Novel methods to quantify changes in schizophrenia symptoms using PANSS interview transcripts

Jeff Cochran

Director, Data Scientist

Otsuka Pharmaceutical Development & Commercialization, Inc.

Disclosure

- This presentation represents my own views and opinions.
- I am not representing or speaking on behalf of Otsuka America Pharmaceutical, Inc. (OAPI), Otsuka Pharmaceutical Development & Commercialization, Inc. (OPDC), Otsuka America Inc. (OAI) or its affiliates.

LLMs as a tool for more precise development

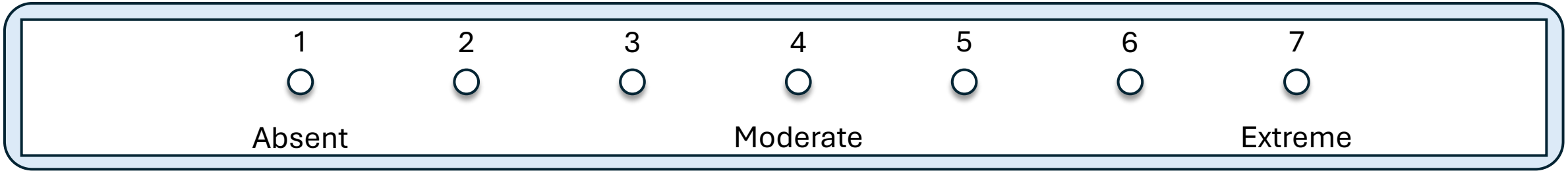
Goal: more precise approaches to drug development, prioritizing the measurement of meaningful differentiated disease and symptom response and functional improvement

Can LLMs be used to support this goal?

Positive and Negative Syndrome Scale (PANSS)

PANSS is a 30-item, clinician rated scale of schizophrenia symptoms that serves as the gold-standard for schizophrenia clinical trials

Each item scored on a scale from 1 to 7



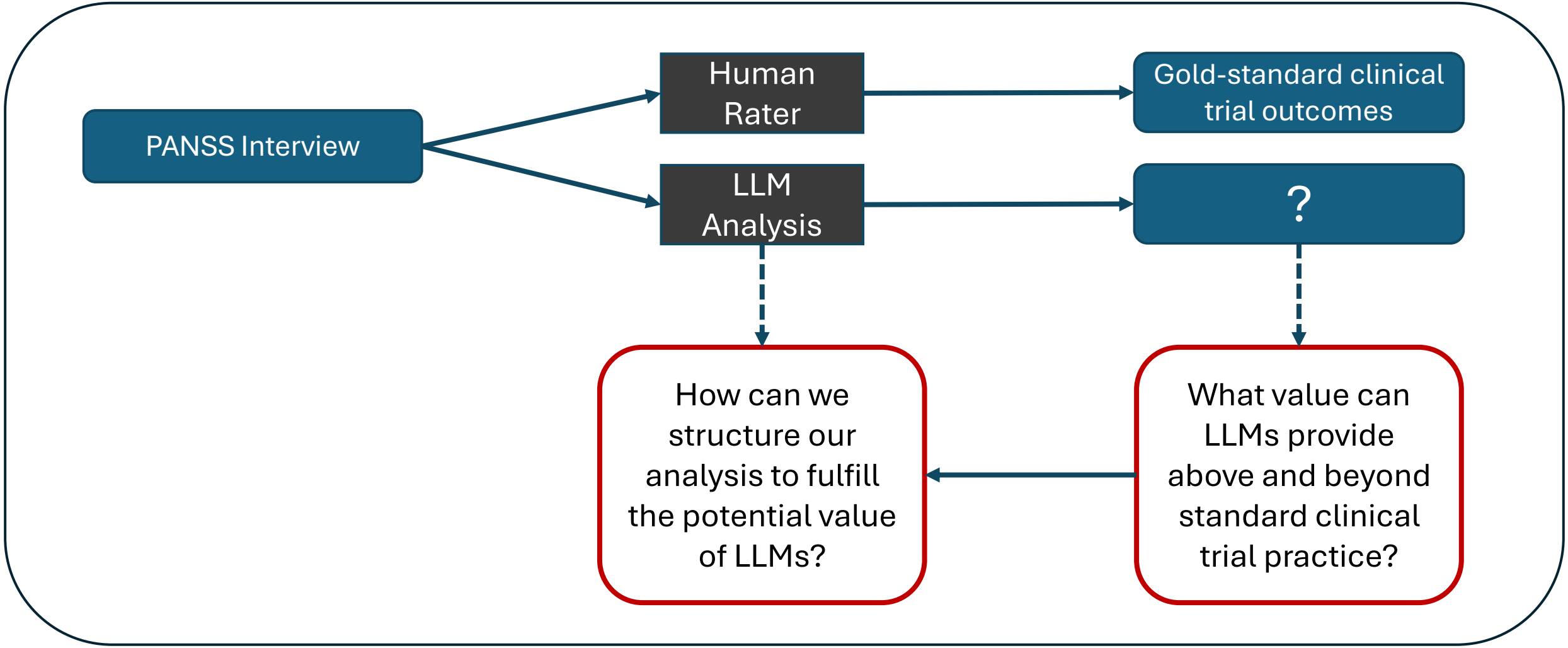
3 Sub-Scales

Positive (e.g., Delusions, Conceptual Disorganization)	**Negative** (e.g., Blunted Affect, Emotional Withdrawal)	**General Psychopathology** (e.g., Somatic Concern, Anxiety)

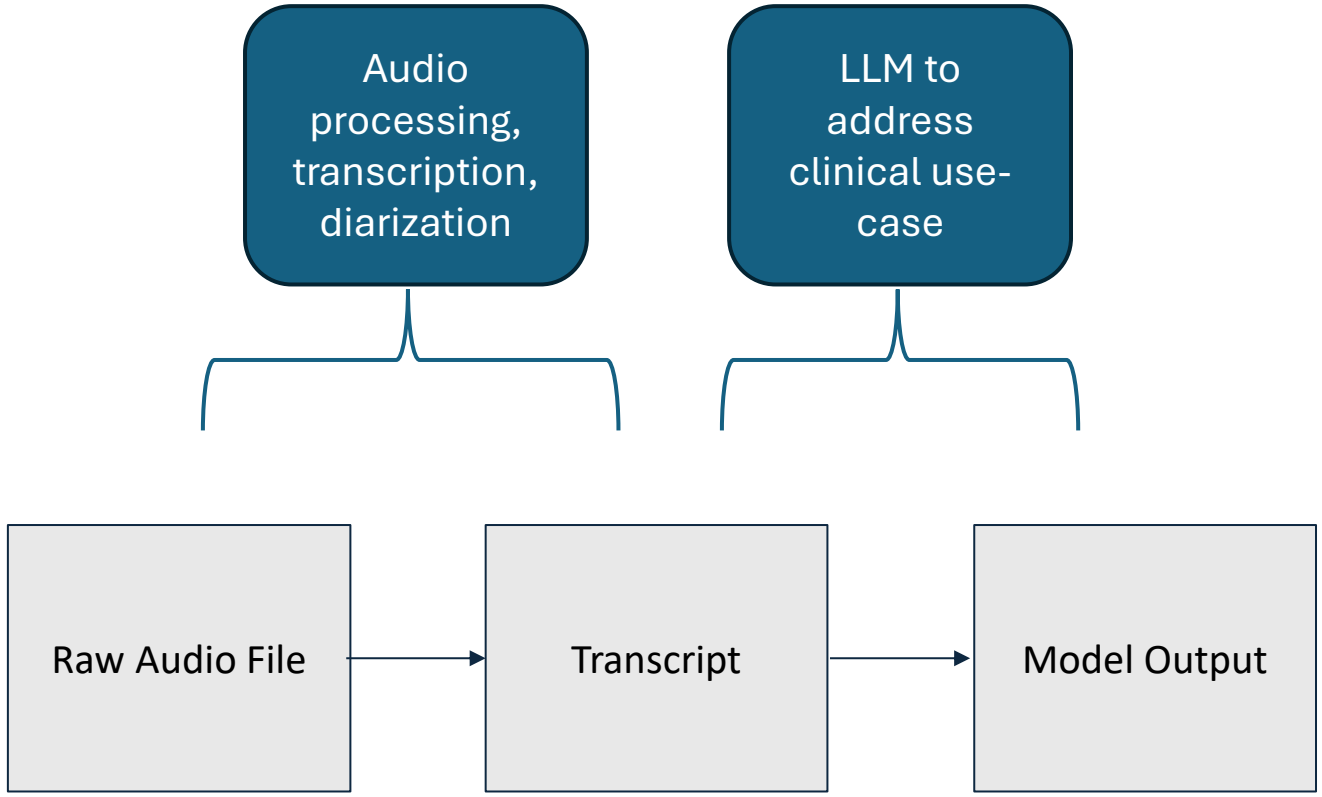
Some item-level information won't be captured in transcripts (e.g. N4, G16)

- Kay, *et al*. *Schizophren Bull*, 1987.
- Opler, *et al*. *Innov Clin Neurosci*, 2017.

How can LLMs supplement available insights from clinical outcomes assessments?



