altered a median 35% of decisions at 10 and 25% at equal capacity. Reversing letters altered 20% and 21%: clinically meaningless relabelling changed both how many and who. The psychiatrist persona changed least among wordings (7%, 8%). Full dialogue rather than participant speech changed 17% of equal-capacity selections. The unmentioned-symptom rule and direct-total scoring chiefly moved thresholds: 21% versus 14%, and 24% versus 15%, respectively.

The dominance of offsets suggested a practical repair: use local labels to reset the thresholds. This repaired much of the disagreement. Across 200 random splits, 40 labelled participants admitted raters with $AUC \geq 0.70$ (mean 538 of 880) and recalibrated them by five rules: subtracting mean over-rating, linear regression, logistic regression, the attainable

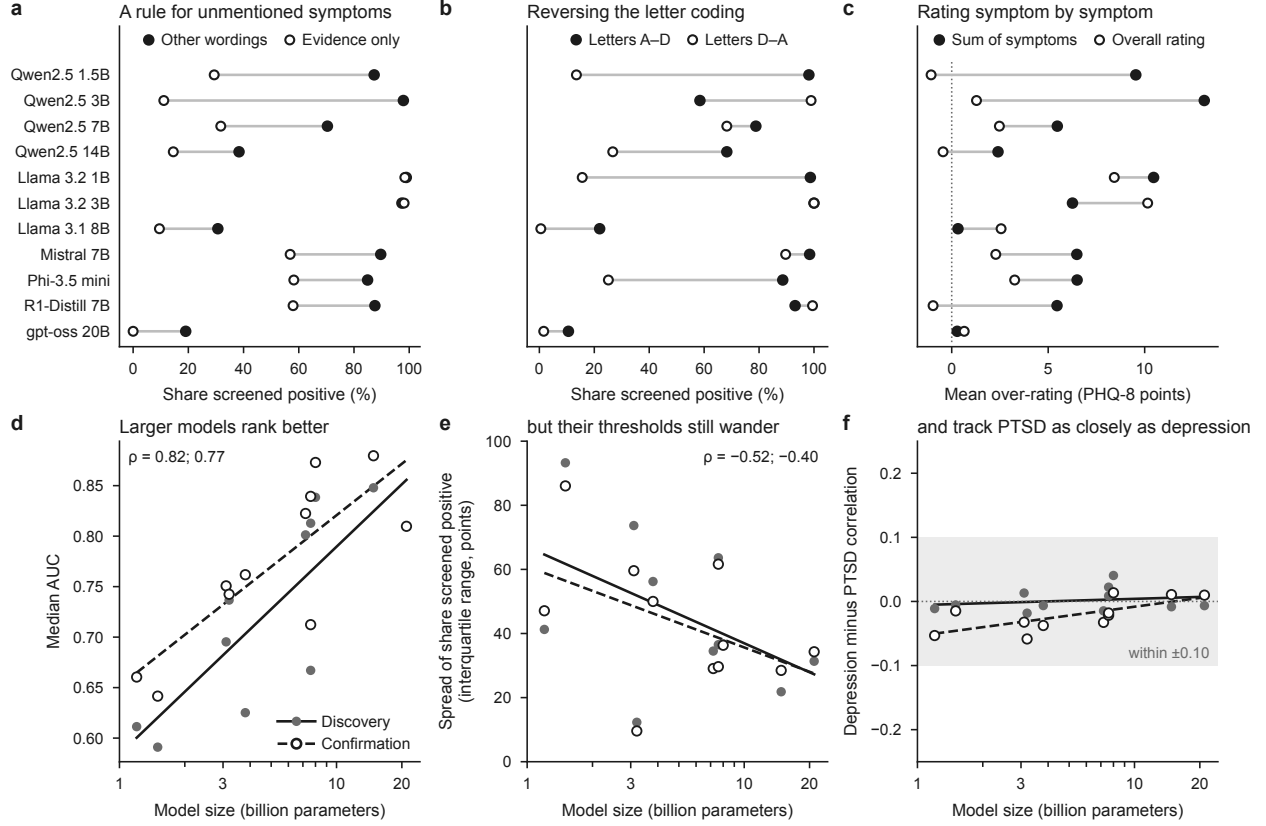


Figure 4: Small choices move the threshold, and size buys ranking but not calibration or specificity. **a–c**, For each model, the median share screened positive with the other wordings (filled) and with the instruction to rate unmentioned symptoms as absent (open) (**a**), and with answer letters A–D meaning ‘not at all’ to ‘nearly every day’ (filled) and the reverse (open) (**b**); and the median over-rating when symptoms are summed (filled) or the total is rated directly (open) (**c**). **d–f**, Against model size, one point per model: median AUC (**d**), interquartile range across a model’s raters of the share screened positive (**e**), and the median difference between a rater’s correlation with the PHQ-8 and with PTSD severity (**f**); grey band, differences within ± 0.10 . Lines, least-squares fits against log size; filled markers and solid lines, discovery; open markers and dashed lines, confirmation; ρ , Spearman correlations across models.

threshold closest to labelled prevalence, or the highest threshold retaining at least 80% of labelled positives. Evaluation used the remaining participants, whose labels informed no step. Prevalence matching raised accuracy from 61% to 75% (confirmation 59% to 75%) and halved disagreement, from 40% to 20% (40% to 21%; Fig. 7d). Every rule improved accuracy.

Lower disagreement sometimes meant more missed cases: linear recalibration gave 16% disagreement and median sensitivity 42%. Retaining 80% of labelled positives gave held-out sensitivity 78% and disagreement 27%. Each rule sets a cut-off on the rater’s score. Increasing fits preserve rankings, as did all but 14 of 107,540 linear fits. Residual disagreement after prevalence matching remained near one in five, resembling equal-capacity disagreement with all labels available. It persisted with 20 or 80 labels (22%, 20%), fixed configurations (24%) and whole-cohort admission (19%; Extended Data Fig. 8).

Prevalence matching brought the share flagged within a median 7 percentage points of self-reported prevalence (confirmation 8), with a central 95% split-to-split range of 3–19 points. Score ties sometimes prevented matching labelled prevalence within 5 points (achieved in 88% of cases). For prevalence estimation, the 40 labels alone gave intervals nearly as narrow as prediction-powered estimates using ratings⁴² (25 versus 27 percentage points), with similar coverage (Extended Data Fig. 4c).

Calibration did not always travel. We transferred discovery calibrations unchanged for configurations with discovery AUC ≥ 0.70 , from manual transcripts and operator-controlled interviews to automatic transcripts and autonomous interviews. Subtracting discovery offsets over-corrected: median share flagged fell from 51% to 23%, against 34%

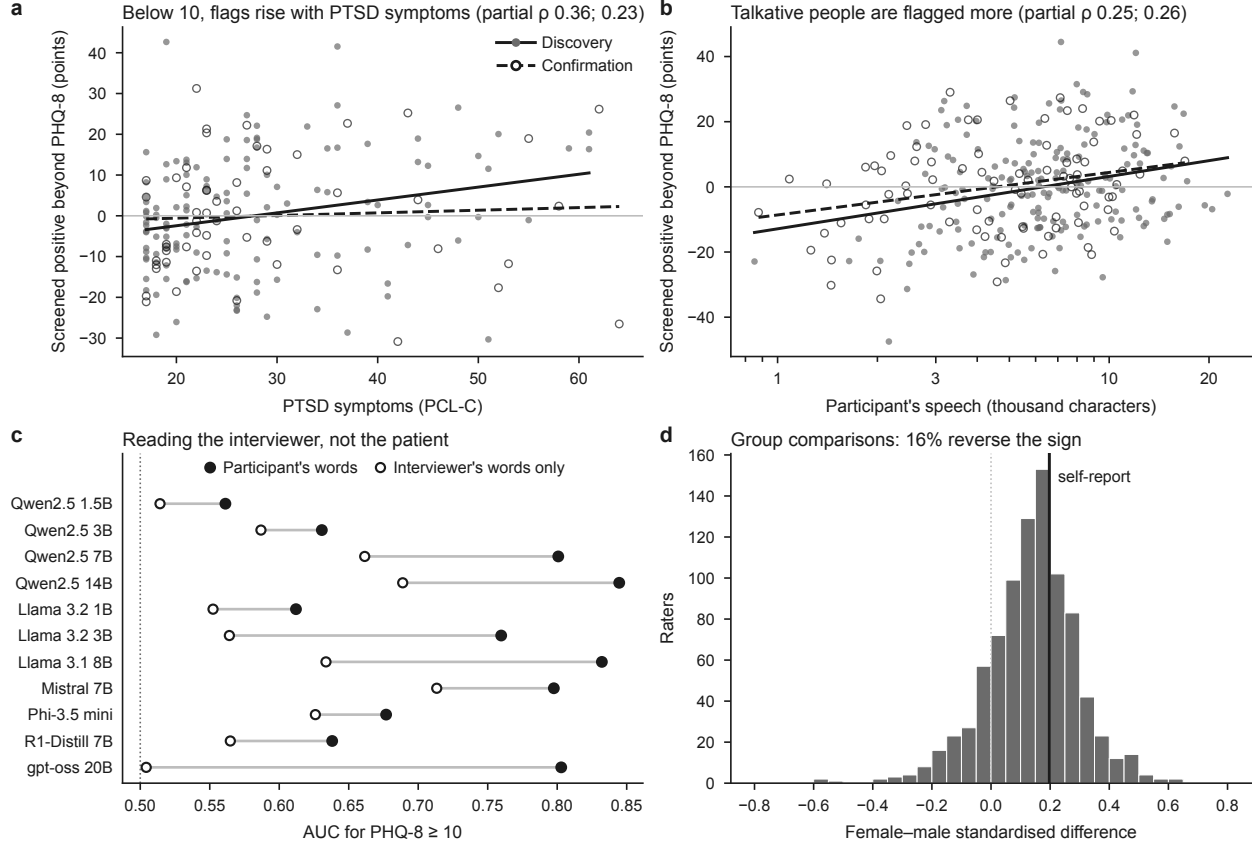


Figure 5: **What the raters read.** **a**, Among participants below the screening threshold, the share of raters with $\text{AUC} \geq 0.70$ screening them positive beyond what their PHQ-8 predicts (residual of a linear fit, percentage points), against PTSD symptoms (PCL-C). **b**, The same for all participants, against the amount of speech (logarithmic scale). Points and lines show residuals and linear fits on the original scales; the statistics in the panel titles are partial Spearman correlations on ranks (discovery; confirmation). Filled markers and solid lines, discovery; open markers and dashed lines, confirmation. **c**, AUC of each model from the participant's speech (filled) and from the interviewer's speech alone (open), median over wordings and formats of the expected-value item sum (discovery). **d**, Female-male standardised difference estimated by every rater (discovery); black line, the difference in self-reported PHQ-8.

prevalence. Prevalence-matched thresholds transferred reasonably (37% flagged, median sensitivity 72%). The best discovery configuration (Qwen2.5 14B), with its discovery threshold, flagged 43%, with 79% accuracy and 83% sensitivity. In this single related transfer, transportability depended on the calibration and required checking.

Could disagreement guide deferral? Pooling and abstention have been proposed for medical predictive multiplicity⁴³, but selective prediction can fail particular groups⁴⁴. We compared rules on the same held-out participants at matched coverage (Extended Data Fig. 9). Majority accuracy was 78%, versus 75% for a single recalibrated rater (confirmation 80%, 75%). Requiring 80% consensus retained 71% of participants, with majority accuracy 86% (89%). Single raters achieved 83% on those participants (84%), or 82% (82%) when deferring their own closest calls at equal coverage. Consensus selection outperformed a rater's own margin, but most apparent gain came from excluding difficult participants, disproportionately positive screens: 24 of 45 positives remained, versus 82 of 104 negatives. Majority sensitivity changed from 58% overall to 56% among retained participants (confirmation 65% to 70%). These sensitivities describe different groups. Deferred participants are neither detected nor missed until reviewed. Without costs for review and missed cases, these comparisons do not define a triage policy.

The regularities held in new interviews

The locked procedure was applied once to 86 E-DAIC interviews (sessions 600–718) from the same laboratory, with an autonomous interviewer and automatic transcription (440 participant-speech raters). AUC ranged from 0.34 to 0.91 (median 0.78), and share flagged from 0% to 100%, against 33.7% prevalence (Extended Data Fig. 5). $\text{AUC} \geq 0.70$

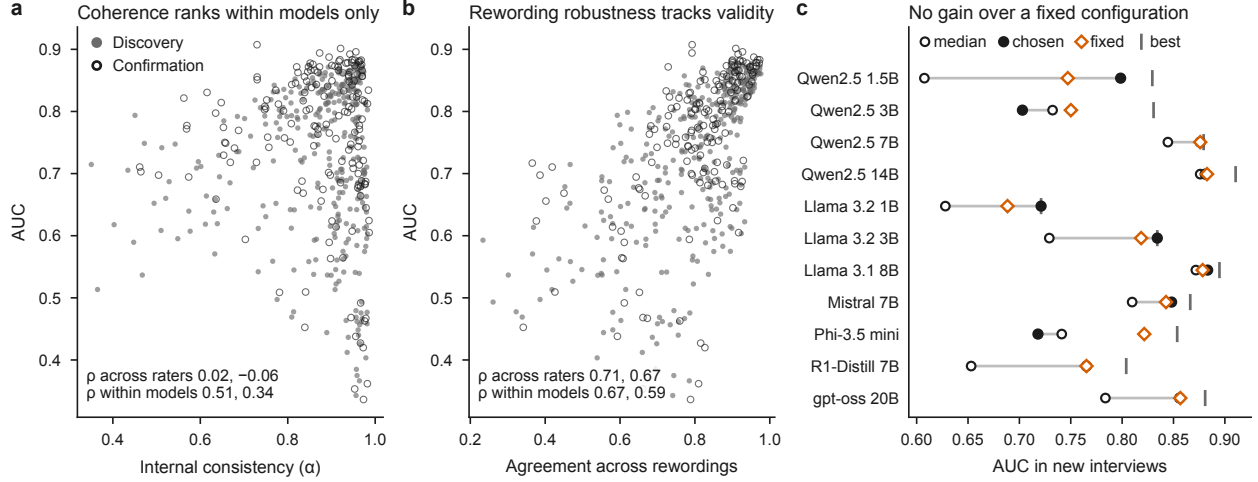


Figure 6: **Checks that need no labels.** **a,b**, AUC of each item-sum rater against its internal consistency (**a**) and against its agreement with the same model and format under the other wordings (**b**); filled, discovery; open, confirmation; Spearman correlations across all raters and medians of the correlations within models, for discovery and then confirmation. **c**, In the confirmation set, the AUC of each model’s median item-sum configuration, of the configuration chosen without labels by agreement across rewordings, of a fixed neutral configuration (the questionnaire’s wording, digit answers, the participant’s speech, expected value) and of its best configuration; the rule was specified before the confirmation analysis.

raters disagreed on 40%. Participant differences explained 13.0% of variance, model choice 26.2%. Of 6 evaluable pre-registered predictions, 5 were supported. As in discovery, 40-label calibration did not halve prevalence spread. Discovery supported 5 of 7 predictions, also falling short on interviewer speech. All 9 of 9 discovery regularities met criteria specified before confirmation (Extended Data Table 1).

