# The development and validation of an automated secondary review model for clinical endpoint interviews

**Michelle Worthington**[1,2], Jacob Ouyang[1], Kieran McVeigh[1], Miriam Evans[3], Vijay Yadav[1,4], Anzar Abbas[1]

[1]**Brooklyn Health**, [2]Yale School of Medicine Department of Psychiatry, [3]Adams Clinical Site Network, [4]University of New South Wales

## Introduction

- Rater variability and subjective scoring compromise data quality and can negatively impact outcomes in CNS trials.
- Currently, human secondary review is used to mitigate rater variability; however, this manual review process poses challenges in standardization and scalability.
- WillisAI, an LLM-based AI model, can bridge this quality control gap and supplement manual secondary review by analyzing transcript context and scoring responses using standard guidelines.
- This study validates an automated LLM-based secondary review model for reviews and scoring of endpoint MADRS interviews in an open-label depression study.

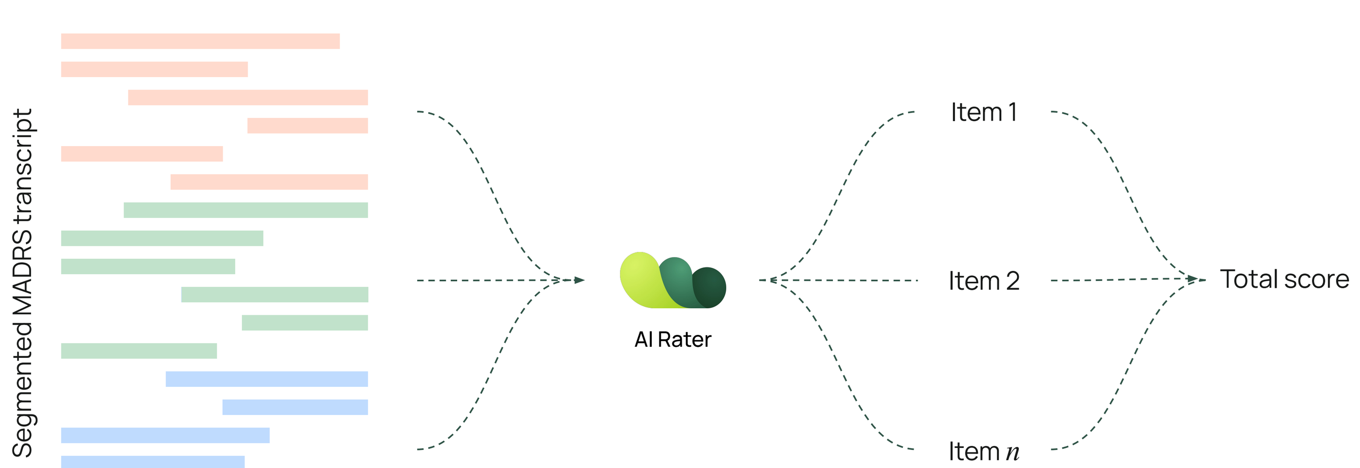

Segmented MADRS transcript
AI Rater
Item 1
Item 2
Item n
Total score

**Figure 1.** AI-based score prediction method based on transcript segmentation and individual item-level score prediction.

## Methods

- 790 MADRS assessments were transcribed and diarized using AWS Transcribe, and segmented into item-level interview segments based on SIGMA interview prompts.
- Each item was scored from 0 to 6 by an LLM (Claude Sonnet 4.5) using MADRS anchors and criteria. Item scores were summed to calculate total scores.
- Scoring agreement was evaluated against human ratings using Pearson's R, MAE, RMSE, and Bland-Altman plots.