Approach #1: LLM analyzes full transcript



Zero-Shot Learning

Input full transcript into pre-trained LLM with minimal prompting

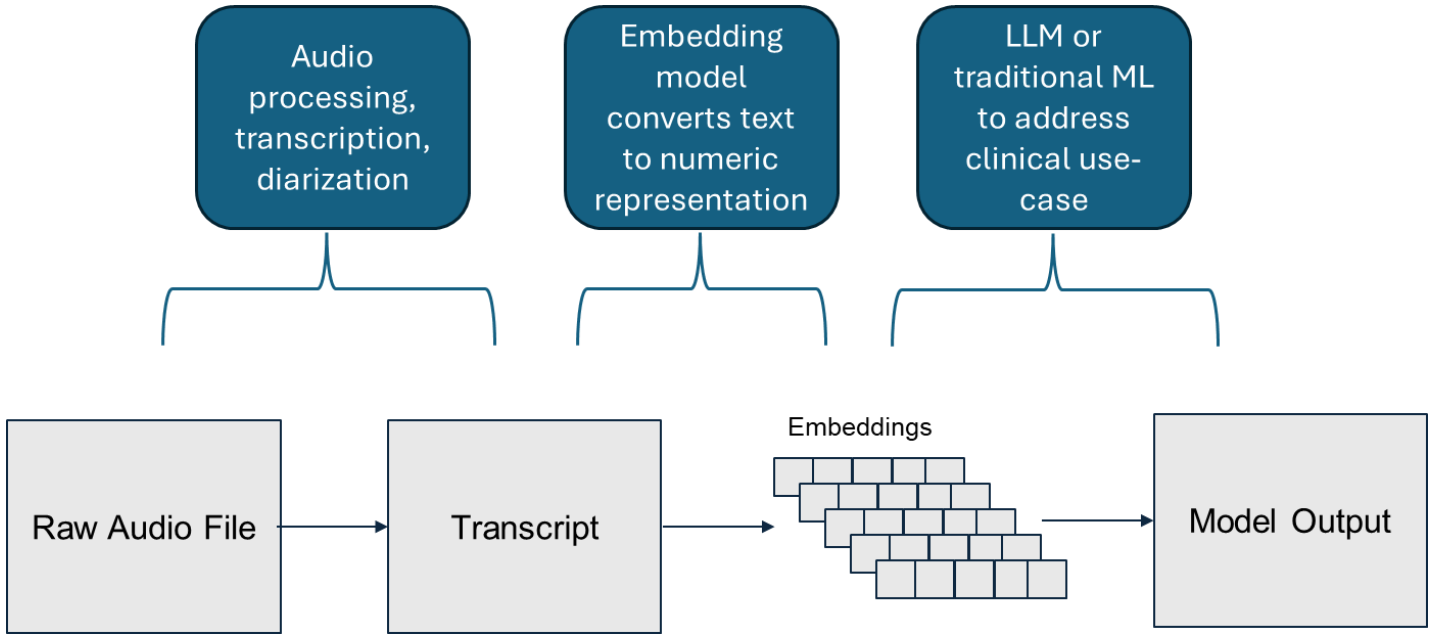
Pros

- Least engineering work
- Doesn't require training datasets

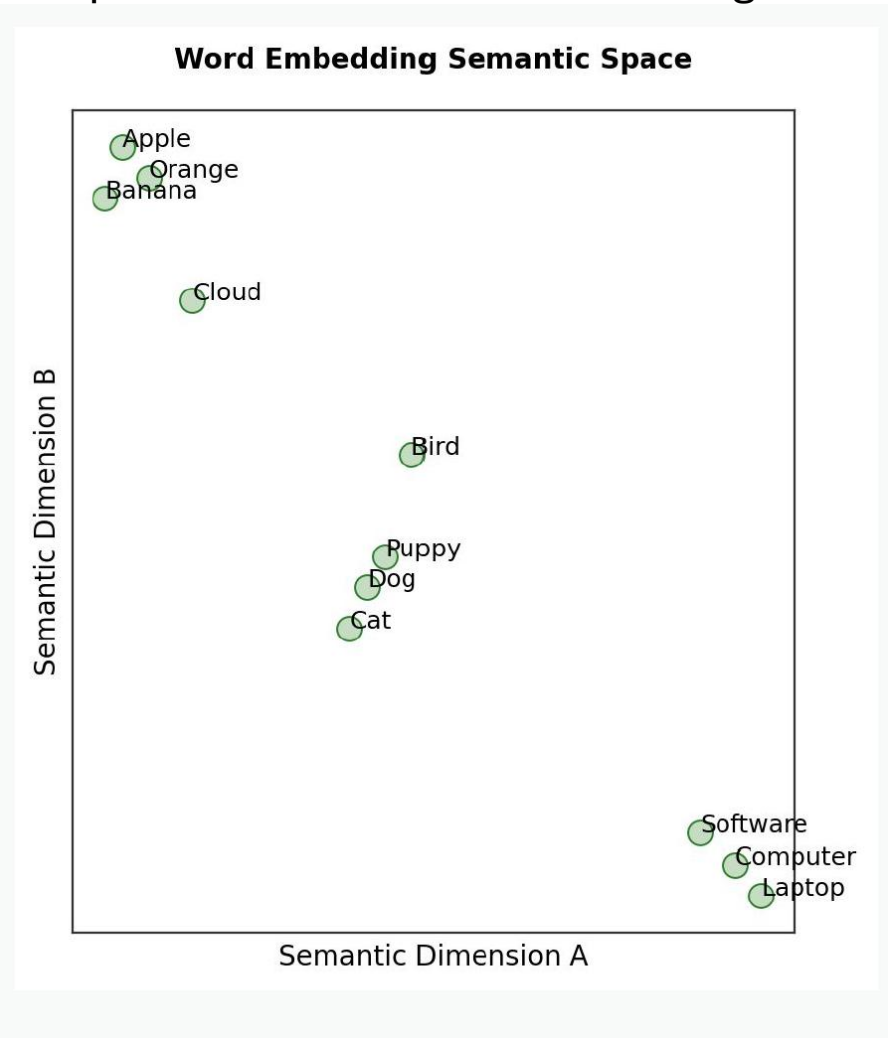
Cons

- Transcripts may be too similar to meaningful differentiate severity
- Minimal interpretability

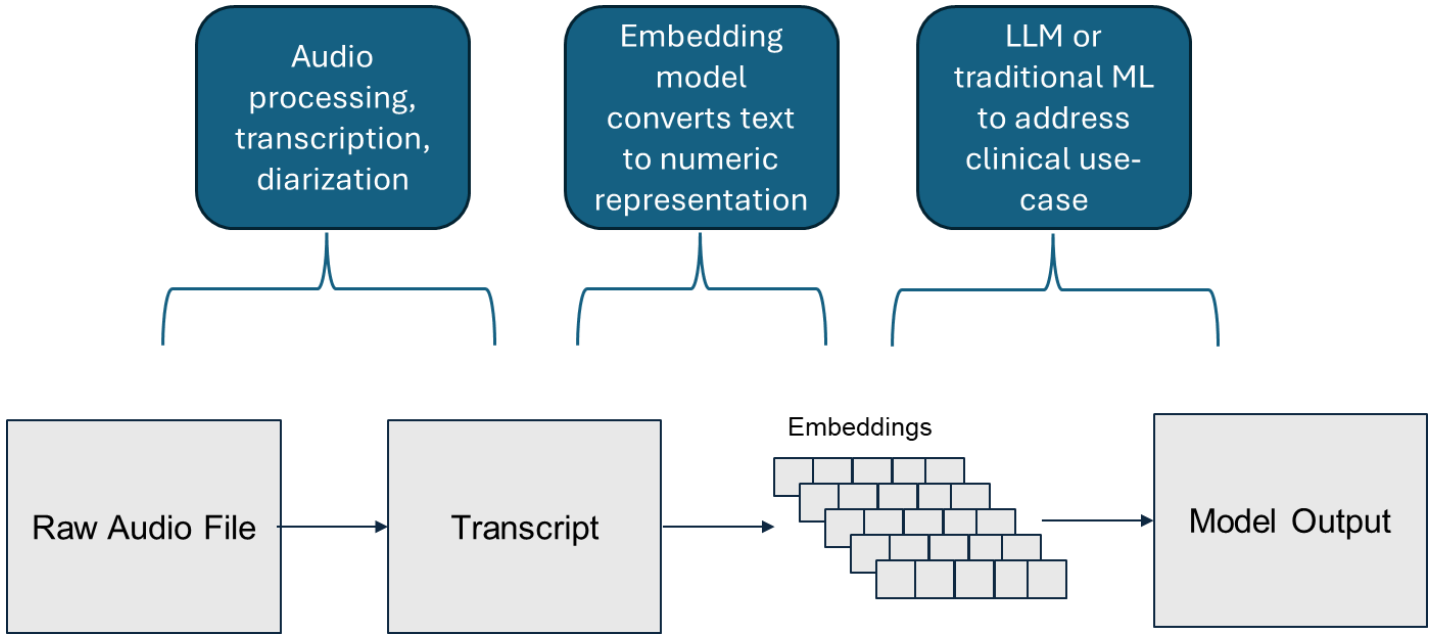
Approach #2: Embedding & Machine Learning