Discussion

A language model asked to rate depression does not report a property of the patient alone. In these interviews, model choice explained about three times more summed-symptom score variation than stable differences between participants. Raters that ranked participants well still disagreed about many of them, and how many people were flagged depended on the offset rather than the accuracy. Psychiatry spent decades responding to unreliable diagnosis with explicit criteria and structured interviews^{1–4}. Automation promises relief from variation between human raters. Here, the rater returns as a configuration: a model, an instruction and a choice of letters, often undocumented. Choices made when building an instrument can become differences in how it judges a person.

Earlier findings of accurate, consistent LLM ratings^{9–11} remain compatible with ours: they usually examined few configurations, and our best were accurate too. Ranking versus calibration³⁷, predictive multiplicity^{34,35} and downstream decision disagreement³⁶ are established distinctions. The multiverse makes these distinctions tangible for one clinical decision. It measures how much a decision about a person depends on choices that an evaluation usually leaves in the background, and how much of that dependence a simple local repair removes. Clinics act at thresholds, where AUC is blind to offsets. Internal consistency did not rank configurations across models, and rewording robustness identified fragility without improving on a neutral default. For prevalence estimation, ratings added little precision beyond the 40 questionnaire labels alone in this cohort. The value of automation therefore needs to be established against what the same local measurement effort can achieve without it. Alliance estimates also inform classification⁴⁵ and conversational recommendations^{46–48}.

Calibration exposed the distinction between agreement about populations and agreement about people. A few dozen local labels repaired much of the disagreement: decisions became more accurate and more alike, and prevalence estimates became better calibrated. But agreement about how many is not agreement about who. An increasing recalibration leaves a rater’s ranking untouched, and different configurations ranked participants differently enough to make opposite selection decisions for about one participant in five even when selecting equal numbers. Resetting a scale’s zero cannot reconcile two scales that disagree about who is heavier. In a service using these ratings to allocate a fixed number of assessment appointments, the implication would be different names on a referral list of the same length.

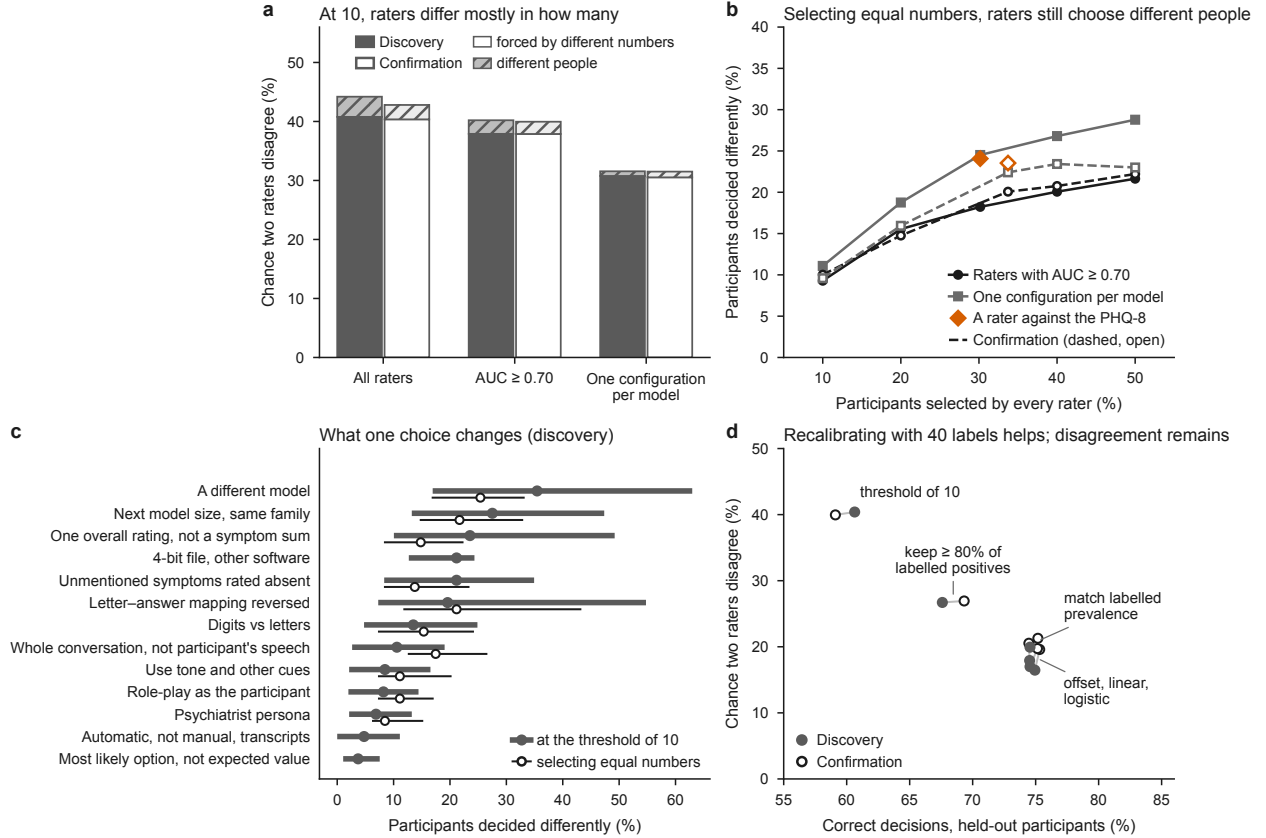


Figure 7: How many versus who. **a**, The chance that two different raters disagree about a participant at the threshold of 10, split by the identity $P(A \neq B) = |p_A - p_B| + 2 \min\{P(A=1, B=0), P(A=0, B=1)\}$ into the part forced by their different positive rates (solid) and the rest, from flagging different people (hatched), for all raters, raters with $AUC \geq 0.70$ and one fixed configuration per model; dark grey, discovery; white, confirmation. **b**, Participants decided differently when every rater selects the same share of participants, its highest-scoring ones, by two expected-value raters with $AUC \geq 0.70$ or two fixed configurations (one per model); diamonds, participants decided differently by a rater and the self-reported criterion when the rater selects as many as screen positive. Filled markers and solid lines, discovery; open markers and dashed lines, confirmation. **c**, Participants decided differently by configurations that differ in one choice only (discovery; changes of wording are relative to the questionnaire's own wording): dot, median over pairs of configurations; bar, interquartile range; filled, at the threshold of 10; open, selecting as many as screen positive (expected-value scores only); deployment-stack and transcript comparisons are for decisions only. **d**, For raters admitted by their AUC on 40 labelled participants and recalibrated with the same 40 labels, the chance that two raters disagree against the share of correct decisions, both on the held-out participants (means over 200 splits); filled, discovery; open, confirmation. Exploratory analyses.

Pooling improved accuracy; consensus deferral improved it further at matched coverage, while disproportionately excluding positive screens. Decision-focused reconciliation³⁶ may narrow the residual. Until demonstrated, individual agreement must be measured and calibration checked locally. Agreement itself is also insufficient: raters shared departures from PHQ-8, and choosing a threshold trades missed positive screens against unnecessary flags. The clinical objective must therefore be specified alongside the calibration rule.

The ratings also carried information beyond depression, raising the question of what an apparently successful score measures. They tracked PTSD as closely as PHQ-8. A preprint on these corpora found symptom geometry closer to questionnaire wording than self-report⁴⁹. Below the screening threshold, participants with more trauma symptoms were flagged more often than their PHQ-8 predicted. Greater amounts of speech also predicted additional flags beyond PHQ-8. Specificity remains unresolved where depression and PTSD co-occur, including trauma-exposed and veteran populations. Ambient scribes now record entire clinical visits⁵⁰, making the clinician's contribution to model inputs a practical concern beyond research interviews.

Box 1 | Five checks before a language-model rating informs care

Report the range, not one number. Rate with several defensible configurations and report how often their decisions disagree, including among configurations that pass checks of accuracy, sensitivity and specificity, and whether they disagree about how many people to flag or about which ones (Figs 1, 2 and 7).

Fix a neutral configuration, and treat fragility as an alarm. Choose the wording, answer format and scoring in advance, for example the questionnaire's own wording with digit answers. Reword the request: a rating that changes with the wording is fragile. Internal consistency did not rank configurations across models (Fig. 6).

Calibrate locally, check transfer, and measure what remains. A few dozen local patients who completed a validated questionnaire can set the threshold and give honest estimates of prevalence. Check whether a calibration carries over to a new setting rather than assuming it. A recalibration that preserves a configuration's ranking changes how many people are flagged, not whom it ranks highest, so measure the remaining disagreement about individuals (Figs 2c and 7).

Ask what else it reads. Test specificity against neighbouring conditions such as PTSD with criteria that can separate them, and whether the clinician's words drive the rating (Fig. 5).

Freeze the whole stack. Treat any change of model, precision, inference software, transcription or wording as a new instrument, and revalidate it (Fig. 4).

If the configuration is the instrument, changing a component changes the instrument. Changes in model precision and inference software, model size, instructions and transcription each altered substantial numbers of decisions. Five practices follow (Box 1). Report decision ranges across defensible configurations, extending multiverse and many-analyst approaches^{32,51,52} to clinical LLM reporting⁵³. Fix a neutral configuration and test rewording robustness. Calibrate against a validated criterion in a few dozen local patients, check transfer and report residual individual disagreement. Test specificity against neighbouring conditions. Freeze the complete stack, especially when following a patient over time. Otherwise, a future change in a patient's score could reflect a changed instrument as well as a changed condition. Longitudinal validation would need to distinguish these possibilities.

Several limits define the scope. Corpus licences required local processing, restricting us to open models up to 21 billion parameters. We scored normalised answer probabilities after a fixed cue, without generated reasoning. Frontier models used as intended may differ. Size improved ranking and narrowed threshold variation without eliminating it; larger models may reduce these effects, but disappearance cannot be assumed. Our finite, deliberately broad grid makes inclusion consequential, motivating stricter sets, fixed configurations and single-choice comparisons. PHQ-8 was a pre-interview self-report screening criterion, not a diagnosis, although rater disagreement requires no criterion. No clinicians rated these transcripts, precluding direct comparison with human disagreement.

Both English-language sets followed one laboratory's protocol. Confirmation was small, with few women, and changed participants, interviewer and transcription, not clinical setting. Calibration transfer was therefore one related comparison. Analyses of how many versus who, recalibration, transfer, deferral and fairer label-free comparisons were added after confirmation and remain exploratory. Repeated splits of small cohorts are not independent samples. Finally, frozen code corrected termination for only some direct totals, and batching affected inference reproducibility. Both changed scores by fractions of a point in the re-scored sample (Methods).

Open questions extend beyond depression. Do frontier models escape the rater effect? Does disagreement persist across clinically equivalent paraphrases, and how does it compare with clinicians reading identical transcripts? How large are configuration effects relative to treatment-related change? Can raters be reconciled, or designed, to agree about individuals while remaining accurate against an independent clinical assessment? The same decomposition could inform assessments of essays, job applicants and online speech wherever a model's judgement becomes a measurement of a person.

A questionnaire makes its eight questions and cut-off visible, so readers can inspect and contest the terms on which it judges a person. A language-model rater can conceal equally consequential choices in software that appears merely to read the interview. Here, those choices shaped how many people were flagged and which people they were. Calibration made the raters more nearly right on average, yet left them divided about individuals. Clinical care meets people one at a time. When changing how we ask changes whom we identify, the instrument becomes part of the explanation. Responsible measurement makes that contribution visible, so that a judgement about a person can be understood and questioned by the person who must live with it.

Online Methods

Data. The discovery set was the Distress Analysis Interview Corpus – Wizard of Oz (DAIC-WOZ)^{54,55}: 189 semi-structured interviews of adults, conducted by an animated virtual interviewer that a human operator controlled from

another room, with manual transcripts of the interviewer’s and the participant’s turns. Before the interview, participants completed the PHQ-8, which scores eight symptoms over the preceding two weeks from 0 (not at all) to 3 (nearly every day), for a total of 0–24; a total of 10 or more is the standard screening threshold³⁸. A participant-level meta-analysis also supports this threshold for the related PHQ-9⁵⁶. A participant screened positive if their PHQ-8 total was 10 or more; this registered definition differs from the binary flag distributed with the corpora for one discovery and five confirmation participants (Supplementary Information). Totals were available for all 189 participants and item scores for 142 (complete for 141; the challenge training and development partitions⁵⁷). The same participants’ PTSD Checklist – Civilian Version (PCL-C) severity was available from the extended corpus (E-DAIC)⁵⁸. The confirmation set comprised the 86 E-DAIC interviews (sessions 600–718) conducted by a fully autonomous version of the virtual interviewer, with automatic speech recognition (ASR) transcripts of the participant’s speech and PHQ-8 and PCL-C scores (PCL-C missing for two). In the discovery set, 30.2% of participants screened positive (57 of 189; 87 women and 102 men); in the confirmation set, 33.7% (29 of 86; 18 women and 68 men). The corpora were used under their end-user licence agreements; all processing ran on a local workstation, and no transcript text was sent to any external service.