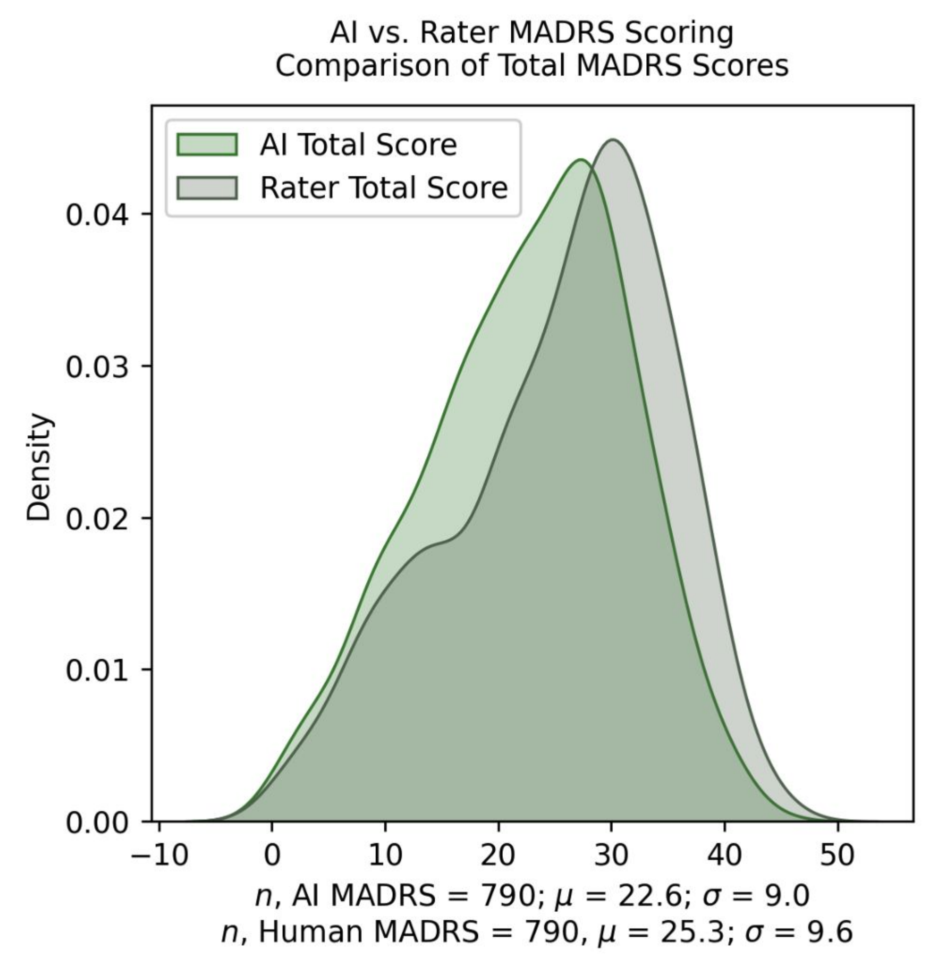

AI vs. Rater MADRS Scoring
Comparison of Total MADRS Scores
AI Total Score
Rater Total Score
Density
n, AI MADRS = 790; μ = 22.6; σ = 9.0
n, Human MADRS = 790, μ = 25.3; σ = 9.6

**Figure 2.** Score distributions for AI-predicted total MADRS scores and Rater total MADRS scores. AI-predicted scores were, on average, **2.7 points** below Rater scores.