LLM converts language to numeric representations called embeddings



Approach #2: Embedding & Machine Learning



Embedding Regression

Convert text to numeric embeddings and model based on numeric features

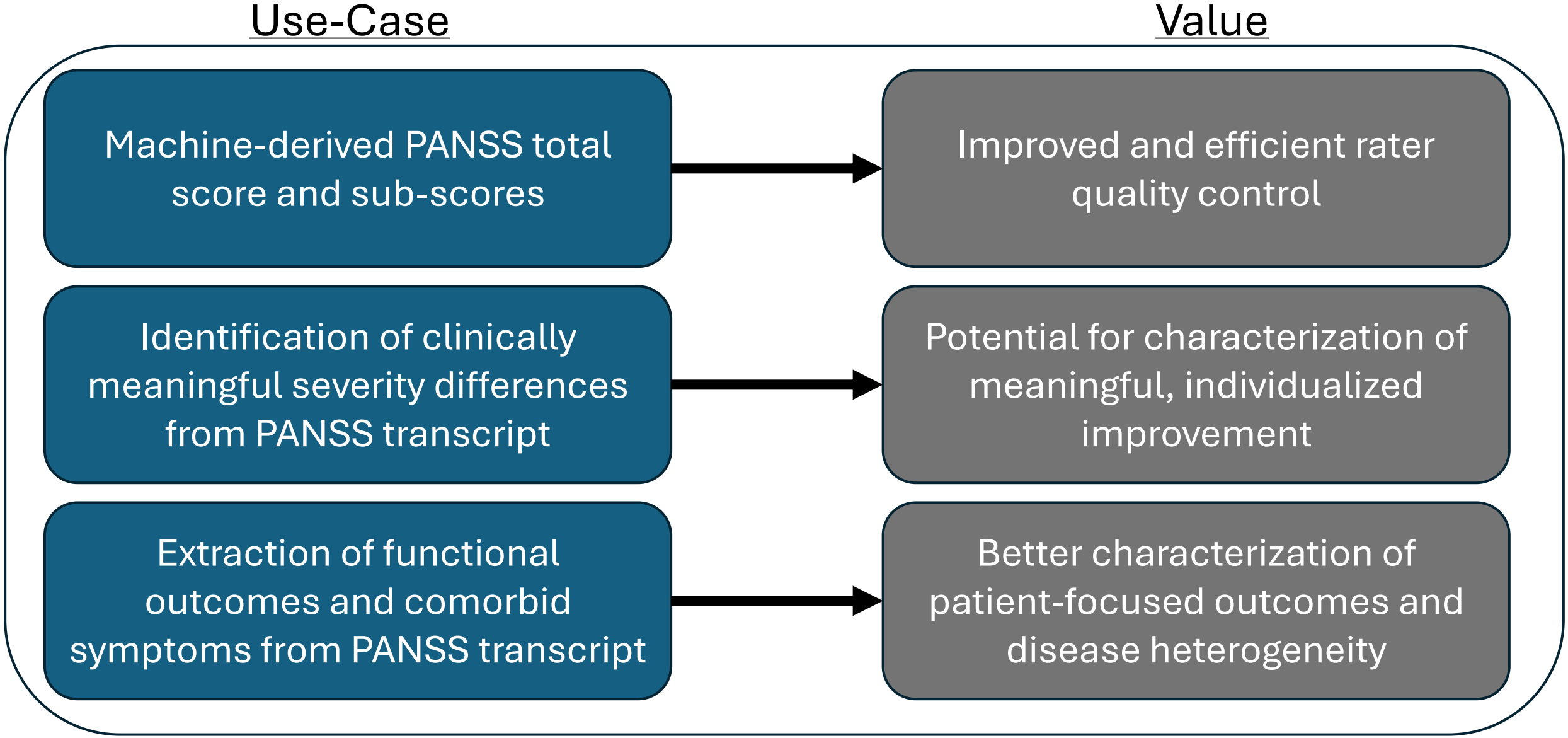
Pros

- Potentially improved conceptual representation
- Potentially more interpretable