Models. We used 11 open-weight instruction-tuned models from six developers, chosen because their weights were already on the workstation: Qwen2.5-Instruct with 1.5, 3, 7 and 14 billion parameters⁵⁹; Llama-3.2-Instruct 1B and 3B and Llama-3.1-8B-Instruct⁶⁰; Mistral-7B-Instruct-v0.3⁶¹; Phi-3.5-mini-instruct⁶²; DeepSeek-R1-Distill-Qwen-7B⁶³; and gpt-oss-20b⁶⁴. Qwen2.5 1.5B, 3B and 7B ran in bfloat16 with the Transformers library⁶⁵ on the Apple GPU; the others ran as 8-bit (Llama-3.2-1B) or 4-bit quantized GGUF files with llama.cpp. For the reasoning model the reasoning block was left empty, and gpt-oss-20b answered in its final channel at low reasoning effort. As a check on the deployment stack, Qwen2.5-7B was also run as a 4-bit GGUF file with llama.cpp.

Prompts. Each prompt, sent in the model’s own chat format, began with a one-sentence description of the transcript and the transcript itself, followed by a task instruction (wording), a description of the response options (format) and one PHQ-8 item. The five wordings represented distinct instructions: *instrument*, complete the PHQ-8 on the participant’s behalf using the interview as the only source; *clinician*, an experienced psychiatrist rating the participant; *evidence*, rate only symptoms the participant explicitly reports and rate undiscussed symptoms as absent; *role-play*, answer as the participant would about themselves; and *inferential*, use every cue, including tone, word choice and avoidance. The evidence, role-play and inferential wordings change the assessment policy as well as the words; the clinician wording changes only who is said to rate. The three formats were digits 0–3, letters A–D with A meaning ‘not at all’, and letters A–D with A meaning ‘nearly every day’, each with the PHQ-8 anchor labels. Items used the PHQ-8 wording with the standard two-week stem. For the direct total, the wording was followed by a request for the participant’s total PHQ-8 score from 0 to 24, answered with digits. Exact texts are given in Supplementary Information and the code.

Reading the answer. We did not sample. The assistant’s reply was prefilled with an answer cue (“Answer:” for items, “Score:” for totals), and each answer option was scored by its probability as the next tokens, summed over tokenisation variants (with and without a leading space) and renormalised over the options. The unnormalised total was retained as a measure of format compliance. This is normalised candidate-label likelihood scoring. The expected-value rule uses the whole distribution, and the most-likely-option rule takes the option with the highest normalised probability, which need not coincide with the answer that free or token-by-token greedy generation would produce. For totals, we scored the 25 numbers 0–24 as token sequences. The intended rule multiplies each number’s probability by the probability that it ends where it does (the next token is not a digit); the frozen code applied this factor only to numbers whose tokens begin a longer candidate. Re-scoring the direct totals of 25 participants spread over the PHQ-8 range with both rules from the same forward pass, for four models and the participant view, changed expected totals by at most 0.008 points in that sample and no screening decision (Supplementary Information); this measures the defect in a sample rather than repairing it, and item sums are unaffected. All prompts for one transcript share the transcript as a prefix; we encoded it once and evaluated all continuations in a few forward passes with tree-structured attention, which reproduced one-prompt-at-a-time evaluation to 10^{-4} in log-probability in 32-bit arithmetic for the prompts tested. Inference was not bit-for-bit reproducible across batch compositions: re-running the same direct-total prompts reproduced the saved expected totals to a median of 0.003 points (maximum 0.68), changing 1 of 500 decisions.

Raters and scores. A rater (configuration) is a combination of model (11), transcript view (participant speech only; full dialogue), and either a wording, a format and a scoring rule applied to the eight items, summed ($5 \times 3 \times 2$), or a wording and a scoring rule applied to the direct total (5×2): $11 \times 2 \times 40 = 880$ raters. The expected-value rule scores an item as $\sum_k p_k$ and the mode rule as $\arg \max_k p_k$. A rater screens a participant positive if its score is 10 or more. Two further views served as checks: the interviewer’s speech only, which contains no report from the participant but is known to leak depression status through the interviewer’s follow-up questions⁴¹, and ASR transcripts of the same participants’ speech.

Quality checks. Two secondary outputs of the first run, for Qwen2.5 1.5B (the interviewer’s speech for one participant and the automatic transcript for another), had no finite probabilities, and the frozen analysis’s median over models silently omitted that model. We regenerated both with the frozen code, which gave finite outputs with nearly all probability on the answer options, kept the original files, and re-ran the frozen analysis: primary results were unchanged, and the prediction about the interviewer’s speech, previously met, was no longer supported. A check of every raw output against an explicit manifest of the expected models, conditions and participants now runs before the frozen analysis. The only permitted exclusions are three participants whose interviewer transcripts were empty. The check also verifies the registered dimensions and that the stored option probabilities sum to one within their rounding. It rejects empty, truncated or incomplete files and flags the two original outputs. In six discovery transcripts, empty turns had been converted to the literal token ‘nan’ (1–12 tokens each); excluding these participants changed the median AUC from 0.753 to 0.757.

Estimands. For each rater we computed the area under the ROC curve (AUC) for $\text{PHQ-8} \geq 10$ ⁶⁶, Spearman’s correlation with the PHQ-8 total, the mean absolute error and bias (mean over-rating), and the share screened positive. For each participant, π_i is the share of raters that screen them positive; a participant is discordant if $0 < \pi_i < 1$ and contested if $0.2 \leq \pi_i \leq 0.8$. The chance that two different raters drawn at random from K disagree about a participant is $2\pi_i(1 - \pi_i)K/(K - 1)$; we report its mean over participants (the frozen regularities used $2\pi_i(1 - \pi_i)$, which differs negligibly). The expected share of participants receiving both decisions from k raters drawn at random without replacement is computed exactly, $1 - [(m_i) + (K - m_i)] / \binom{K}{k}$ averaged over participants, with m_i the number of raters screening participant i positive. For two raters with positive rates p_A and p_B , $P(A \neq B) = |p_A - p_B| + 2 \min\{P(A=1, B=0), P(A=0, B=1)\}$, an algebraic identity that separates the least disagreement forced by unequal rates from the rest, without assigning causes. At equal capacity, each rater selects its q highest-scoring participants, with q the number of participants screening positive by self-report unless stated; expected-value scores had no ties at the boundaries used. We compared the between-rater variance of the AUC with its participant-bootstrap sampling variance. The generalizability study treated person, model, wording, format and view as crossed random facets of the expected item sum (330 raters in discovery), with variance components estimated from expected mean squares^{30,31}. We report the person main effect and, separately, all components involving the person. Decision studies give the generalizability ($E\rho^2$) and dependability (Φ) coefficients when scores are averaged over several models, wordings and formats; in discovery they also average both views unless stated. For each item-sum rater we computed Cronbach’s α across the eight items; the correlation between two independent runs sampled at temperature 1, which follows analytically from the answer distributions; and the mean correlation with the same rater under the other wordings. We related each index to the AUC across raters and within models. For group inference, we computed each rater’s standardised female–male difference (the difference in means divided by the square root of the mean of the two group variances) and compared its sign and size with the difference in the criterion⁶⁷. Convergent and discriminant validity compared each rater’s correlation with the PHQ-8 and with PTSD severity by the Meng–Rosenthal–Rubin test for correlated correlations⁶⁸; Steiger’s pooled form⁶⁹ and a paired participant bootstrap of the difference gave the same conclusions. Agreement between transcript sources and between deployment stacks used decision flips and ICC(2,1)⁷⁰. Pre-registered remedies were averaging across raters or models; isotonic calibration of each rater’s total score to the PHQ-8 total with 20 or 40 labelled participants (200 random draws, evaluated on the rest), which calibrates scores rather than probabilities; and prediction-powered estimation of prevalence with the same labels⁴².

Further analyses. Sensitivity, specificity and the share screened positive follow from each rater’s decisions. The share screened positive was related to the rater’s bias by a normal-ogive curve, $\Phi((b - c)/s)$, fitted by least squares; the number of additional people flagged per point of bias at the centre is $100 \phi(0)/s$. A model’s typical configuration is the one with its median AUC. Effects of wording and letter order compare model medians across the corresponding configurations; effects of aggregation compare item sums and direct totals with digit answers only, because direct totals were asked only with digits (the frozen regularity, Extended Data Table 1, counted item sums in all formats and found 7 of 11 models); and effects of size use the log parameter count. Partial Spearman correlations remove a linear trend in the ranks of the PHQ-8 from both variables and are tested with $n - 3$ degrees of freedom. Decision changes between neighbouring model sizes compare identical configurations. The label-free rule chose, for each model, the expected-value item-sum rater with the highest agreement across wordings (ties broken by sampling reliability), and was compared, as specified, with the median AUC of the model’s item-sum raters. These analyses were exploratory in the discovery set. The label-free rule, with the prediction that it would beat the model median in at least 9 of 11 models, and nine regularities, each with a pass criterion, were specified in time-stamped addenda after the discovery analysis and before any confirmation output was compared with the criterion, while confirmation outputs were still being generated; all are reported. The criteria are descriptive, and because the regularities share participants and raters they are not independent replications.

Analyses added after the confirmation analysis. The following analyses were not pre-specified and are exploratory in both sets. Pairwise disagreement in stricter sets of raters ($AUC \geq 0.70$ and 5–95% screened positive; AUC , sensitivity and specificity all ≥ 0.70), and the exact expected discordance among k raters. The identity for the original decisions, and equal-capacity selection at 10–50% of participants, for expected-value raters with $AUC \geq 0.70$ (a selection that uses the whole cohort’s labels), all expected-value raters and one fixed configuration per model (instrument wording, digits, participant’s speech, expected value), with each rater’s disagreement with the criterion at the self-reported prevalence. Single-choice comparisons over all pairs of configurations identical except in one choice. The model-to-person variance ratio with digit answers only, for the instrument wording with digits and the participant’s speech, and with each developer’s models left out. The label-free rule against the median of expected-value item sums, the median of those with digit answers, and the fixed neutral configuration. Recalibration on held-out participants: in each of 200 random splits, 20, 40 or 80 participants (20 or 40 in confirmation) formed a labelled sample and the rest were held out; three rater sets were recalibrated, namely raters with $AUC \geq 0.70$ in the labelled sample, one fixed configuration per model, and, for comparison, raters with $AUC \geq 0.70$ in the whole cohort, whose admission uses the held-out labels; the rules were subtracting the mean over-rating, linear regression of the PHQ-8 total on the score, logistic regression for PHQ-8 ≥ 10 with a weak ridge penalty (deciding at a probability of 0.5), the attainable threshold whose labelled positive count is closest to the number of labelled positives, compared exactly as integers with ties going to the higher threshold (ties among scores can make the target unattainable; achieved counts are reported, a miss of 5 percentage points counting as within 5), and the highest threshold that keeps at least 80% of the labelled positives (the exact order statistic). We report held-out accuracy, sensitivity, specificity, pairwise disagreement, the median absolute error of the share flagged with its 2.5th and 97.5th percentiles across splits, and achieved training targets. A defensible core (instrument or clinician wording, digits, expected values, symptom sums; every view) was analysed alongside these sets without any selection by outcome. Deferral was compared at matched coverage on the same held-out participants: a single recalibrated rater, the majority of raters, the majority after setting aside participants on whom fewer than a given share of raters agreed, single raters on the same retained participants, and each single rater setting aside its own closest calls, those nearest its cut-off, to reach the same coverage; retained participants with and without a positive screen are counted. Where disagreement falls was summarised by self-reported PHQ-8 band, at the threshold and at equal capacity. Three headline estimands were bootstrapped over participants with the rater sets held fixed. Transfer carried each configuration’s discovery offset or discovery thresholds (matching the discovery prevalence, or keeping 80% of discovery positives), for configurations with $AUC \geq 0.70$ in discovery and for the single most accurate discovery configuration, unchanged to confirmation. Prevalence intervals from 40 labels (2,000 draws): prediction-powered with the registered formula, prediction-powered for the finite cohort, and from the labels alone. The discovery normal-ogive curve applied unchanged to confirmation. The decision rules and their tie handling are implemented and tested in the analysis code.

Pre-registration and confirmation. The study was pre-registered on OSF Registries (<https://osf.io/8qg2m/>) before any model output was compared with the criterion, and the scoring and analysis code were fixed at the same time. The pre-registered hypotheses were that the validity, individual screening decisions and implied prevalence of the ratings depend substantially on configuration choices, and that indices of rating consistency do not indicate which configurations are valid. Model outputs for the confirmation set were generated after the discovery analysis was complete, with the locked code, and analysed once. Deviations from the plan, including the regenerated outputs and the scoring defect, are listed in Supplementary Information.

Statistics. Intervals for AUCs and for the gain of the label-free rule are percentile 95% intervals from 2,000 bootstrap resamples of participants. The pre-registered calibration and prediction-powered analyses used 200 random draws of labelled participants (50 for the display in Extended Data Fig. 4b); the exploratory analyses used 200 splits for recalibration and deferral, 2,000 draws for prevalence intervals, 1,000 paired bootstrap resamples for the specificity check and 1,000 participant resamples, with the rater sets held fixed, for the headline intervals, which are therefore conditional on the raters. Split-to-split ranges describe variation over repeated splits of one cohort and are not confidence intervals. All tests are two-sided unless stated.

Data availability

DAIC-WOZ and E-DAIC are available from the University of Southern California Institute for Creative Technologies under a data use agreement. Only aggregate results are distributed with the code.

Code availability

The scoring and analysis code will be made available on publication at a public repository.

Acknowledgements

This work was supported in part by a 2025 NARSAD Young Investigator Grant from the Brain & Behavior Research Foundation, funded by The Sidney R. Baer, Jr. Foundation (Grant ID 33941). B.L. is also supported by the U.S. National Institutes of Health (NIH), the U.S. Department of Veterans Affairs (VA), the Hasso Plattner Foundation, and the Windreich Family Foundation. We thank the University of Southern California Institute for Creative Technologies for making the interview corpora available.

Author contributions

B.L. conceived the study, performed the analyses, and wrote the manuscript.