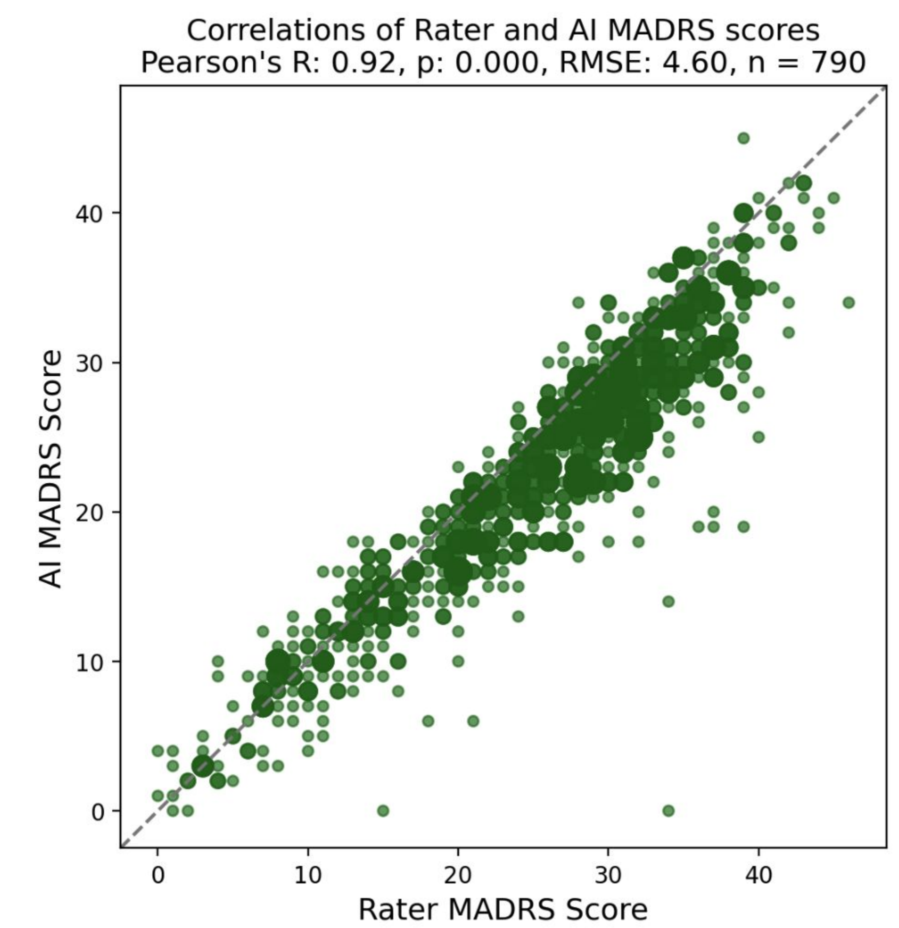

Correlations of Rater and AI MADRS scores
Pearson's R: 0.92, p: 0.000, RMSE: 4.60, n = 790
AI MADRS Score
Rater MADRS Score

**Figure 3.** Scatter plot showing the relationship between Rater total MADRS scores and AI-predicted total MADRS scores. Overall correlation was **0.92.**