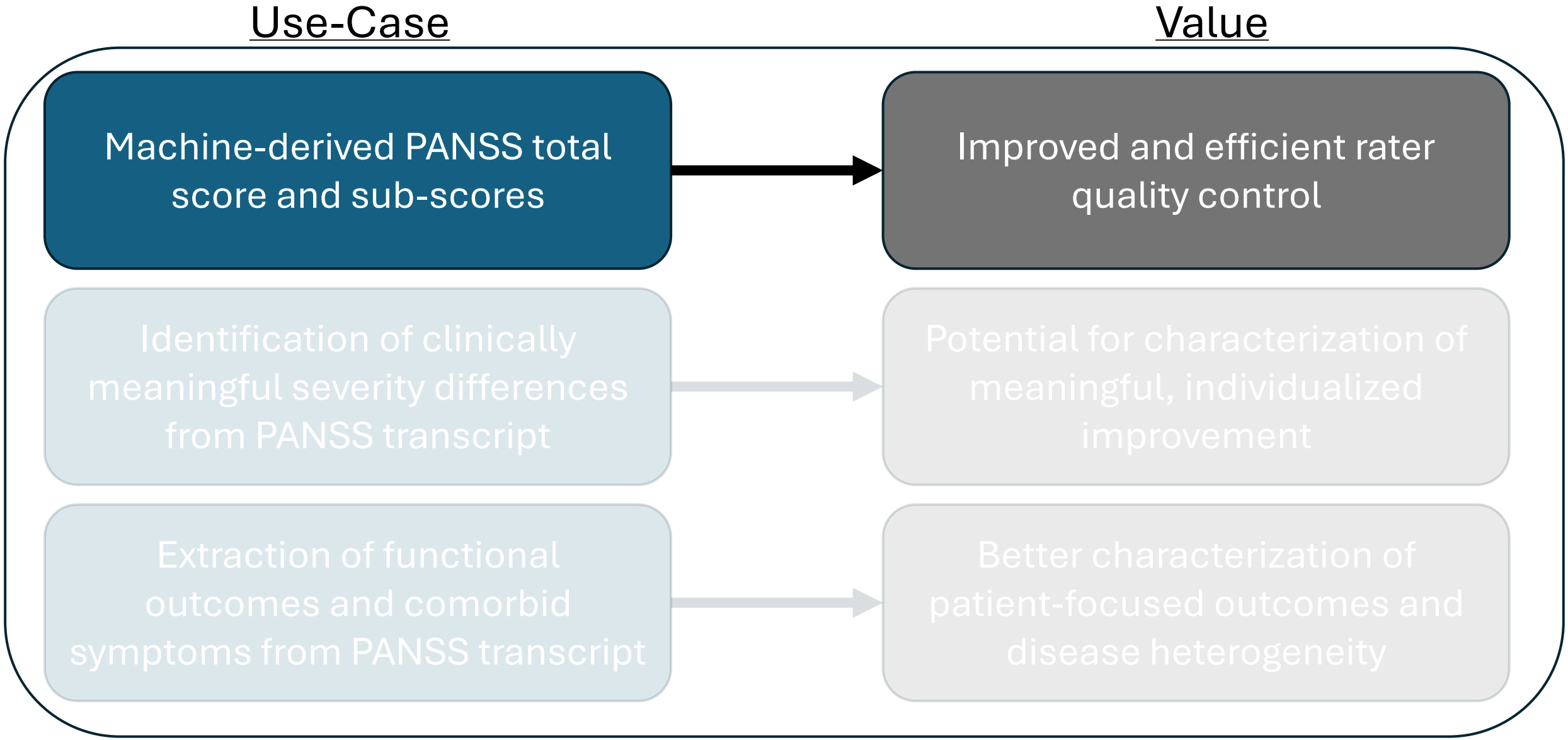
Cons

- More engineering
- Requires model training

Three potential use cases for LLM analysis of PANSS



Three potential use cases for LLM analysis of PANSS



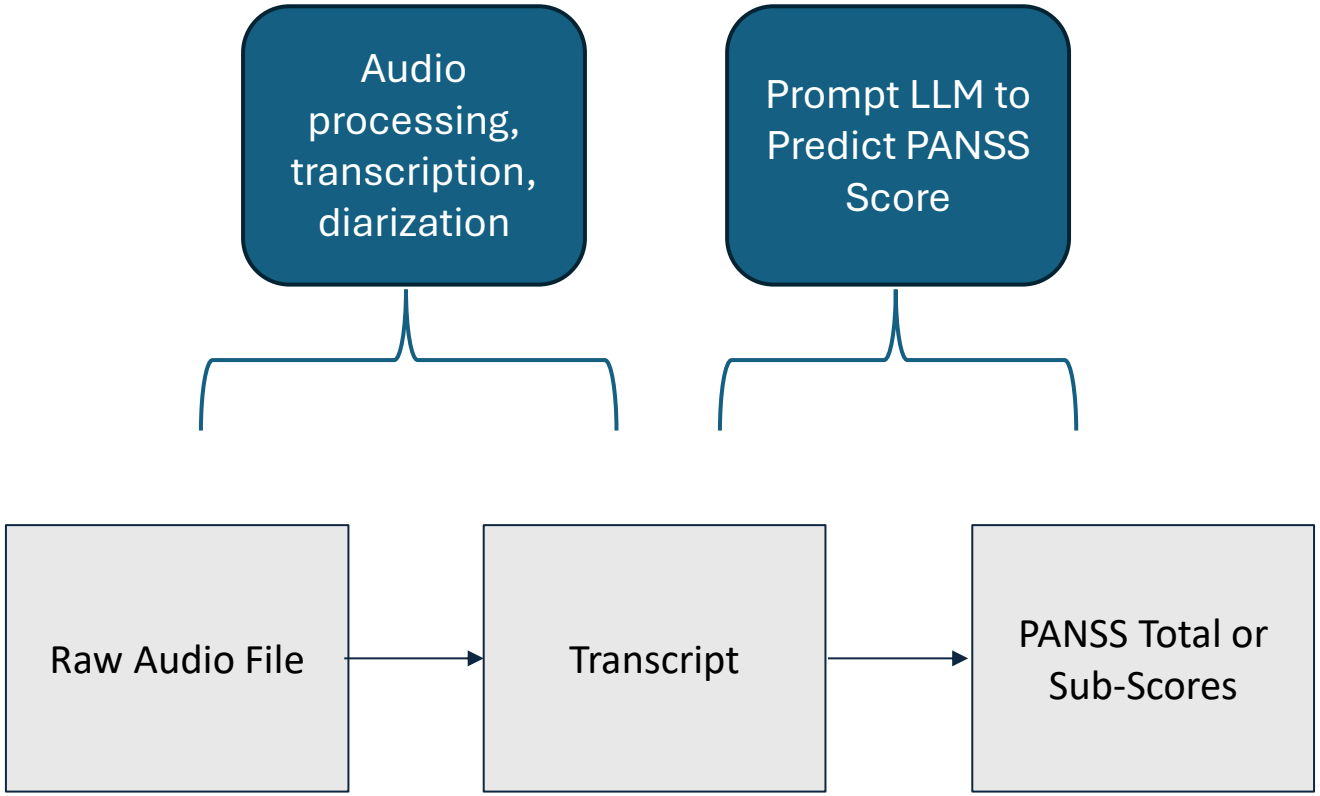
Value: reducing rater variability increases study power

FDA Guidance and best practice is to ensure standardization and comparability of rater assessments

Patient-Focused Drug Development: Incorporating Clinical Outcome Assessments Into Endpoints For Regulatory Decision-Making

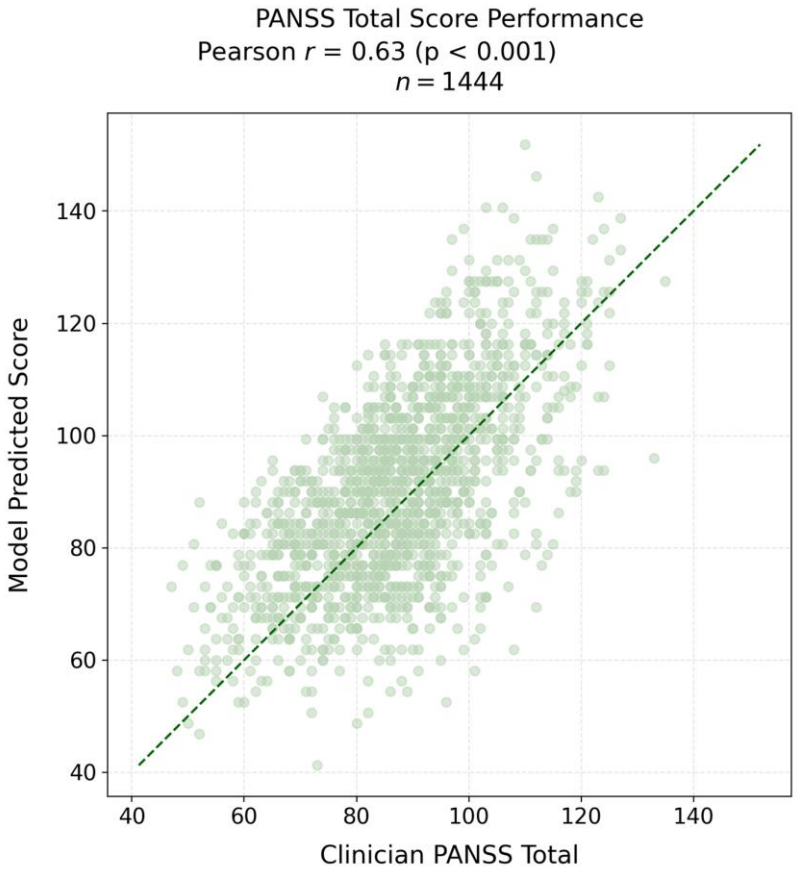
- There might be differences in the measures used to assess the concept(s) of interest, method of COA administration, and/or the COA assessment frequency/schedule that could lead to differences between the groups that is unrelated to the effect of treatment. It is important to establish comparability of the COAs across the groups, to use well-defined and reliable COA-based endpoints in conjunction with standardized rater training and instructions for administration within each comparator arm and across comparator arms. Every effort should be made to ensure comparability in the assessment methods and timing of COA administration, together with the use of standardized data collection methods (e.g., standardized modes of administration).

Approach: LLM directly predicts PANSS scores from transcripts



Results: LLM Predicts Total PANSS Score

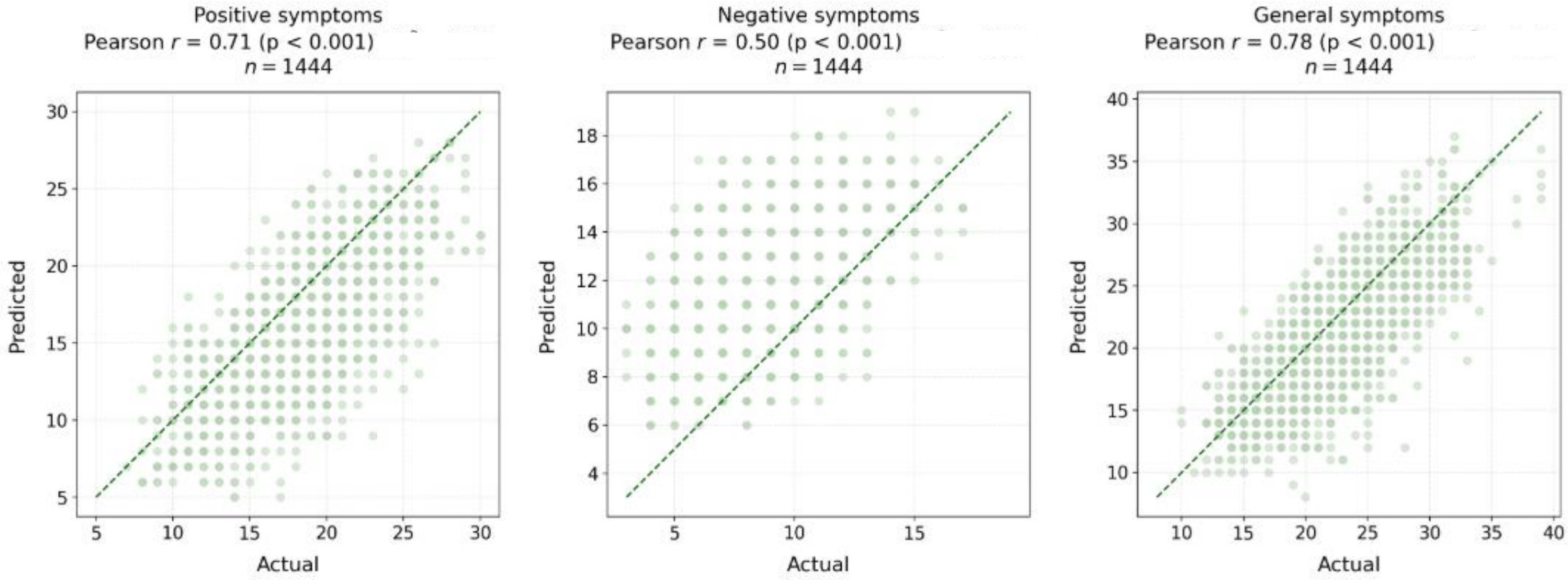
Predicting total PANSS score via LLM with zero-shot prompting



- Model predictions shows strong, but imperfect, correlations with actual PANSS scores
- Prediction error could be due to significant amount of information that isn't captured in the transcript