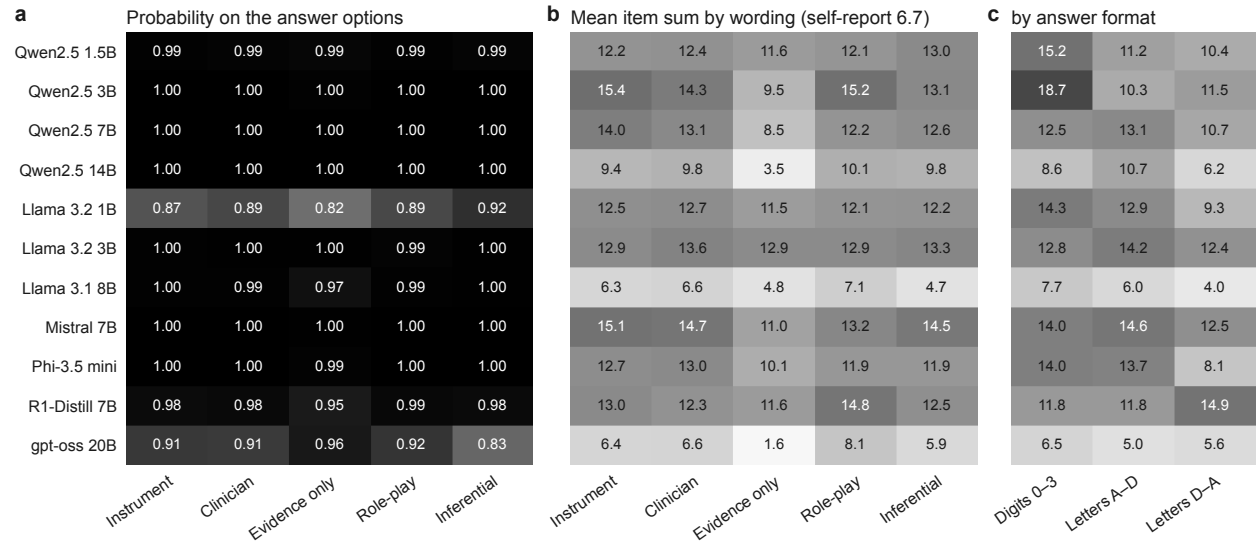
Competing interests

The author declares no competing interests.

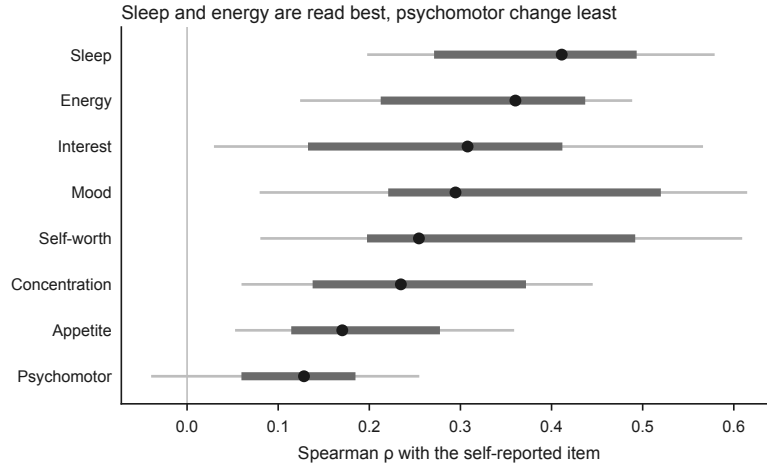
Extended Data

Extended Data Table 1: **Regularities found in the discovery set and checked in new interviews.** Discovery values are exploratory; the criteria were specified before the confirmation analysis. π , a participant's share of raters with AUC ≥ 0.70 screening them positive. For the first regularity, confirmation values refit the curve; applied unchanged, the discovery curve gives $R^2 = 0.90$. P values for partial correlations use $n - 3$ degrees of freedom ($P = 0.049$ for the PTSD regularity in confirmation). With digit answers only, item sums over-rate in 8 of 11 models in both sets. The regularities share participants and raters, so they are not independent tests.

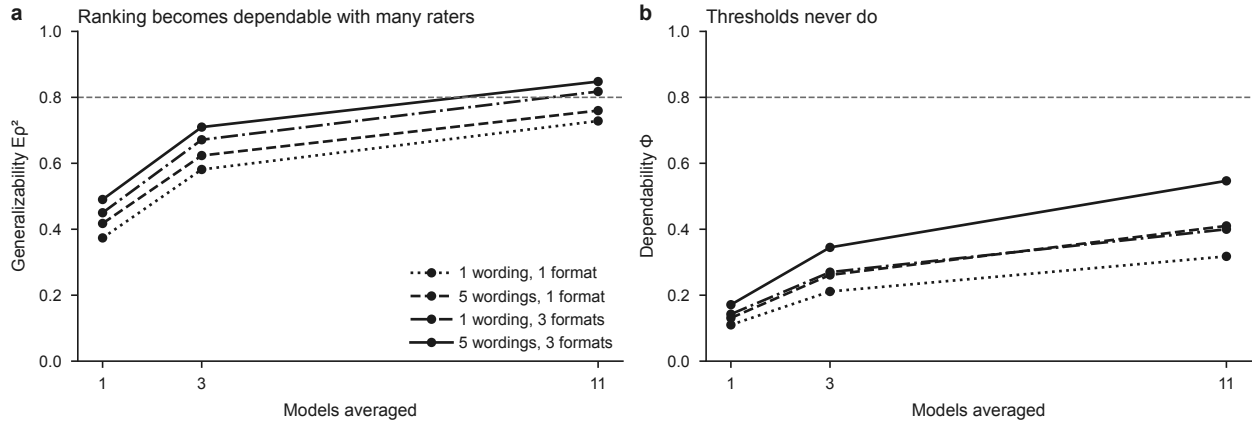
Regularity	Criterion	Discovery	Confirmation	Held
The share screened positive follows a configuration's mean over-rating (probit R^2 ; referrals per 100 per point)	$R^2 \geq 0.75$	0.91; 13	0.91; 14	yes
Configurations with the same AUC (± 0.02) differ in share screened positive (median, points)	≥ 15	28	28	yes
Participants reporting few symptoms (PHQ-8 ≤ 4) are screened positive (mean share of configurations with AUC ≥ 0.70)	$\geq 20\%$	35%	44%	yes
Larger models discriminate better (Spearman, size vs median AUC), yet no model's ratings correlate with the PHQ-8 more than 0.10 above their correlation with PTSD severity (median difference; not a test of diagnostic separation)	> 0.5 ; none > 0.10	0.82; max 0.04	0.77; max 0.01	yes
Below the threshold, more PTSD symptoms go with a larger share screened positive at equal PHQ-8 (partial Spearman)	> 0 , $P < 0.05$	0.36	0.23	yes
Rating only explicitly reported symptoms lowers referrals; reversing the answer letters moves them (median points)	8 and 6 of 11 models	27; 21	31; 24	yes
Summing item ratings over-rates relative to rating the total directly (models; as frozen, item sums in all formats)	≥ 6 of 11	7 of 11	7 of 11	yes
Talkative participants are screened positive more often at equal PHQ-8 (partial Spearman)	> 0 , $P < 0.05$	0.25	0.26	yes
The most accurate model's own configurations disagree about participants (mean $2\pi(1 - \pi)$, as frozen)	≥ 0.15	0.24	0.26	yes



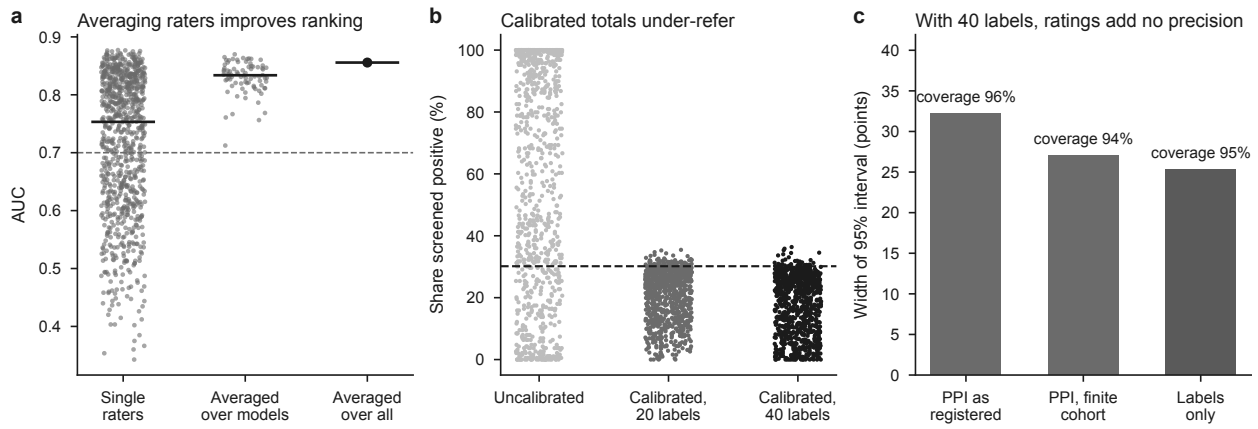
Extended Data Fig. 1: **Format compliance and rated severity.** **a**, Median probability that each model placed on the answer options, by wording (participant view). **b**, Mean expected item sum by model and wording; the self-reported mean PHQ-8 is given in the panel title. **c**, Mean expected item sum by model and answer format.



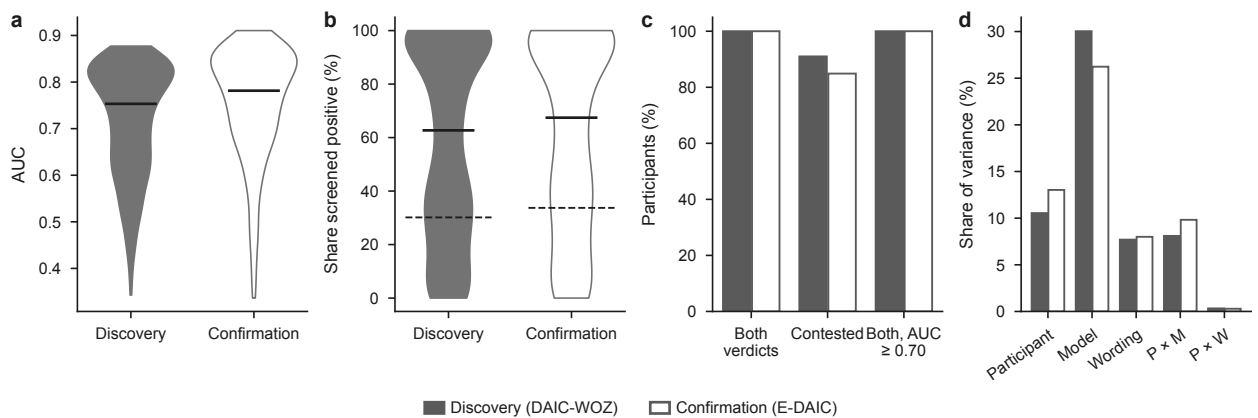
Extended Data Fig. 2: **Agreement with individual symptoms.** Spearman correlation between each model's expected item rating and the participant's self-reported score on the same item, for the 141 participants with complete item scores (participant view), per symptom: dot, median model; thick line, interquartile range over models; thin line, range over models; each model's value is its median over wordings and formats. Exploratory analysis.



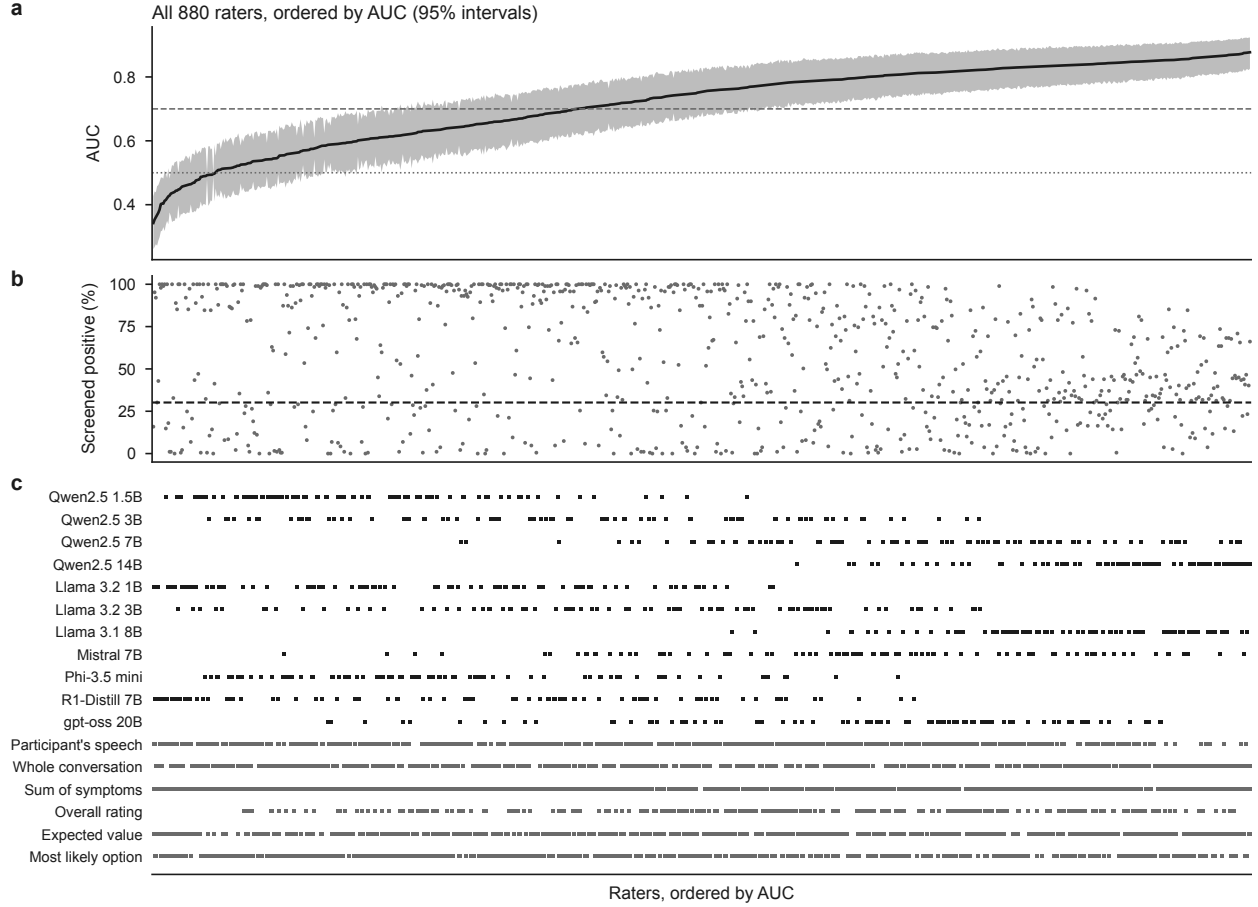
Extended Data Fig. 3: **Decision studies.** Generalizability ($E\rho^2$; **a**) and dependability (Φ ; **b**) of the expected item sum (discovery), when scores are averaged over the given number of models and over one or all wordings and formats, with both views of the conversation averaged. Dashed line, 0.80.