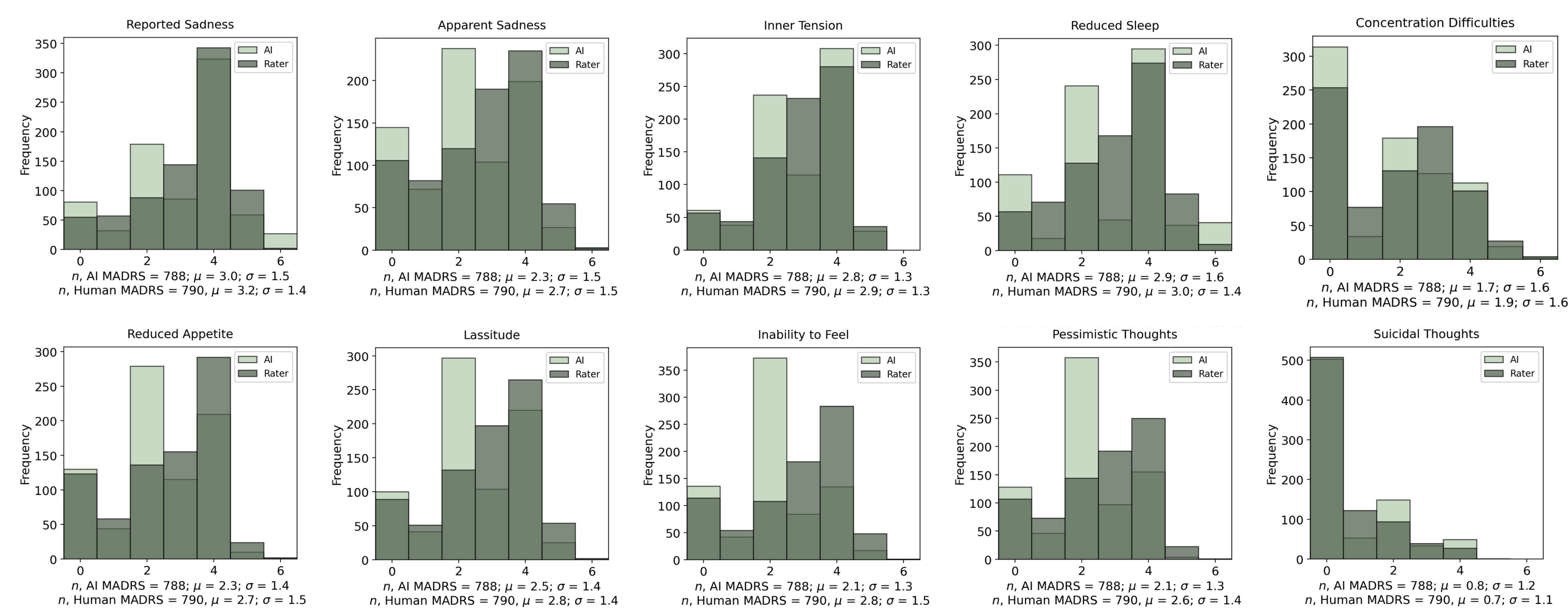

Reported Sadness
n, AI MADRS = 788; μ = 3.0; σ = 1.5
n, Human MADRS = 790, μ = 3.2; σ = 1.4
Apparent Sadness
n, AI MADRS = 788; μ = 2.3; σ = 1.5
n, Human MADRS = 790, μ = 2.7; σ = 1.5
Inner Tension
n, AI MADRS = 788; μ = 2.8; σ = 1.3
n, Human MADRS = 790, μ = 2.9; σ = 1.3
Reduced Sleep
n, AI MADRS = 788; μ = 2.9; σ = 1.6
n, Human MADRS = 790, μ = 3.0; σ = 1.4
Concentration Difficulties
n, AI MADRS = 788; μ = 1.7; σ = 1.6
n, Human MADRS = 790, μ = 1.9; σ = 1.6
Reduced Appetite
n, AI MADRS = 788; μ = 2.3; σ = 1.4
n, Human MADRS = 790, μ = 2.7; σ = 1.5
Lassitude
n, AI MADRS = 788; μ = 2.5; σ = 1.4
n, Human MADRS = 790, μ = 2.8; σ = 1.4
Inability to Feel
n, AI MADRS = 788; μ = 2.1; σ = 1.3
n, Human MADRS = 790, μ = 2.8; σ = 1.5
Pessimistic Thoughts
n, AI MADRS = 788; μ = 2.1; σ = 1.3
n, Human MADRS = 790, μ = 2.6; σ = 1.4
Suicidal Thoughts
n, AI MADRS = 788; μ = 0.8; σ = 1.2
n, Human MADRS = 790, μ = 0.7; σ = 1.1
AI
Rater
Frequency

**Figure 4.** Histograms for distributions of AI-predicted item scores (light green) and Rater item scores (dark green). AI-predicted item scores were, on average, **0.27 points** below Rater item scores.