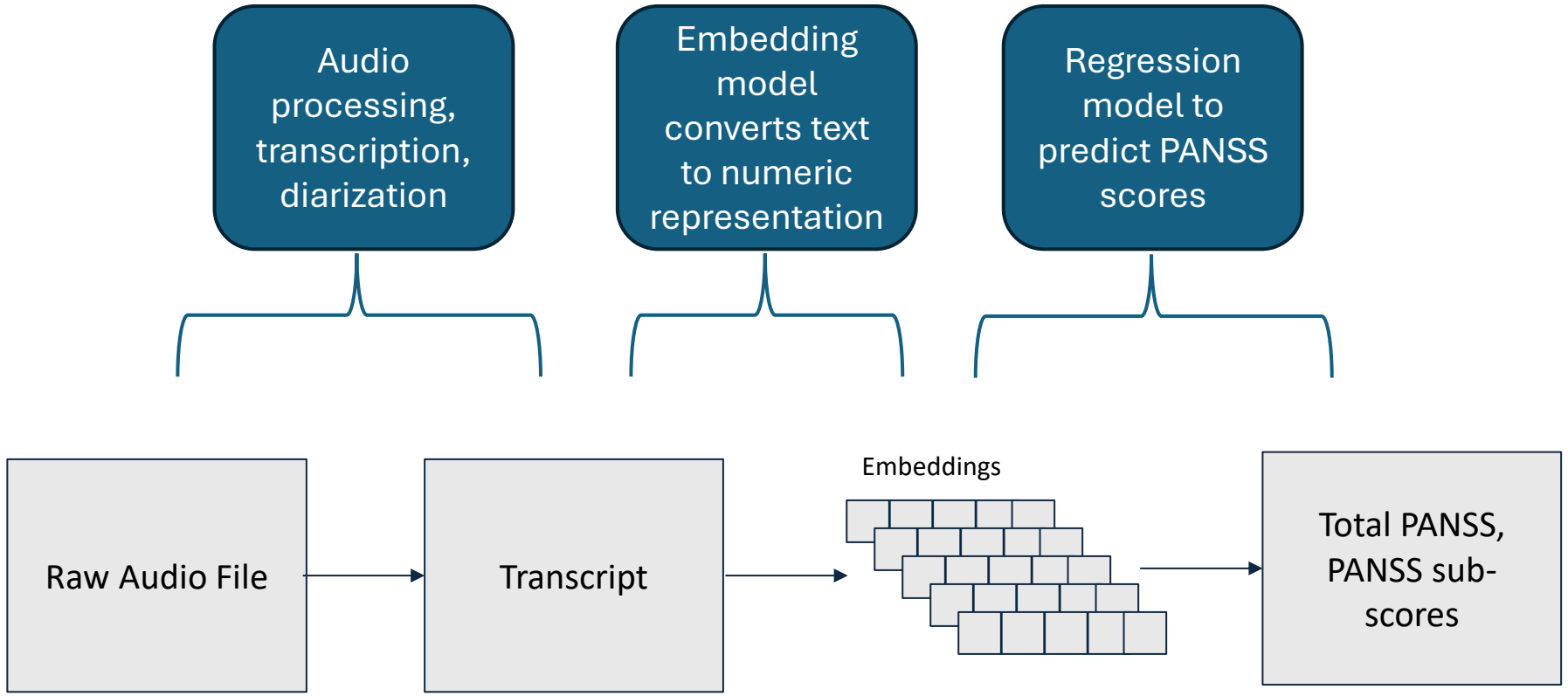
Results: LLM Predicts PANSS Sub-Scores

Predicting PANSS sub-scores via LLM with zero-shot prompting



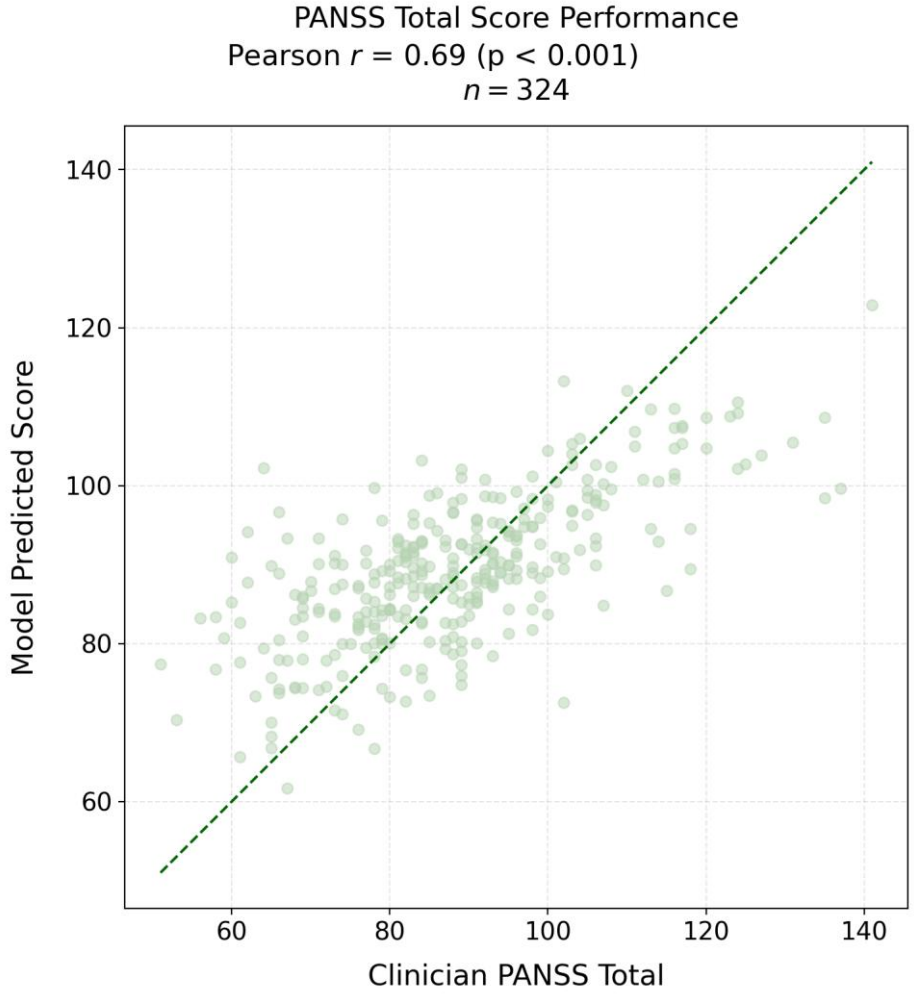
Performance is better for Positive and General sub-scores than for Negative – negative symptom scores may be more dependent on non-verbal information that isn't captured in transcript

Approach: Predict PANSS scores from embeddings



Results: Embedding-based prediction of Total PANSS Score

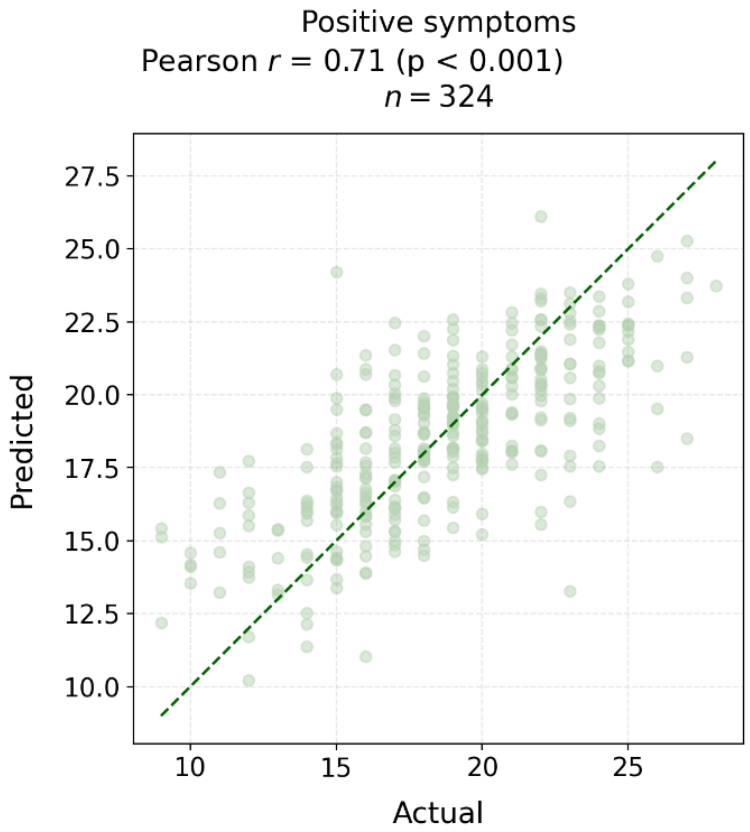
Predicting total PANSS score via regression on embedding features