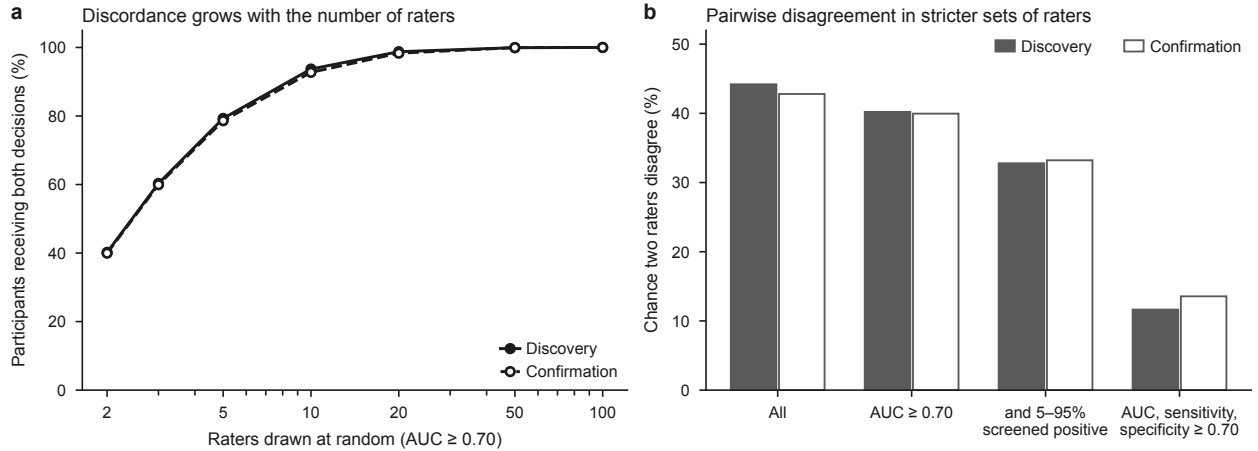
Extended Data Fig. 4: **What averaging and labels do** (discovery). **a**, AUC of single raters, of scores averaged over models for each configuration of the other choices, and of the average over all raters; lines, medians; dashed line, 0.70. **b**, Share screened positive by each rater before and after isotonic calibration of its total score against 20 or 40 labelled participants (mean over 50 draws); dashed line, self-reported prevalence. **c**, Mean width of 95% intervals for the prevalence from 40 labelled participants (2,000 draws): prediction-powered intervals with the registered formula and for the finite cohort, and intervals from the labels alone; observed coverage above the bars.



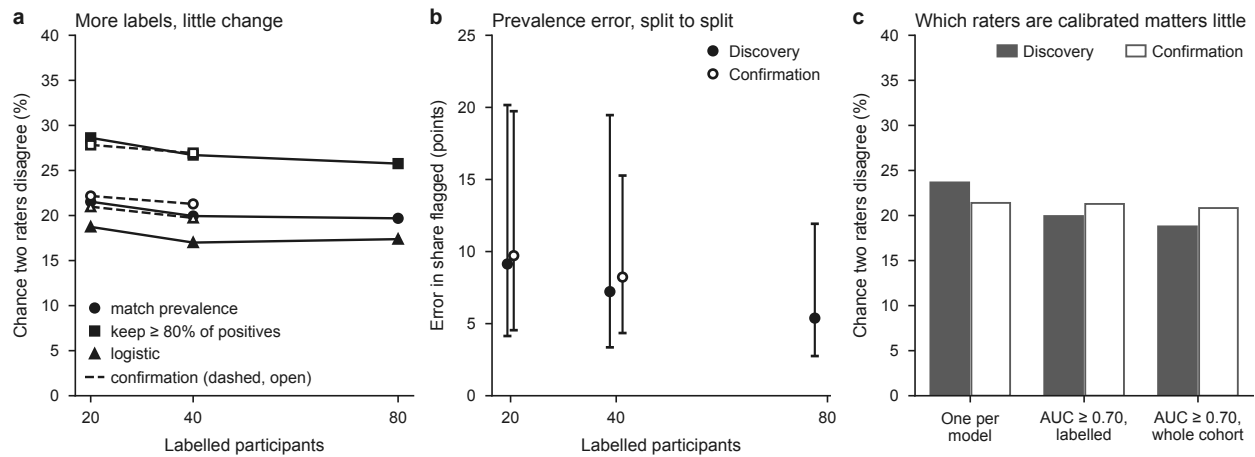
Extended Data Fig. 5: **Discovery and confirmation**. Discovery (DAIC-WOZ, dark grey) and confirmation (E-DAIC sessions 600–718, automatic transcripts, white). **a**, AUC of every rater; lines, medians. **b**, Share screened positive; dashed lines, self-reported prevalence. **c**, Participants receiving both decisions, contested decisions (20–80% of raters), and both decisions among raters with AUC ≥ 0.70 . **d**, Shares of the variance of the item-sum score: participant, model, wording, participant $\times$ model ($P \times M$) and participant $\times$ wording ($P \times W$).



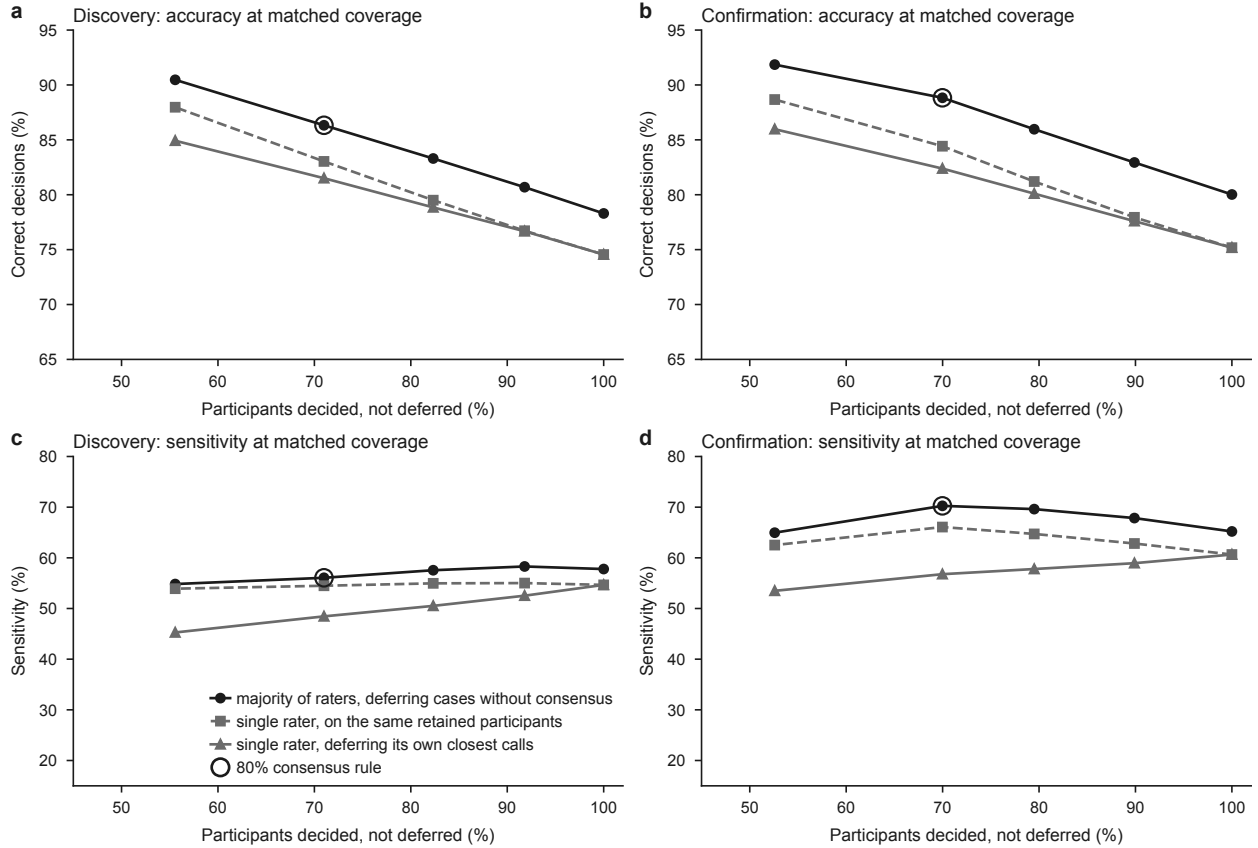
Extended Data Fig. 6: **All raters, ordered by accuracy** (discovery). **a**, AUC of every rater, ordered, with 95% bootstrap intervals; dotted line, chance; dashed line, 0.70. **b**, Share screened positive by the same raters; dashed line, self-reported prevalence. **c**, The choices that make up each rater.



Extended Data Fig. 7: **How disagreement depends on which raters are counted**. **a**, Expected share of participants who receive both decisions from k raters with AUC ≥ 0.70 drawn at random without replacement (exact), against k ; filled markers and solid line, discovery; open markers and dashed line, confirmation. **b**, Chance that two different raters disagree about a participant, among all raters, raters with AUC ≥ 0.70 , those that also screen 5–95% of participants positive, and those with AUC, sensitivity and specificity all ≥ 0.70 ; dark grey, discovery; white, confirmation. Exploratory analyses.



Extended Data Fig. 8: **Recalibration in detail.** **a**, The chance that two recalibrated raters disagree about a held-out participant against the number of labelled participants, for raters admitted by their AUC in the labelled sample and recalibrated to match the labelled prevalence, to keep at least 80% of labelled positives, or by logistic regression; solid and filled, discovery; dashed and open, confirmation. **b**, Median absolute error of the share flagged with the prevalence-matched threshold; bars, 2.5th to 97.5th percentiles across splits (not confidence intervals). **c**, Disagreement after matching the prevalence with 40 labels for three rater sets: one fixed configuration per model, raters admitted by their AUC in the labelled sample, and raters admitted by their AUC in the whole cohort, a selection that uses the held-out labels. Means over 200 splits; exploratory analyses.



Extended Data Fig. 9: **Deferral at matched coverage.** For raters admitted by their AUC on 40 labelled participants and recalibrated to the labelled prevalence, accuracy (**a**, discovery; **b**, confirmation) and sensitivity (**c**, discovery; **d**, confirmation) on the held-out participants who are decided rather than deferred, against the share decided: the majority of raters for all participants (50% consensus) and after setting aside participants on whom fewer than 60, 70, 80 or 90% of raters agree (black; unanimous consensus is not shown because it left fewer than 3% of participants); single raters on the same retained participants (grey, dashed); and single raters setting aside their own closest calls to reach the same coverage (grey, solid). Large circle, the 80% consensus rule. Means over 200 splits; exploratory analyses.

Supplementary Information

1 Supplementary Methods

1.1 Prompts

Each prompt was a single user turn in the model’s own chat format. Its components appeared in a fixed order: a one-sentence description of the transcript view; the transcript, enclosed in <transcript> tags, with turns labelled “Interviewer:” and “Participant:” in the full-dialogue view, “Interviewer:” in the interviewer view and unlabelled in the participant view; the template’s task instruction; the response format; and one PHQ-8 item preceded by the stem “Over the last 2 weeks, how often has the participant been bothered by the following problem?”. The assistant’s reply began with the answer cue specified by that format. For the direct total, the format and item were replaced by “Considering all eight PHQ-8 problems together, what is the participant’s total PHQ-8 score (from 0 to 24)? Reply in the form ‘Score: N’.”, and the reply was prefilled with “Score:”. Supplementary Table 1 gives every text.

Supplementary Table 1: Prompt texts. The transcript follows the first line; the template, the format and the item follow the transcript.

Transcript view (first line of every prompt)	
participant	Below is a transcript of a clinical research interview. Only the participant’s speech is shown.
full	Below is a transcript of a clinical research interview between an interviewer and a participant.
interviewer	Below is a transcript of a clinical research interview. Only the interviewer’s speech is shown.
Template	
instrument	Task: complete the PHQ-8 depression questionnaire on behalf of the participant, using the interview as your only source of information.
clinician	You are an experienced psychiatrist. Using your clinical judgement about this interview, rate the participant on the PHQ-8 depression scale.
evidence	Rate each PHQ-8 symptom only from what the participant explicitly reports in the interview. If a symptom was not discussed, rate it as ‘Not at all’. Do not speculate.
role-play	Imagine that you are the participant in this interview. Answer the PHQ-8 questionnaire the way they would answer it about themselves, given everything they said.
inferential	Use every available cue, including indirect ones such as tone, word choice and what the participant avoids talking about, to infer how the participant would score on the PHQ-8.
Response format (answer cue prefilled in the reply)	
digits	Answer each question with a single digit: 0 = Not at all, 1 = Several days, 2 = More than half the days, 3 = Nearly every day. Reply with the digit only. Cue: “Answer:”
letters	Answer each question with a single letter: A = Not at all, B = Several days, C = More than half the days, D = Nearly every day. Reply with the letter only. Cue: “Answer:”
letters reversed	Answer each question with a single letter: A = Nearly every day, B = More than half the days, C = Several days, D = Not at all. Reply with the letter only. Cue: “Answer:”
Item stem and items	
stem	Over the last 2 weeks, how often has the participant been bothered by the following problem?
item 1	Little interest or pleasure in doing things.
item 2	Feeling down, depressed, or hopeless.
item 3	Trouble falling or staying asleep, or sleeping too much.
item 4	Feeling tired or having little energy.
item 5	Poor appetite or overeating.
item 6	Feeling bad about yourself, or that you are a failure or have let yourself or your family down.
item 7	Trouble concentrating on things, such as reading the newspaper or watching television.
item 8	Moving or speaking so slowly that other people could have noticed, or the opposite, being so fidgety or restless that you have been moving around a lot more than usual.
direct total	Considering all eight PHQ-8 problems together, what is the participant’s total PHQ-8 score (from 0 to 24)? Reply in the form ‘Score: N’. Cue: “Score:”

1.2 Models

Supplementary Table 2 lists the models, their runtimes and the exact weights. Transformers models ran on the Apple GPU (Metal Performance Shaders) in bfloat16; GGUF files ran with llama.cpp (llama-cpp-python 0.3.16) with all

layers on the GPU. The reasoning model received an empty reasoning block, and gpt-oss-20b was set to low reasoning effort and read in its final channel, so that every model answered directly.