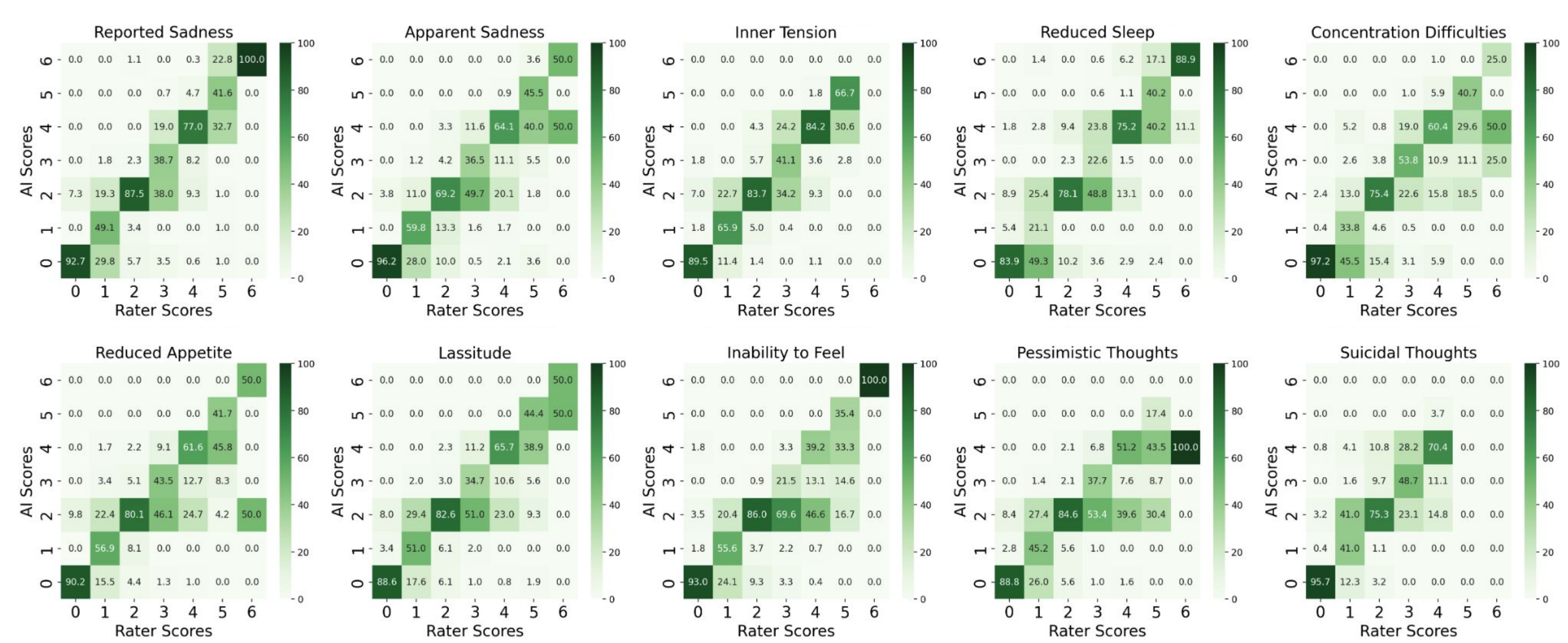

Reported Sadness
Apparent Sadness
Inner Tension
Reduced Sleep
Concentration Difficulties
Reduced Appetite
Lassitude
Inability to Feel
Pessimistic Thoughts
Suicidal Thoughts
AI Scores
Rater Scores

**Figure 5.** Frequency heatmaps showing alignment for individual item score severities between AI-predicted item scores and Rater scores. Frequencies sum to 100% within each level of Rater-assigned item scores to demonstrate how AI-predicted scores tend to vary for each given Rater score.

## Results

- Rater-assigned and AI total scores demonstrated high aggregate correlation (Pearson's R = 0.92, p ‹ 0.001, RMSE = 4.60, Bland-Altman bias = 2.76), which remained consistent longitudinally and across individual raters.
- Across items, strong agreement between rater-assigned and AI item scores was observed (item-level MAE = 0.27), although items requiring greater clinical inference were slightly less aligned.