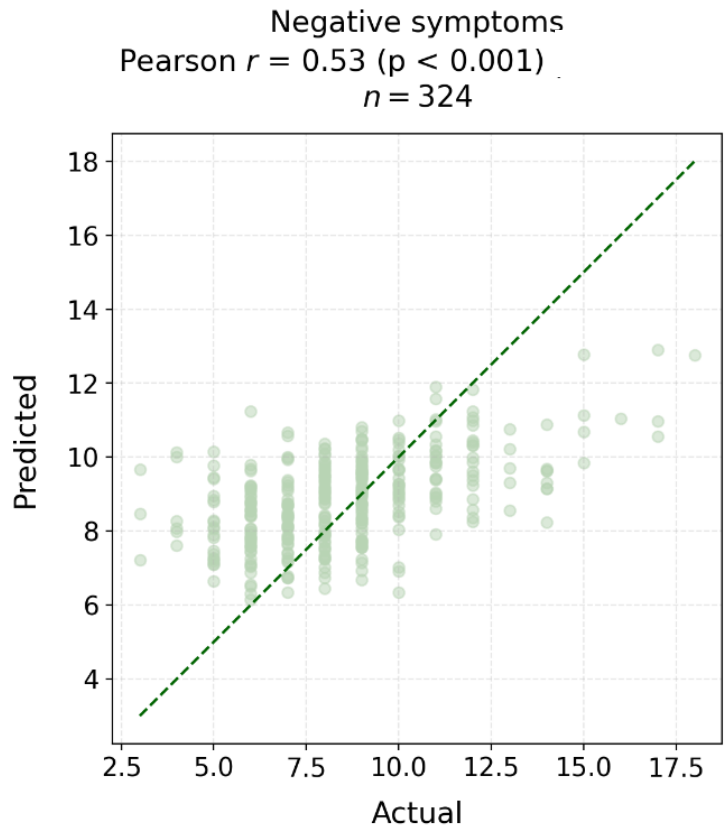


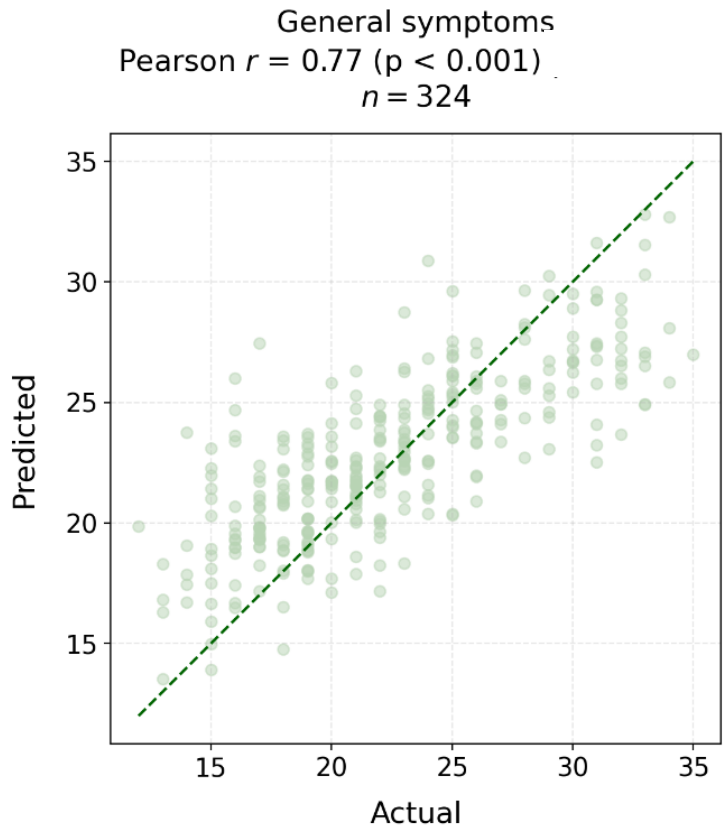
- Similar, but slightly improved overall performance relative to LLM model
- Distribution of PANSS data limits performance for more extreme PANSS scores

Results: Embedding-based prediction of PANSS Sub-Scores

Predicting PANSS sub-scores via regression on embedding features







Embedding-based regression models produce very similar overall performance across all three sub-scores to the LLM based models

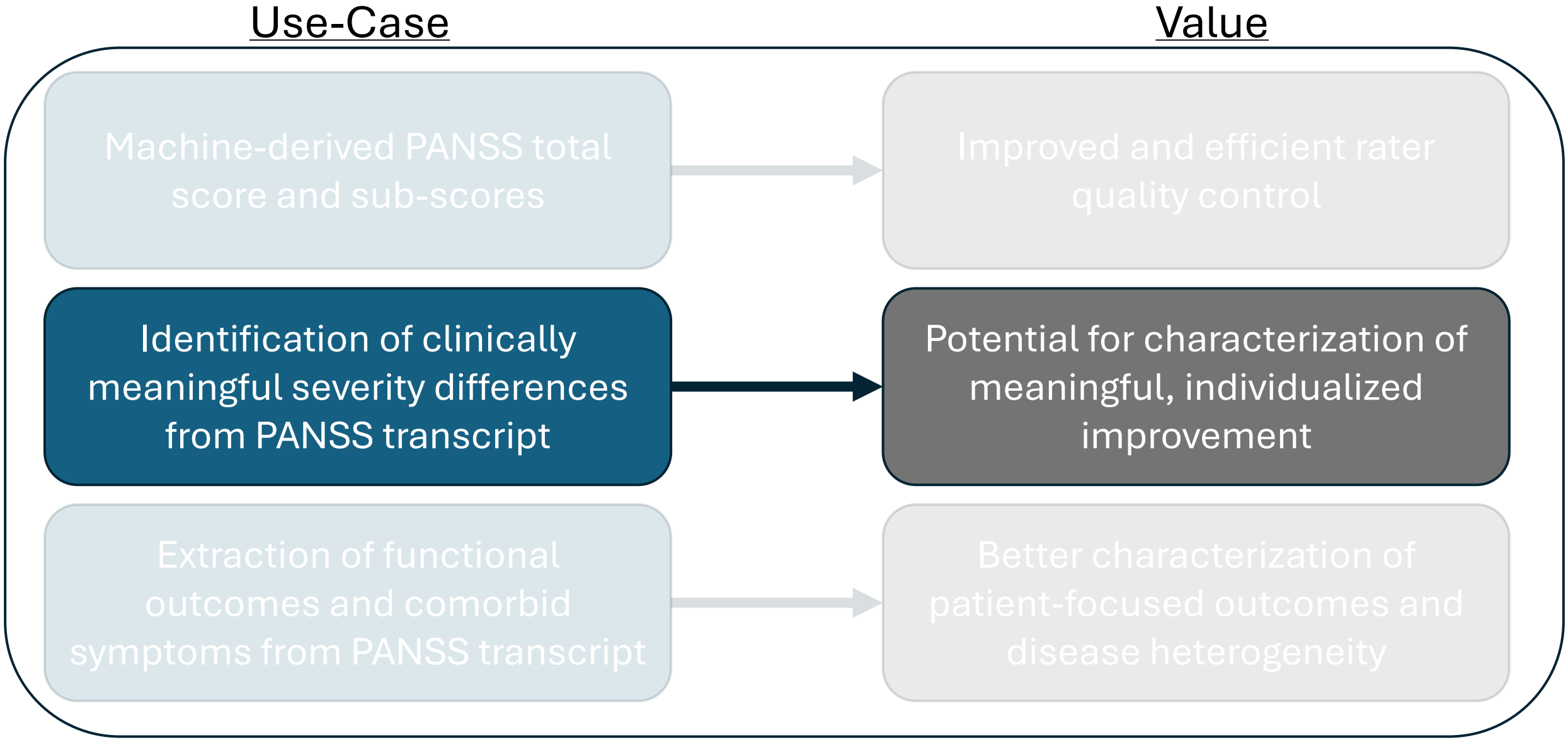
Limitations

- Not all information that informs scoring is captured in interview transcript (clinical impressions, body language, etc.) – could limit accuracy of machine scoring relative to other scales/assessment

Conclusions

- One-shot prompting LLM models perform very similarly to embedding-based regression models even without additional fine-tuning and/or training

Three potential use cases for LLM analysis of PANSS



Value: Focus on clinically meaningful difference prioritizes differentiated development

Assessing clinical meaningfulness can lead to more robust interpretation of treatment response

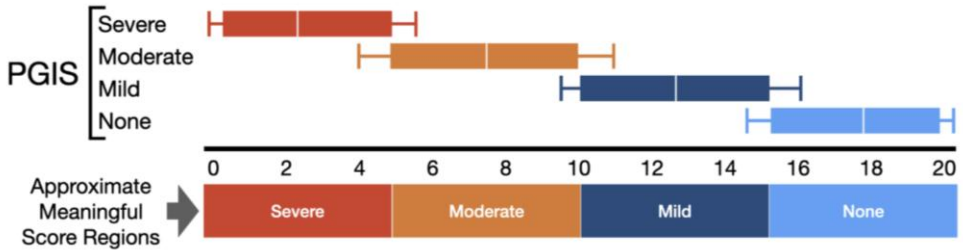
Patient-Focused Drug Development: Incorporating Clinical Outcome Assessments Into Endpoints For Regulatory Decision-Making

1. Interpreting in Terms of Meaningful Score Differences

This first approach identifies what size difference between any two COA scores would be viewed as meaningful for patients. This will be referred to as the *meaningful score difference* (*MSD*). Often, *MSD* is determined based on what patients would regard as a clinically meaningful within-patient change (i.e., improvement or deterioration from the patient's perspective), but other approaches might also be appropriate (e.g., those based on the patient's perception of the differences between hypothetical vignettes representing different degrees of symptom severity or functioning). Note that patients differ in their views of what might count as *MSD*, but for purposes of evaluating the results of clinical trials, a range of *MSD* should be selected that reflects most patients.

Regardless of the approach used to determine the *MSD*, the *MSD* can be used in at least two ways: (1) to evaluate the expected treatment effect for the average patient in some target population; or (2) to use as a threshold in descriptive analyses that identify individual patients who might have changed by a meaningful amount. Both of these applications will be discussed (see III.C) following a review of approaches for selecting a value or range of values for *MSD*.

Figure 1. Example of Approach for Interpreting COA Scores in Terms of Meaningful Score Regions Corresponding to Patient Global Impression of Severity (PGIS).