Supplementary Table 2: Models and weights. [†]Numerical-precision check only; not part of the multiverse.

Model	Developer	Parameters (B)	Runtime and precision	Weights
Qwen2.5 1.5B	Alibaba	1.5	Transformers, bf16	Qwen/Qwen2.5-1.5B-Instruct (revision 989aa79)
Qwen2.5 3B	Alibaba	3.1	Transformers, bf16	Qwen/Qwen2.5-3B-Instruct (revision aa8e725)
Qwen2.5 7B	Alibaba	7.6	Transformers, bf16	Qwen/Qwen2.5-7B-Instruct (revision a09a354)
Qwen2.5 14B	Alibaba	14.8	llama.cpp, Q4_K_M	Ollama library qwen2.5:14b; SHA-256 2049f5674b1e
R1-Distill 7B	DeepSeek	7.6	llama.cpp, Q4_K_M	DeepSeek-R1-Distill-Qwen-7B-Q4_K_M.gguf; SHA-256 731ece8d06dc
Llama 3.2 1B	Meta	1.2	llama.cpp, Q8_0	Ollama library llama3.2:1b; SHA-256 74701a8c35f6
Llama 3.2 3B	Meta	3.2	llama.cpp, Q4_K_M	llama-3.2-3b-instruct.Q4_K_M.gguf; SHA-256 6c1a2b411610
Llama 3.1 8B	Meta	8	llama.cpp, Q4_K_M	Ollama library llama3.1:8b; SHA-256 667b0c1932bc
Phi-3.5 mini	Microsoft	3.8	llama.cpp, Q4_0	Ollama library phi3.5:latest; SHA-256 b5374915da53
Mistral 7B	Mistral AI	7.2	llama.cpp, Q4_K_M	mistral-7b-instruct-v0.3.Q4_K_M.gguf; SHA-256 1270d22c0fbb
gpt-oss 20B	OpenAI	21	llama.cpp, MXFP4/Q4_K_M	gpt-oss-20b-Q4_K_M.gguf; SHA-256 c27536640e41
Qwen2.5 7B (4-bit) [†]	Alibaba	7.6	llama.cpp, Q4_K_M	Ollama library qwen2.5:7b; SHA-256 2bada8a74506

1.3 Reading the answer

For each item we read the probability of every response option at the position after the answer cue, summing the probabilities of tokenisation variants (with and without a leading space). The four probabilities were renormalised to sum to one; their unnormalised sum (the option mass) measures how completely the model complied with the format. For the direct total, each number from 0 to 24 was scored as a token sequence. The intended rule multiplies its probability by the probability that the next token is not a digit, so that, for example, “1” is not counted when the model goes on to write “12”. The frozen code applied this factor only to numbers whose token sequence begins a longer candidate. Supplementary Table 3 gives the size of the difference: we re-scored the direct totals of 25 participants spread over the PHQ-8 range for four models, assembling both rules from the same forward pass. The 25 probabilities were renormalised in the same way. The expected-value rule scores an item as $\sum_k k p_k$ and the mode rule as $\arg \max_k p_k$. This is normalised candidate-label likelihood scoring: each option’s probability sums its tokenisation variants, and the mode is the option with the highest normalised probability. It need not coincide with the answer of free or token-by-token greedy generation; for example, an option whose two token variants each have probability 0.25 is chosen over an option with a single variant of probability 0.40, whereas greedy generation would emit the latter.

Supplementary Table 3: Direct totals re-scored for 25 participants (five wordings each). Termination rule: the factor applied to every candidate against the frozen rule, from the same forward pass; flips, screening decisions that change under either scoring rule. Re-run against saved: the frozen rule in the new pass against the saved outputs (different batch composition).

Model	Totals	Termination rule		Re-run against saved		
		max $ \Delta $	Flips	median $ \Delta $	max $ \Delta $	Flips
gpt-oss 20B	125	0.0000	0	0.006	0.68	0
Llama 3.1 8B	125	0.0000	0	0.001	0.01	0
Mistral 7B	125	0.0080	0	0.001	0.01	0
Qwen2.5 7B	125	0.0028	0	0.090	0.45	1

1.4 Tree-structured evaluation

All prompts for one transcript and view share the transcript as a prefix. We encoded the shared prefix once, then evaluated the continuations (templates, formats and items) in a few forward passes. An attention mask restricted each continuation to its own ancestors in the prompt tree. Position indices continued from that continuation’s prefix. For Transformers models this was a four-dimensional attention mask; for llama.cpp, one sequence per continuation sharing the cached prefix. Against one-prompt-at-a-time evaluation, the log-probabilities agreed to 10^{-4} in 32-bit arithmetic; in bfloat16 option probabilities differed by at most 0.03, and with llama.cpp log-probabilities agreed within 0.007. The checks covered one participant and 16 sampled continuations (the 20 most probable tokens of each). They diagnose numerical agreement without guaranteeing it for every prompt; Supplementary Table 3 shows the reproducibility of whole scores across batch compositions.

1.5 Generalizability analysis

The data were the expected item sums X_{pmfvt} for person p , model m , template t , format f and view v , in a fully crossed design with one observation per cell. We computed mean squares for all 31 effects and solved their expectations for the variance components, working downwards from the highest-order interaction. Negative estimates were retained when solving for lower-order components, then set to zero for reporting³¹. The highest-order interaction is confounded with residual error. For a decision study that averages over n'_m models, n'_t templates and n'_f formats (and, in discovery, over both views unless stated), the relative error variance σ_δ^2 is the sum of the interactions involving the person, each divided by the numbers of conditions averaged over its other facets; the absolute error variance σ_Δ^2 adds the effects not involving the person, divided in the same way. The coefficients are $E\rho^2 = \sigma_p^2/(\sigma_p^2 + \sigma_\delta^2)$ and $\Phi = \sigma_p^2/(\sigma_p^2 + \sigma_\Delta^2)$.

1.6 Consistency indices

For an item-sum configuration with answer distributions p_{ik} for participant i and item k , the expected total is $\mu_i = \sum_k E[X_{ik}]$ and, with items answered independently at temperature 1, the sampling variance of the total is $s_i^2 = \sum_k \text{Var}(X_{ik})$. The expected correlation between two independent sampled runs is $\text{Var}_i(\mu_i)/[\text{Var}_i(\mu_i) + s_i^2]$. Cronbach’s α was computed across the eight expected item scores. Agreement across templates is the mean Pearson correlation of the configuration’s expected totals with those of the same model, format and view under each of the other four templates.

1.7 Remedies

For calibration, we randomly drew n participants ($n = 20$ or 40) and fitted isotonic regression from each configuration’s score to their PHQ-8 totals. We applied the fitted map to the remaining participants and evaluated the resulting decisions and implied prevalence. This was repeated for 200 draws. Prediction-powered estimates of prevalence used the configuration’s decisions on all participants and the labels of the same n participants to correct the bias of the model-based prevalence, with the standard variance estimate⁴². The standard estimate targets a superpopulation. Our coverage assessment instead used the prevalence of the fixed cohort. In the exploratory analyses we therefore also computed a finite-cohort version, $\hat{\theta} = \bar{f} + (\bar{y} - \bar{f})_L$ with variance $(1 - n/N) s_{y-f}^2/n$, and an estimate from the labels alone, $\bar{y}_L$ with variance $(1 - n/N) s_y^2/n$, where f is the configuration’s decision, y the label, L the labelled participants and N the cohort size.

1.8 Decision analyses (exploratory)

How many versus who. For two raters with decisions A and B and positive rates p_A and p_B , $P(A \neq B) = P(A=1, B=0) + P(A=0, B=1)$ and $p_A - p_B = P(A=1, B=0) - P(A=0, B=1)$, so $P(A \neq B) = |p_A - p_B| + 2 \min\{P(A=1, B=0), P(A=0, B=1)\}$. The first term is the least disagreement that the two positive rates force; the second, the excess, is zero exactly when one rater’s positives are a subset of the other’s. We average both over all pairs of distinct raters in a set. The identity is elementary and assigns no causes: a pure difference of scale between two raters with identical rankings and equal means produces disagreement at a fixed threshold, so the residual after subtracting each rater’s mean bias is not a measure of ranking disagreement. At equal capacity, every rater selects its q highest-scoring participants; we used expected-value scores, whose boundaries had no ties at the capacities reported (ties would be broken by participant order and are counted in the code). Single-choice comparisons use every pair of configurations that are identical except in the named choice, at the threshold of 10 and, for expected-value pairs, at equal capacity.

Recalibration. In each of 200 random splits (the same splits for every rule and rater set), 20, 40 or 80 participants formed a labelled sample (20 or 40 in confirmation) and the rest were held out. Three rater sets were recalibrated: raters with $AUC \geq 0.70$ in the labelled sample; one fixed configuration per model (instrument wording, digits, participant's speech, expected value), which uses no selection; and, for comparison, raters with $AUC \geq 0.70$ in the whole cohort, whose admission uses the held-out labels. The rules were: subtracting the mean difference between the score and the PHQ-8 total in the labelled sample; linear regression of the PHQ-8 total on the score, thresholded at 10; logistic regression of $PHQ-8 \geq 10$ on the standardised score with a weak ridge penalty (10^{-3} , intercept included; Newton's method, convergence recorded), deciding at a probability of 0.5; the threshold, among the attainable ones (the distinct labelled scores), whose labelled positive count is closest to the number of labelled positives, compared exactly as integers, the higher threshold on equal distance; and the highest threshold keeping at least 80% of the labelled positives, that is, the $\lceil 0.8 n_+ \rceil$ -th largest labelled positive score. Ties among scores can make the prevalence target unattainable, and the achieved labelled rates are reported. Every rule sets a cut-off on the rater's own score. An increasing fitted relation preserves its ranking, a decreasing fit reverses it, and a constant fit flags everyone or no one. With 40 labels and raters admitted on the labelled sample, the linear fit was increasing in all but 14 of 107,540 rater-split combinations in discovery and in all 62,743 in confirmation. On the held-out participants we computed the chance that two recalibrated raters disagree, accuracy, the median sensitivity and specificity across raters, and the median absolute error of the share flagged, and averaged over splits; the central 95% range across splits (2.5th to 97.5th percentiles) describes split-to-split variation within one cohort and is not a confidence interval (Supplementary Table 4). Deferral set aside the held-out participants flagged by more than 20% and fewer than 80% of the recalibrated raters, retaining those at exactly 20% or 80%, and scored the majority decision on the rest.

Defensible core. Besides the full grid, raters with $AUC \geq 0.70$ and one fixed configuration per model, we report a clinically defensible core: the questionnaire's own instruction or the psychiatrist wording (neither changes the assessment policy), digit answers, expected values and symptom sums, in every view (44 raters in discovery, 22 in confirmation). It uses no outcome labels to choose raters (Supplementary Table 5).

Where disagreement falls. For each participant, the chance that two raters decide differently, $2\pi_i(1 - \pi_i)K/(K - 1)$, was averaged within bands of self-reported PHQ-8 (0–4, 5–9, 10–14, 15–24), at the threshold of 10 and at equal capacity, and each band's share of the total was computed (Supplementary Table 6). This describes the criterion status of the participants on whom raters differ; it does not identify which decision is clinically right.

Deferral at matched coverage. On the held-out participants of each split, with raters admitted on the 40 labels and recalibrated to the labelled prevalence, we compared on the same participants: a single recalibrated rater (averaged over raters); the majority of raters; the majority after deferring participants on whom fewer than a share c of raters agreed ($c = 0.6, 0.7, 0.8, 0.9, 1.0$; $c = 0.8$ defers those flagged by strictly more than 20% and fewer than 80% of raters, retaining those at exactly 20% or 80%; unanimous consensus retained fewer than 3% of participants and is not displayed); single raters evaluated on the same retained participants, which separates the effect of deferral from that of pooling; and each single rater deferring its own closest calls, the participants whose scores lie nearest its cut-off, to reach the same coverage. We report accuracy, sensitivity, specificity and the numbers of retained participants with and without a positive screen (Supplementary Table 7; Extended Data Fig. 9). Neither deferral rule has defined costs for review or for missed cases, so these are descriptive comparisons, not a clinical policy evaluation.

Uncertainty for headline estimands. For three headline estimands (pairwise disagreement at 10 among raters with $AUC \geq 0.70$, and equal-capacity disagreement among expected-value raters with $AUC \geq 0.70$ and among one fixed configuration per model) we resampled participants 1,000 times with the rater sets held fixed; the percentile intervals are therefore conditional on the raters and describe sampling of participants, not of configurations.

Transfer. For configurations with $AUC \geq 0.70$ in discovery (participant view), three discovery calibrations were applied unchanged to the same configuration in confirmation: subtracting the discovery mean over-rating; the attainable discovery threshold closest to the discovery prevalence; and the discovery threshold keeping at least 80% of discovery positives. The discovery set had manual transcripts and an operator-controlled interviewer, and the confirmation set automatic transcripts and an autonomous interviewer, so this is one related transfer, not a comparison of independent sites.