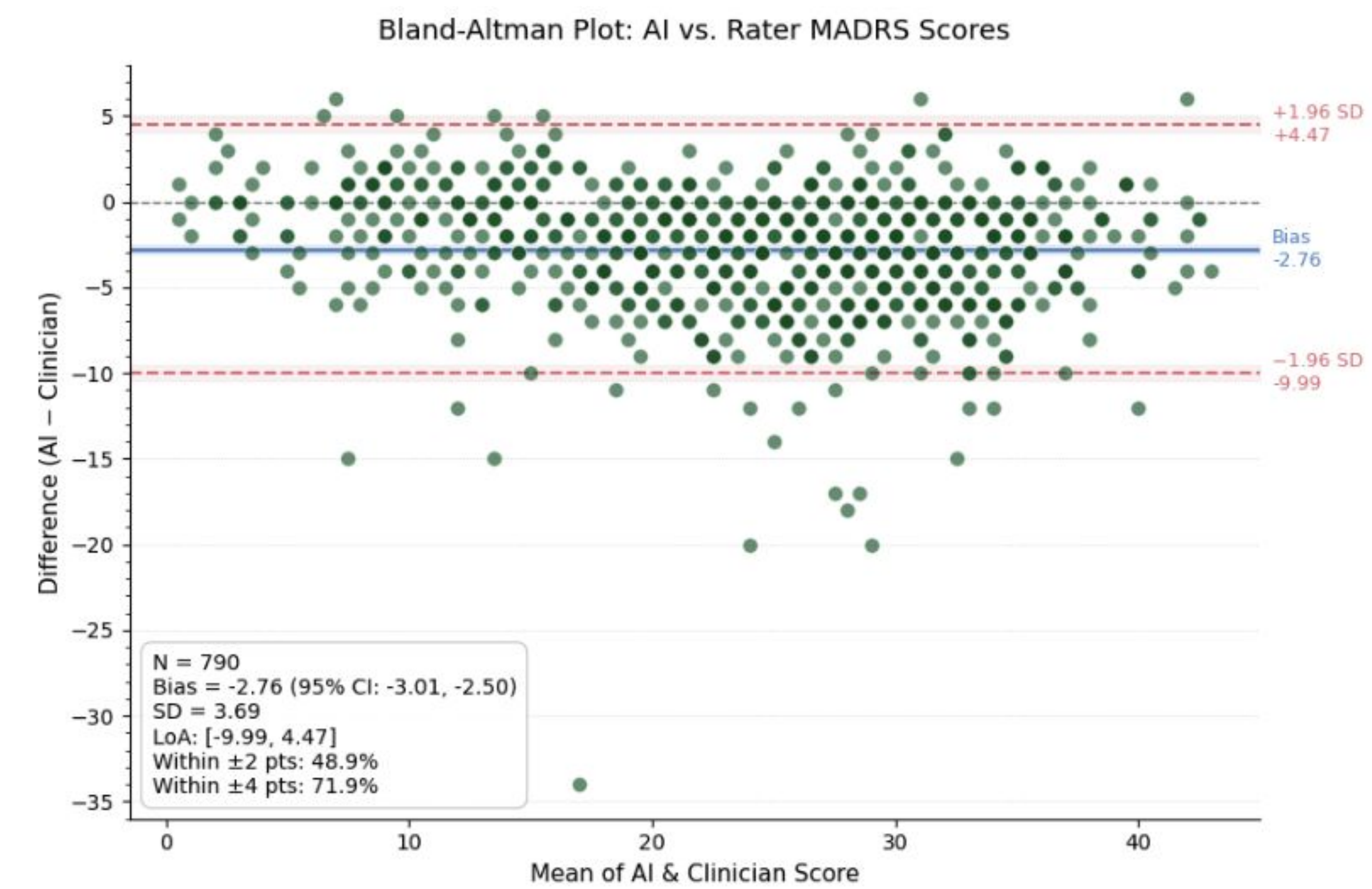

Bland-Altman Plot: AI vs. Rater MADRS Scores
+1.96 SD
+4.47
Bias
-2.76
−1.96 SD
-9.99
Difference (AI − Clinician)
Mean of AI & Clinician Score
N = 790
Bias = -2.76 (95% CI: -3.01, -2.50)
SD = 3.69
LoA: [-9.99, 4.47]
Within ±2 pts: 48.9%
Within ±4 pts: 71.9%

**Figure 6.** Bland-Altman plot showing that AI scores consistently predicted scores lower than Raters by an average of 2.76 points across all severity levels.

## Conclusions

- LLM-based automated secondary review is a scalable, continuous, and low-burden method to independently score clinical interviews.
- High concurrent alignment with human ratings establishes the feasibility of using LLMs as continuous data-auditing mechanisms to protect data quality in CNS trials.

MW, JO, KM, VY, and AA are all current or former employees of and hold equity in Brooklyn Health, a private company that developed and indirectly benefits from the model described in this poster. ME is an employee at Adams Clinical Site Network, where data was collected for the validation study.

**Presented at the 2026 ISCTM Autumn Conference, Philadelphia, PA.**