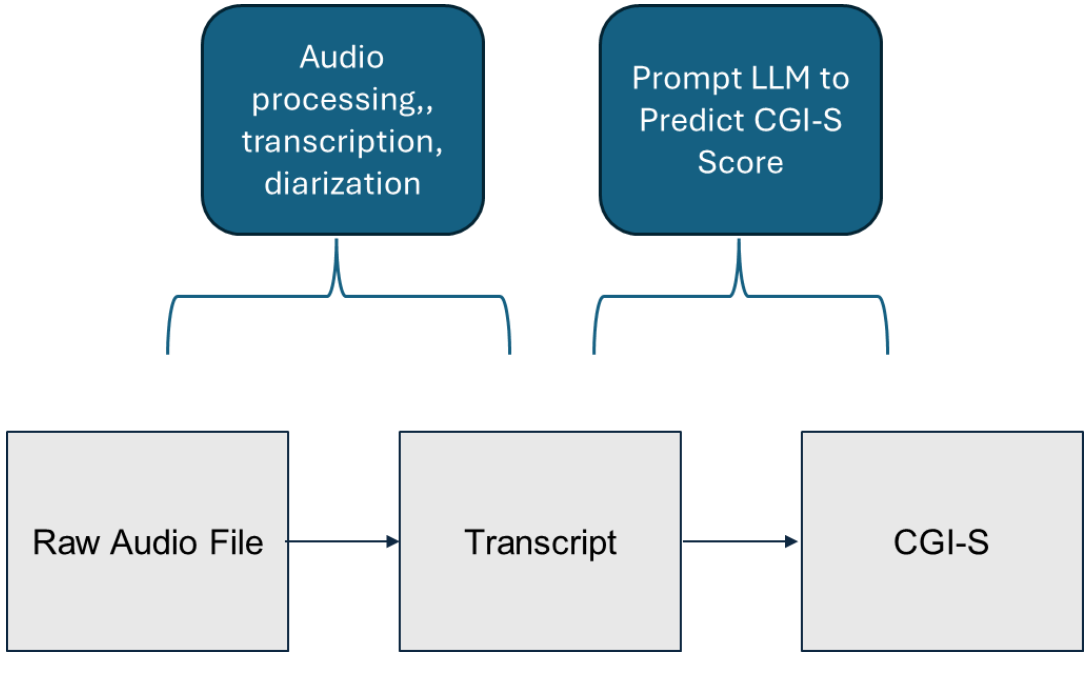
Value

- Average symptom improvement alone can obscure patterns of response
- Within-participant differences may provide personalized indicator of response

Approach: LLM predicts CGI-S or ΔCGI-S from transcripts

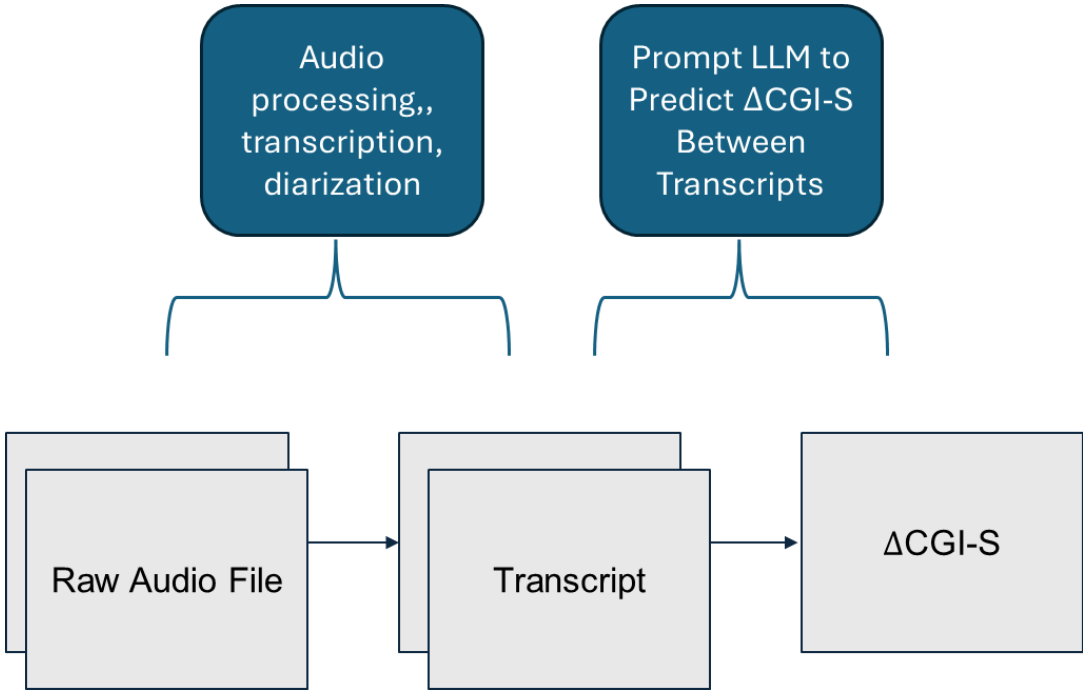
Approach 1

Predict CGI-S from PANSS transcript



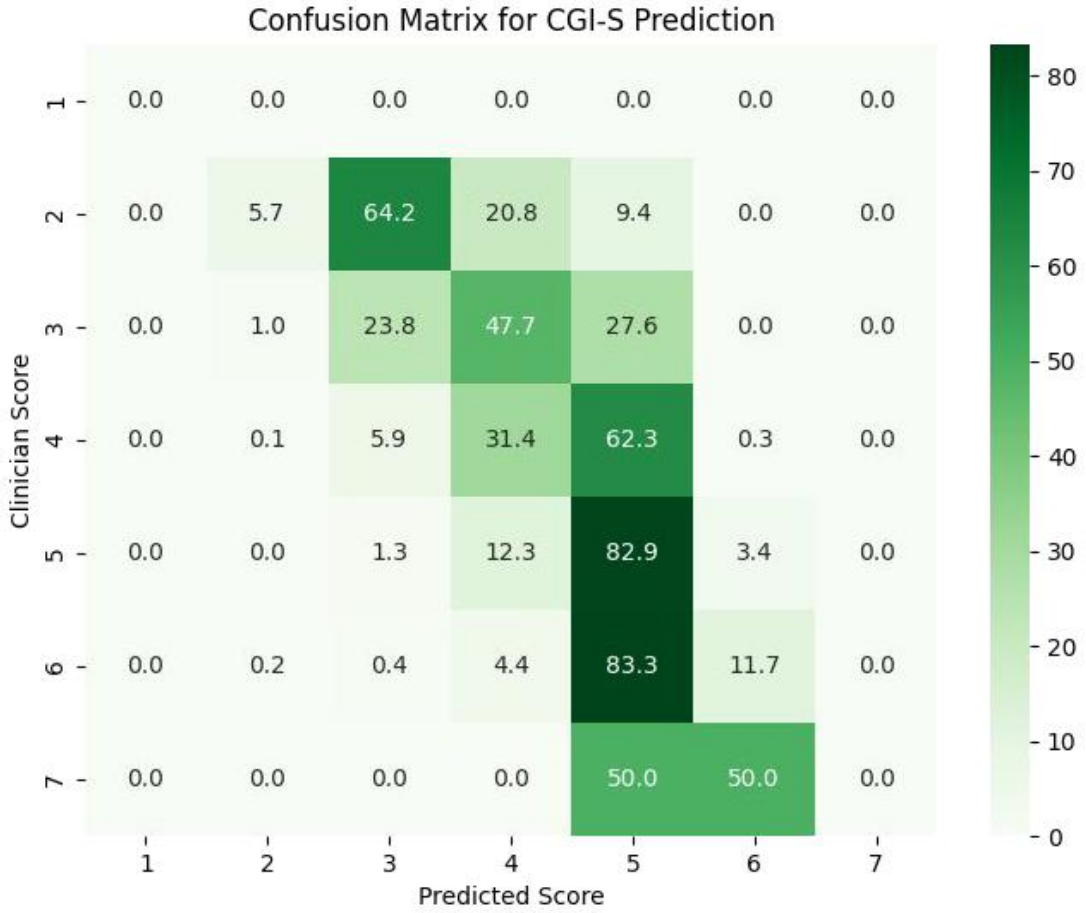
Approach 2

Predict difference in CGI-S between two PANSS transcripts



Results: LLM prediction of CGI-S

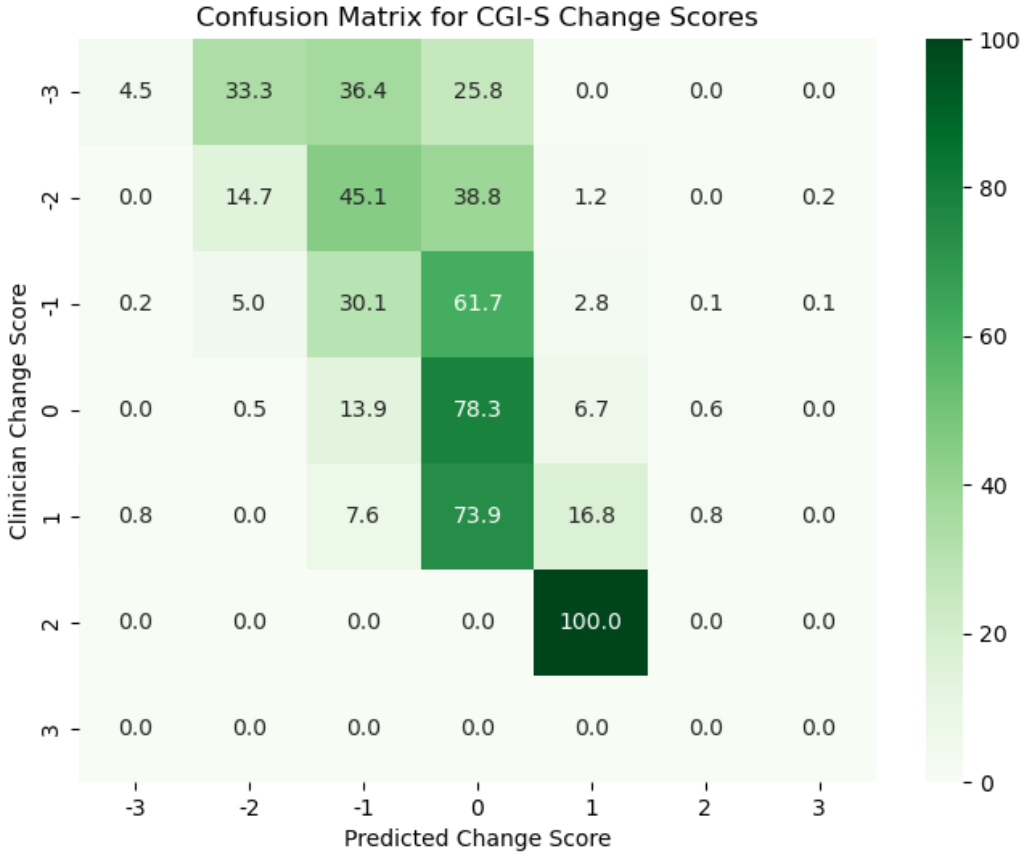
Predicting CGI-S score via LLM with zero-shot prompting