Supplementary Table 4: Recalibration with 40 labelled participants, evaluated on the held-out participants (raters admitted by their AUC in the labelled sample; means over 200 splits). Disagreement, the chance that two recalibrated raters decide a held-out participant differently; error, median absolute error of the share flagged (2.5th–97.5th percentile across splits); target, the share of raters and splits whose labelled positive rate came within 5 points of the labelled prevalence (prevalence rule) or that kept at least 80% of labelled positives (sensitivity rule).

Set	Rule	Disagree (%)	Acc. (%)	Sens. (%)	Spec. (%)	Error (points)	Target (%)
Discovery	Threshold of 10	40	61	83	67	28 (21–32)	
	Mean over-rating	18	75	39	92	14 (7–25)	
	Linear	16	75	42	91	12 (5–27)	
	Logistic	17	75	45	89	11 (4–28)	
	Labelled prevalence	20	75	56	84	7 (3–19)	88
	≥ 80% of positives	27	68	78	68	17 (7–33)	100
Confirmation	Threshold of 10	40	59	91	53	33 (26–39)	
	Mean over-rating	21	74	49	92	14 (7–24)	
	Linear	20	75	54	89	11 (4–24)	
	Logistic	20	75	56	87	10 (4–22)	
	Labelled prevalence	21	75	62	84	8 (4–15)	86
	≥ 80% of positives	27	69	79	70	15 (7–28)	100

Supplementary Table 5: How the decision results change under restrictions of the grid. K , raters; at 10, the chance that two raters decide a participant differently at the threshold; forced, the share of that disagreement forced by different positive rates; equal no., the chance at equal capacity among expected-value raters of the set (all expected-value raters for the full grid); recal. and acc., held-out disagreement and accuracy after matching the labelled prevalence with 40 labels (no selection for the defensible core and the fixed configurations; whole-cohort AUC for the $AUC \geq 0.70$ row).

Set	Raters	K	At 10 (%)	Forced (%)	Equal no. (%)	Recal. (%)	Acc. (%)
Discovery	Full grid	880	44	92	27	–	–
	$AUC \geq 0.70$	540	40	94	18	19	75
	Defensible core	44	31	95	21	21	74
	One per model	11	32	98	25	24	72
Confirmation	Full grid	440	43	94	26	–	–
	$AUC \geq 0.70$	319	40	95	20	21	76
	Defensible core	22	30	97	21	20	76
	One per model	11	32	97	22	21	75

Supplementary Table 6: Where the disagreement falls, by self-reported PHQ-8, for expected-value raters with $AUC \geq 0.70$: each participant's chance of being decided differently by two raters, averaged within the band (mean) and as the band's share of all such disagreement (share), at the threshold of 10 and at equal capacity.

Set	PHQ-8	n	At 10		Equal numbers	
			Mean (%)	Share (%)	Mean (%)	Share (%)
Discovery	0–4	86	44	48	9	24
	5–9	46	43	25	25	33
	10–14	30	40	15	27	24
	15–24	27	33	11	25	20
Confirmation	0–4	36	47	48	13	28
	5–9	21	42	25	24	29
	10–14	13	38	14	30	23
	15–24	16	27	12	22	21

Supplementary Table 7: Deferral at matched coverage on held-out participants (raters admitted on 40 labels and recalibrated to the labelled prevalence; means over 200 splits). Coverage, participants decided rather than deferred; counts in parentheses, mean retained participants with and without a positive screen.

Set	Rule	Coverage (%)	Accuracy (%)	Sens. (%)	Spec. (%)
Discovery	Single rater, all	100	75	55	83
	Majority, all	100	78	58	88
	Majority, consensus ≥ 0.8 (24+, 82–)	71	86	56	95
	single rater, same participants	71	83	54	91
	single rater, own closest calls deferred	71	82	48	92
	Majority, consensus ≥ 0.9 (13+, 70–)	56	90	55	97
	single rater, same participants	56	88	54	94
	single rater, own closest calls deferred	56	85	45	94
	Single rater, all	100	75	61	83
Confirmation	Majority, all	100	80	65	88
	Majority, consensus ≥ 0.8 (9+, 24–)	70	89	70	95
	single rater, same participants	70	84	66	91
	single rater, own closest calls deferred	70	82	57	92
	Majority, consensus ≥ 0.9 (5+, 19–)	53	92	65	98
	single rater, same participants	53	89	63	95
	single rater, own closest calls deferred	53	86	53	94

Supplementary Table 8: Participants decided differently by configurations that differ in one choice only: median (interquartile range) over pairs, at the threshold of 10 and, for expected-value pairs, selecting as many participants as screen positive by self-report.

Choice changed	Discovery		Confirmation	
	At 10	Equal numbers	At 10	Equal numbers
Letter–answer mapping reversed	20 (7–55)	21 (12–43)	26 (6–49)	21 (12–36)
Digits vs letters	13 (5–25)	15 (7–24)	12 (5–22)	14 (9–22)
Psychiatrist persona	7 (2–13)	8 (6–15)	6 (2–12)	10 (7–16)
Unmentioned symptoms rated absent	21 (8–35)	14 (8–23)	23 (9–36)	14 (9–21)
Role-play as the participant	8 (2–14)	11 (7–17)	8 (3–14)	10 (7–19)
Use tone and other cues	8 (2–17)	11 (7–20)	9 (2–14)	14 (7–19)
Most likely option, not expected value	4 (1–8)	–	3 (1–8)	–
One overall rating, not a symptom sum	24 (10–49)	15 (8–22)	21 (10–35)	16 (9–23)
Whole conversation, not participant’s speech	11 (3–19)	17 (13–26)	–	–
Next model size, same family	28 (13–47)	22 (15–33)	30 (16–48)	23 (16–30)
A different model	35 (17–63)	25 (17–33)	35 (15–60)	23 (16–33)

2 Supplementary Results

Supplementary Table 9: Results by model (discovery set; all 80 configurations of each model). Sens. and Spec., median sensitivity and specificity for PHQ-8 ≥ 10 over the model’s configurations (descriptive). Bias, median over configurations of the mean difference between the configuration’s score and the PHQ-8 total. Option mass, median over participants, templates and formats.

Model	AUC, median (range)	AUC ≥ 0.70 (%)	Positive, % (range)	Sens. (%)	Spec. (%)	Bias	Option mass
Qwen2.5 1.5B	0.59 (0.41–0.77)	8	50 (0–100)	55	52	+3.2	0.99
Qwen2.5 3B	0.70 (0.49–0.82)	49	84 (0–100)	96	21	+5.0	1.00
Qwen2.5 7B	0.81 (0.64–0.87)	96	67 (17–98)	93	45	+5.0	1.00
Qwen2.5 14B	0.85 (0.78–0.88)	100	33 (1–82)	71	83	+0.7	1.00
R1-Distill 7B	0.67 (0.35–0.81)	32	80 (1–100)	89	23	+5.3	0.99
Llama 3.2 1B	0.61 (0.34–0.78)	24	99 (0–100)	100	2	+7.3	0.89
Llama 3.2 3B	0.74 (0.45–0.82)	60	97 (58–100)	100	4	+7.2	1.00
Llama 3.1 8B	0.84 (0.76–0.88)	100	24 (0–68)	54	89	-0.4	0.99
Phi-3.5 mini	0.63 (0.49–0.81)	26	79 (3–99)	84	23	+4.6	1.00
Mistral 7B	0.80 (0.55–0.87)	92	85 (21–100)	98	21	+6.0	1.00
gpt-oss 20B	0.81 (0.59–0.85)	88	14 (0–46)	31	95	-0.8	0.91

Supplementary Table 10: Variance components of the expected item sum (person $\times$ model $\times$ template $\times$ format $\times$ view).

Source	$\hat{\sigma}^2$ (raw)	$\hat{\sigma}^2$	Share (%)
model	6.943	6.943	30.0
model $\times$ format	3.645	3.645	15.8
person	2.437	2.437	10.5
person $\times$ model	1.872	1.872	8.1
template	1.779	1.779	7.7
format	1.363	1.363	5.9
model $\times$ template	0.971	0.971	4.2
model $\times$ template $\times$ format	0.691	0.691	3.0
person $\times$ model $\times$ format $\times$ view	0.588	0.588	2.5
person $\times$ model $\times$ format	0.498	0.498	2.2
person $\times$ model $\times$ template $\times$ format $\times$ view	0.429	0.429	1.9
person $\times$ format	0.413	0.413	1.8
person $\times$ model $\times$ template	0.244	0.244	1.1
model $\times$ format $\times$ view	0.234	0.234	1.0
person $\times$ model $\times$ template $\times$ format	0.215	0.215	0.9
person $\times$ model $\times$ template $\times$ view	0.156	0.156	0.7
person $\times$ model $\times$ view	0.155	0.155	0.7
person $\times$ view	0.143	0.143	0.6
model $\times$ template $\times$ format $\times$ view	0.138	0.138	0.6
person $\times$ template	0.076	0.076	0.3
template $\times$ format	0.043	0.043	0.2
person $\times$ template $\times$ format	0.021	0.021	0.1
person $\times$ template $\times$ view	0.014	0.014	0.1
model $\times$ view	0.014	0.014	0.1
model $\times$ template $\times$ view	0.014	0.014	0.1
template $\times$ view	0.014	0.014	0.1
person $\times$ format $\times$ view	0.006	0.006	0.0
person $\times$ template $\times$ format $\times$ view	0.002	0.002	0.0
template $\times$ format $\times$ view	-0.000	0.000	0.0
format $\times$ view	-0.010	0.000	0.0
view	-0.009	0.000	0.0

Supplementary Table 11: Decision studies (discovery; both transcript views averaged). For one view, averaging over all models, templates and formats gives $E\rho^2 = 0.82$ and $\Phi = 0.53$; a single configuration with one view gives $\Phi = 0.11$.

Models	Templates	Formats	$E\rho^2$	Φ
1	1	1	0.37	0.11
3	1	1	0.58	0.21
11	1	1	0.73	0.32
1	1	3	0.45	0.14
3	1	3	0.67	0.27
11	1	3	0.82	0.40
1	5	1	0.42	0.13
3	5	1	0.62	0.26
11	5	1	0.76	0.41
1	5	3	0.49	0.17
3	5	3	0.71	0.34
11	5	3	0.85	0.55

Supplementary Table 12: Directional predictions specified before the analysis, and their outcomes.

	Prediction	Criterion	Discovery		Confirmation	
H1	Configuration contributes more variance to the AUC than participant sampling	ratio > 1	10.6	supported	5.6	supported
H2	Participants with discordant decisions among acceptable configurations	$\geq 25\%$	100%	supported	100%	supported
H3	Range of implied prevalence across configurations	≥ 30 pp	100 pp	supported	100 pp	supported
H4	Person share of item-sum variance	< 50%	10.5%	supported	13.0%	supported
H5	Rank correlation of Cronbach's α with the AUC	< 0.30	0.02	supported	-0.06	supported
H6	Median-model AUC from interviewer speech only	> 0.60	0.59	not supported	—	—
H7	Prevalence range after calibration with 40 labels, against before	at most half	58 vs 100 pp	not supported	65 vs 100 pp	not supported

3 Labels and transcripts

All analyses use the registered definition of a positive screen, a PHQ-8 total of 10 or more. It differs from the binary flag distributed with the corpora for one discovery participant and five confirmation participants. Item scores were available for 142 discovery participants and complete for 141. In six discovery transcripts, empty turns had been converted to the literal token ‘nan’ (1–12 tokens per transcript); excluding these six participants changed the median AUC across configurations from 0.753 to 0.757, and the frozen loader was not changed.

4 Deviations from the pre-registered plan

Regenerated outputs. Two secondary outputs of the first run, for Qwen2.5 1.5B (the interviewer’s speech for one participant and the automatic transcript for another), contained no finite probabilities, and the median over models used for the interviewer-speech prediction silently omitted that model. Both outputs were regenerated with the frozen code, giving finite outputs with option mass 0.995 and 0.990; the original files are kept. Re-running the frozen analysis left the primary results unchanged and changed the interviewer-speech prediction from supported (median AUC 0.61) to not supported (0.59).

Direct-total termination. The frozen code applied the end-of-number factor only to numbers whose tokens begin a longer candidate, not to every number as described (Supplementary Table 3). The code was not changed; in the re-scored sample (four models, 25 participants, participant view), the effect on expected totals was at most 0.008 points, with no change of decision; this measures the defect in that sample rather than bounding it for other models or views.

Test attribution and degrees of freedom. The plan names Steiger’s test for comparing a configuration’s correlations with the PHQ-8 and with PTSD severity; the code implements the Meng–Rosenthal–Rubin form⁶⁸. Steiger’s pooled

form⁶⁹ and a paired participant bootstrap give the same shares of configurations (3.9% and 4.3% in discovery; 1.1% and 0.9% in confirmation). The partial Spearman tests of the regularities used $n - 2$ degrees of freedom; with $n - 3$, every regularity still meets its criterion (PTSD regularity in confirmation, $P = 0.049$).

Decision-study views. In discovery the decision studies average over both transcript views; values for one view are given with Supplementary Table 11.

During data collection the analysis code was changed once so that interim runs on an incomplete set of outputs recorded the interviewer-speech prediction as missing instead of stopping. The change was reverted before the registered analyses, which used the code as registered.

The local copies of three model weights (Qwen2.5 1.5B, 3B and 7B) were lost during data collection and obtained again from the same repositories. Re-scoring two participants with the new copy of Qwen2.5 1.5B reproduced the earlier outputs exactly.

Scoring runs were interrupted and resumed several times, and the order in which models were run changed. Answers were read from next-token probabilities without sampling, and every participant was scored once per model and view; across batch compositions, scores agree to the tolerance in Supplementary Table 3.

The two addenda (a label-free rule and nine regularities) were specified after the discovery analysis and before any confirmation output was compared with the criterion, while confirmation outputs were being generated. Analyses added after the confirmation analysis are listed in Methods and are exploratory.

References

1. Kendell, R. E. Diagnostic criteria of american and british psychiatrists. *Archives of General Psychiatry* **25**, 123 (1971). <https://doi.org/10.1001/archpsyc.1971.01750140027006>.
2. Spitzer, R. L. & Fleiss, J. L. A re-analysis of the reliability of psychiatric diagnosis. *British Journal of Psychiatry* **125**, 341–347 (1974). <https://doi.org/10.1192/bjp.125.4.341>.
3. Feighner, J. P. Diagnostic criteria for use in psychiatric research. *Archives of General Psychiatry* **26**, 57 (1972). <https://doi.org/10.1001/archpsyc.1972.01750190059011>.
4. Spitzer, R. L. Research diagnostic criteria: Rationale and reliability. *Archives of General Psychiatry* **35**, 773 (1978). <https://doi.org/10.1001/archpsyc.1978.01770300115013>.
5. Obradovich, N. *et al.* Opportunities and risks of large language models in psychiatry. *NPP—Digital Psychiatry and Neuroscience* **2**, 8 (2024). <https://doi.org/10.1038/s44277-024-00010-z>.
6. Stade, E. C. *et al.* Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *npj Mental Health Research* **3**, 12 (2024). <https://doi.org/10.1038/s44184-024-00056-z>.
7. Demszky, D. *et al.* Using large language models in psychology. *Nature Reviews Psychology* (2023). <https://doi.org/10.1038/s44159-023-00241-5>.
8. Galatzer-Levy, I. R., McDuff, D., Natarajan, V., Karthikesalingam, A. & Malgaroli, M. The capability of large language models to measure psychiatric functioning. Preprint at <https://arxiv.org/abs/2308.01834> (2023).
9. Kebe, G. Y. *et al.* LLAMADRS: Evaluating open-source LLMs on real clinical interviews—to reason or not to reason? *Transactions of the Association for Computational Linguistics* **14**, 1589–1604 (2026). <https://aclanthology.org/2026.tacl-1.71/>.
10. Zhu, T. *et al.* Evaluating large language models for assessment of psychosis risk. *npj Digital Medicine* **9**, 554 (2026). <https://doi.org/10.1038/s41746-026-02928-4>.
11. Chen, A. X., Horga, G. & Escola, S. Using large language models to measure symptom severity scores in patients at-risk for schizophrenia. *Schizophrenia Bulletin* **52**, sbag084 (2026). <https://doi.org/10.1093/schbul/sbag084>.
12. Rosenman, G., Hendler, T. & Wolf, L. LLM questionnaire completion for automatic psychiatric assessment. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 403–415 (Association for Computational Linguistics, 2024). <https://doi.org/10.18653/v1/2024.findings-emnlp.23>.
13. Teferra, B. G. *et al.* Leveraging large language models for automated depression screening. *PLOS Digital Health* **4**, e0000943 (2025). <https://doi.org/10.1371/journal.pdig.0000943>.
14. Lin, B., Cecchi, G. & Bouneffouf, D. Deep annotation of therapeutic working alliance in psychotherapy. In Shaban-Nejad, A., Michalowski, M. & Bianco, S. (eds.) *Artificial Intelligence for Personalized Medicine: Promoting Healthy Living and Longevity*, vol. 1106 of *Studies in Computational Intelligence*, 193–207 (Springer, Cham, 2023).
15. Lin, B. *et al.* COMPASS: Computational mapping of patient-therapist alliance strategies with language modeling. *Translational Psychiatry* **15**, 166 (2025). <https://doi.org/10.1038/s41398-025-03379-3>.
16. Lin, B., Bouneffouf, D., Cecchi, G. & Tejwani, R. Neural topic modeling of psychotherapy sessions. In Shaban-Nejad, A., Michalowski, M. & Bianco, S. (eds.) *Artificial Intelligence for Personalized Medicine: Promoting Healthy Living and Longevity*, vol. 1106 of *Studies in Computational Intelligence*, 209–219 (Springer, Cham, 2023).
17. Sclar, M., Choi, Y., Tsvetkov, Y. & Suhr, A. Quantifying language models’ sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations* (2024). <https://openreview.net/forum?id=RIu5lyNXjT>.
18. Mizrahi, M. *et al.* State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics* **12**, 933–949 (2024). https://doi.org/10.1162/tacl_a_00681.
19. Zheng, C., Zhou, H., Meng, F., Zhou, J. & Huang, M. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations* (2024). <https://openreview.net/forum?id=shr9PXz7T0>.
20. Tjuatja, L., Chen, V., Wu, T., Talwalkar, A. & Neubig, G. Do LLMs exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics* **12**, 1011–1026 (2024). https://doi.org/10.1162/tacl_a_00685.

21. Pugh, S. L. *et al.* Assessing dimensions of thought disorder with large language models: The tradeoff of accuracy and consistency. *Psychiatry Research* **341**, 116119 (2024). <https://doi.org/10.1016/j.psychres.2024.116119>.
22. Atil, B. *et al.* Non-determinism of "deterministic" LLM settings. Preprint at <https://arxiv.org/abs/2408.04667> (2024).
23. Baumann, J. *et al.* Large language model hacking: Quantifying the hidden risks of using LLMs for text annotation. Preprint at <https://arxiv.org/abs/2509.08825> (2025).
24. Lorenzoni, G., Velmovitsky, P. E., Alencar, P. & Cowan, D. Gpt-4 on clinic depression assessment: An llm-based pilot study. Preprint at <https://arxiv.org/abs/2501.00199> (2025).
25. Alkaeed, M., Qayyum, A., Kashreef, N. A., Bilal, M. & Qadir, J. Same patient, different words, different diagnosis? evaluating semantic stability in clinical LLMs. Preprint at <https://arxiv.org/abs/2605.30646> (2026).
26. Xu, J. & Hu, X. Language shapes mental health evaluations in large language models. Preprint at <https://arxiv.org/abs/2603.06910> (2026).
27. Norman, J. D., Rivera, M. U. & Hughes, D. A. Reliability without validity: A systematic, large-scale evaluation of LLM-as-a-Judge models across agreement, consistency, and bias. Preprint at <https://arxiv.org/abs/2606.19544> (2026).
28. Batzdorfer, V. & Santagiustina, C. R. M. A. Reproducibility is not construct validity: LLM measurement of institutionally situated communication. Preprint at <https://arxiv.org/abs/2609.19866> (2026).
29. Lin, Z. From prompts to constructs: A dual-validity framework for large language model research in psychology. *Annual Review of Psychology* (2026). <https://doi.org/10.1146/annurev-psych-100925-034807>.
30. Cronbach, L. J., Gleser, G. C., Nanda, H. & Rajaratnam, N. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles* (Wiley, New York, 1972). <https://gwern.net/doc/psychology/1972-cronbach-the-dependability-of-behavioral-measurements.pdf>.
31. Brennan, R. L. *Generalizability Theory* (Springer New York, 2001). <https://doi.org/10.1007/978-1-4757-3456-0>.
32. Steegen, S., Tuerlinckx, F., Gelman, A. & Vanpaemel, W. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science* **11**, 702–712 (2016). <https://doi.org/10.1177/1745691616658637>.
33. Simonsohn, U., Simmons, J. P. & Nelson, L. D. Specification curve analysis. *Nature Human Behaviour* **4**, 1208–1214 (2020). <https://doi.org/10.1038/s41562-020-0912-z>.
34. Marx, C. T., Calmon, F. d. P. & Ustun, B. Predictive multiplicity in classification. In *Proceedings of the 37th International Conference on Machine Learning*, vol. 119 of *Proceedings of Machine Learning Research*, 6765–6774 (2020). <https://proceedings.mlr.press/v119/marx20a.html>.
35. Black, E., Raghavan, M. & Barocas, S. Model multiplicity: Opportunities, concerns, and solutions. In *2022 ACM Conference on Fairness Accountability and Transparency, FAccT '22*, 850–863 (ACM, 2022). <https://doi.org/10.1145/3531146.3533149>.
36. Du, A. Y., Ngo, D. D. & Wu, Z. S. Reconciling model multiplicity for downstream decision making. In *The Thirteenth International Conference on Learning Representations* (2025). <https://openreview.net/forum?id=uy4EavBEwl>.
37. Van Calster, B. *et al.* Calibration: the achilles heel of predictive analytics. *BMC Medicine* **17**, 230 (2019). <https://doi.org/10.1186/s12916-019-1466-7>.
38. Kroenke, K. *et al.* The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders* **114**, 163–173 (2009). <https://doi.org/10.1016/j.jad.2008.06.026>.
39. Nosek, B. A., Ebersole, C. R., DeHaven, A. C. & Mellor, D. T. The preregistration revolution. *Proceedings of the National Academy of Sciences* **115**, 2600–2606 (2018). <https://doi.org/10.1073/pnas.1708274114>.
40. MacCoun, R. & Perlmutter, S. Blind analysis: Hide results to seek the truth. *Nature* **526**, 187–189 (2015). <https://doi.org/10.1038/526187a>.
41. Burdisso, S. *et al.* DAIC-WOZ: On the validity of using the therapist's prompts in automatic depression detection from clinical interviews. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, 82–90 (Association for Computational Linguistics, 2024). <https://doi.org/10.18653/v1/2024.clinicalnlp-1.8>.
42. Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I. & Zrnic, T. Prediction-powered inference. *Science* **382**, 669–674 (2023). <https://doi.org/10.1126/science.adi6000>.
43. Lockfisch, S. *et al.* On arbitrary predictions from equally valid models. Preprint at <https://arxiv.org/abs/2507.19408> (2025).

44. Lechuga López, L. J., Shamout, F. E. & Rudner, T. G. J. An empirical analysis of calibration and selective prediction in multimodal clinical condition classification. In *Proceedings of the 7th Conference on Health, Inference, and Learning*, vol. 333 of *Proceedings of Machine Learning Research*, 794–833 (2026). <https://proceedings.mlr.press/v333/lopez26a.html>.
45. Lin, B., Cecchi, G. & Bouneffouf, D. Working alliance transformer for psychotherapy dialogue classification. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, 64–69 (Association for Computational Linguistics, Mexico City, 2024).
46. Lin, B., Cecchi, G. & Bouneffouf, D. SupervisorBot: NLP-annotated real-time recommendations of psychotherapy treatment strategies with deep reinforcement learning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 7149–7153 (International Joint Conferences on Artificial Intelligence Organization, 2023).
47. Lin, B., Cecchi, G. & Bouneffouf, D. Psychotherapy AI companion with reinforcement learning recommendations and interpretable policy dynamics. In *Companion Proceedings of the ACM Web Conference 2023*, 932–939 (Association for Computing Machinery, New York, 2023).
48. Gunal, A., Lin, B. & Bouneffouf, D. Conversational topic recommendation in counseling and psychotherapy with decision transformer and large language models. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, 196–201 (Association for Computational Linguistics, Mexico City, 2024).
49. Lin, B. COMPASS 2.0: psychometric representational similarity analysis distinguishes symptom structure from personal signal. Preprint at <https://arxiv.org/abs/2610.06615> (2026). <https://arxiv.org/abs/2610.06615>.
50. Tierney, A. A. *et al.* Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. *NEJM Catalyst* **5** (2024). <https://doi.org/10.1056/cat.23.0404>.
51. Silberzahn, R. *et al.* Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science* **1**, 337–356 (2018). <https://doi.org/10.1177/2515245917747646>.
52. Botvinik-Nezer, R. *et al.* Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* **582**, 84–88 (2020). <https://doi.org/10.1038/s41586-020-2314-9>.
53. Gallifant, J. *et al.* The TRIPOD-LLM reporting guideline for studies using large language models. *Nature Medicine* **31**, 60–69 (2025). <https://doi.org/10.1038/s41591-024-03425-5>.
54. Gratch, J. *et al.* The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 3123–3128 (2014). <https://aclanthology.org/L14-1421/>.
55. DeVault, D. *et al.* SimSensei Kiosk: a virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 1061–1068 (2014). <https://aamas.csc.liv.ac.uk/Proceedings/aamas2014/aamas/p1061.pdf>.
56. Levis, B., Benedetti, A. & Thombs, B. D. Accuracy of patient health questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis. *BMJ* **11476** (2019). <https://doi.org/10.1136/bmj.11476>.
57. Valstar, M. *et al.* AVEC 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, MM '16, 3–10 (ACM, 2016). <https://doi.org/10.1145/2988257.2988258>.
58. Ringeval, F. *et al.* AVEC 2019 workshop and challenge: State-of-mind, detecting depression with AI, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, MM '19, 3–12 (ACM, 2019). <https://doi.org/10.1145/3347320.3357688>.
59. Qwen *et al.* Qwen2.5 technical report. Preprint at <https://arxiv.org/abs/2412.15115> (2024).
60. Grattafiori, A. *et al.* The Llama 3 herd of models. Preprint at <https://arxiv.org/abs/2407.21783> (2024).
61. Jiang, A. Q. *et al.* Mistral 7b. Preprint at <https://arxiv.org/abs/2310.06825> (2023).
62. Abidin, M. *et al.* Phi-3 technical report: A highly capable language model locally on your phone. Preprint at <https://arxiv.org/abs/2404.14219> (2024).
63. Guo, D. *et al.* DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature* **645**, 633–638 (2025). <https://doi.org/10.1038/s41586-025-09422-z>.
64. OpenAI *et al.* gpt-oss-120b & gpt-oss-20b model card. Preprint at <https://arxiv.org/abs/2508.10925> (2025).

65. Wolf, T. *et al.* Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45 (Association for Computational Linguistics, 2020). <https://doi.org/10.18653/v1/2020.emnlp-demos.6>.
66. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36 (1982). <https://doi.org/10.1148/radiology.143.1.7063747>.
67. Gelman, A. & Carlin, J. Beyond power calculations: Assessing Type s (sign) and Type m (magnitude) errors. *Perspectives on Psychological Science* **9**, 641–651 (2014). <https://doi.org/10.1177/1745691614551642>.
68. Meng, X.-l., Rosenthal, R. & Rubin, D. B. Comparing correlated correlation coefficients. *Psychological Bulletin* **111**, 172–175 (1992). <https://doi.org/10.1037/0033-2909.111.1.172>.
69. Steiger, J. H. Tests for comparing elements of a correlation matrix. *Psychological Bulletin* **87**, 245–251 (1980). <https://doi.org/10.1037/0033-2909.87.2.245>.
70. Shrout, P. E. & Fleiss, J. L. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin* **86**, 420–428 (1979). <https://doi.org/10.1037/0033-2909.86.2.420>.