- Model directionally predicts severity
- Error is greater for more extreme values (bias towards the mean)
- May be difficult for LLM to meaningful distinguish severity levels without training

Results: LLM prediction of ΔCGI-S

Predicting ΔCGI-S score via LLM with zero-shot prompting

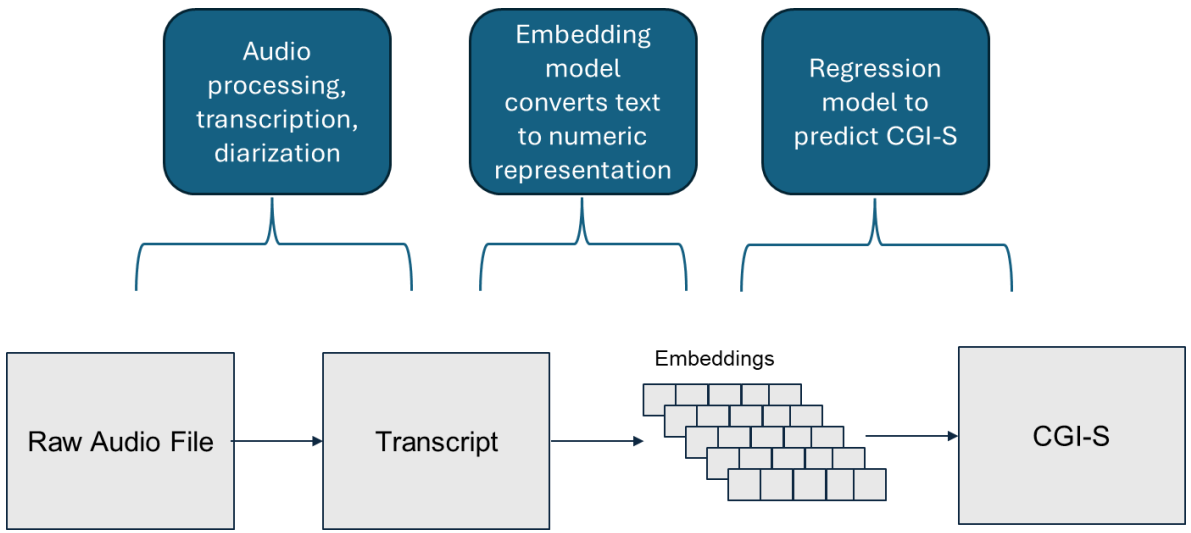


- Directionally predicts severity differences
- Tendency to predict values closer to 0

Approach: Predict CGI-S or ΔCGI-S from embeddings

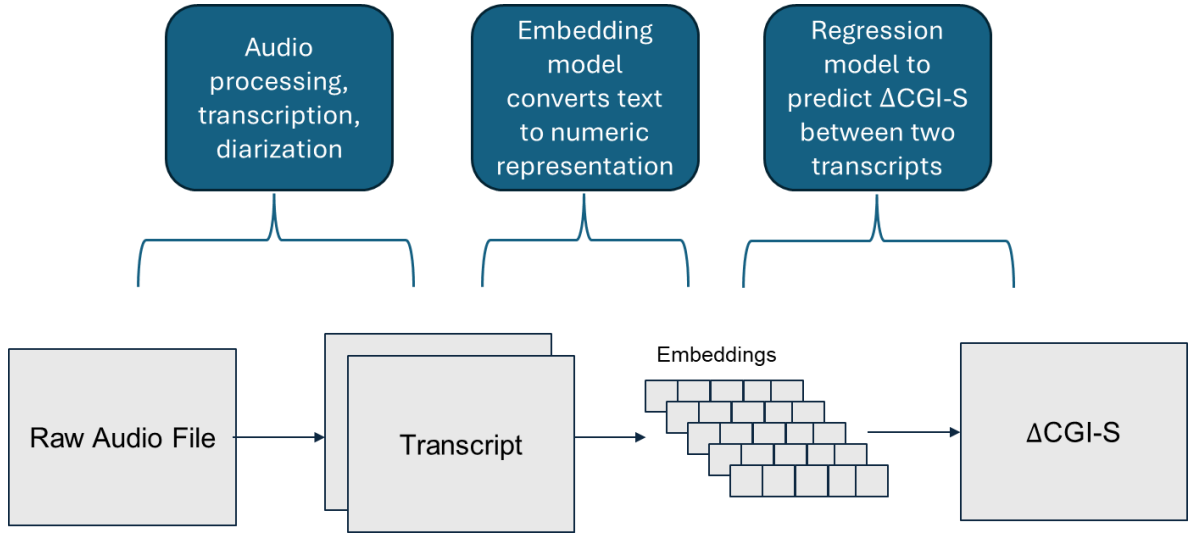
Approach 1

Predict CGI-S from PANSS transcript embeddings



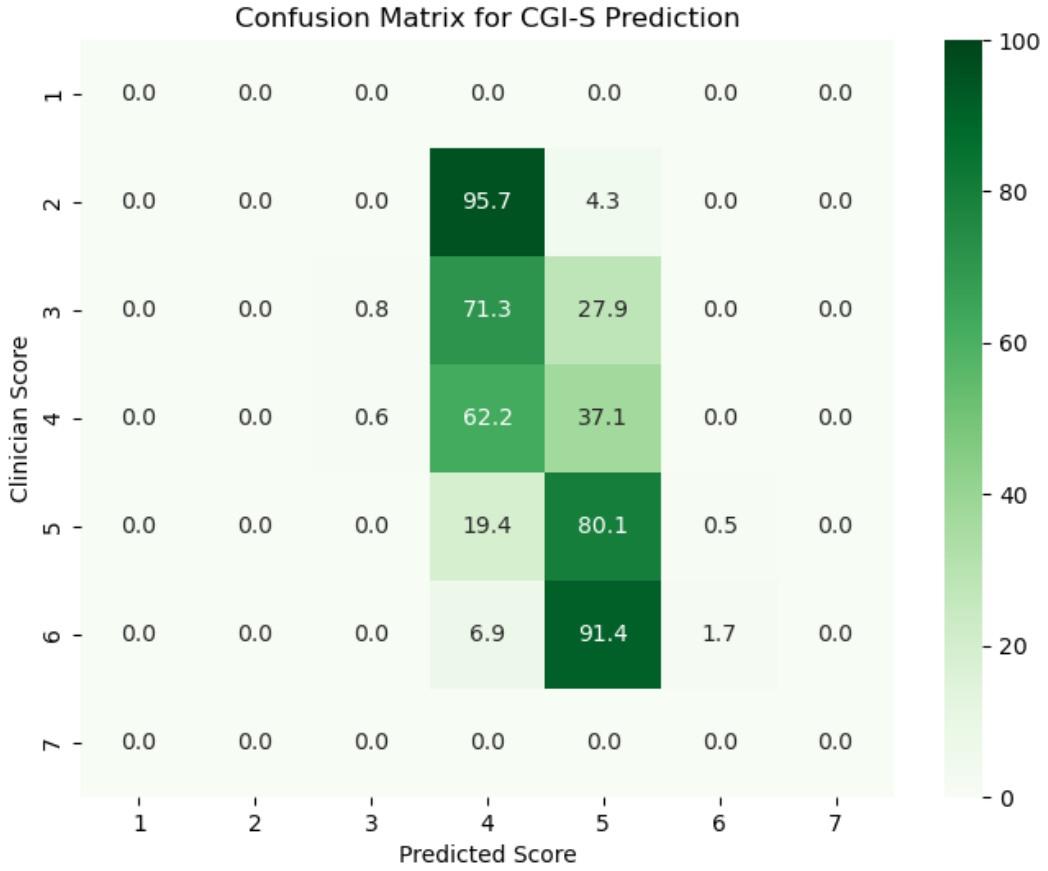
Approach 2

Predict difference in CGI-S between based on embeddings of two PANSS transcripts



Results: Embedding-based prediction of CGI-S

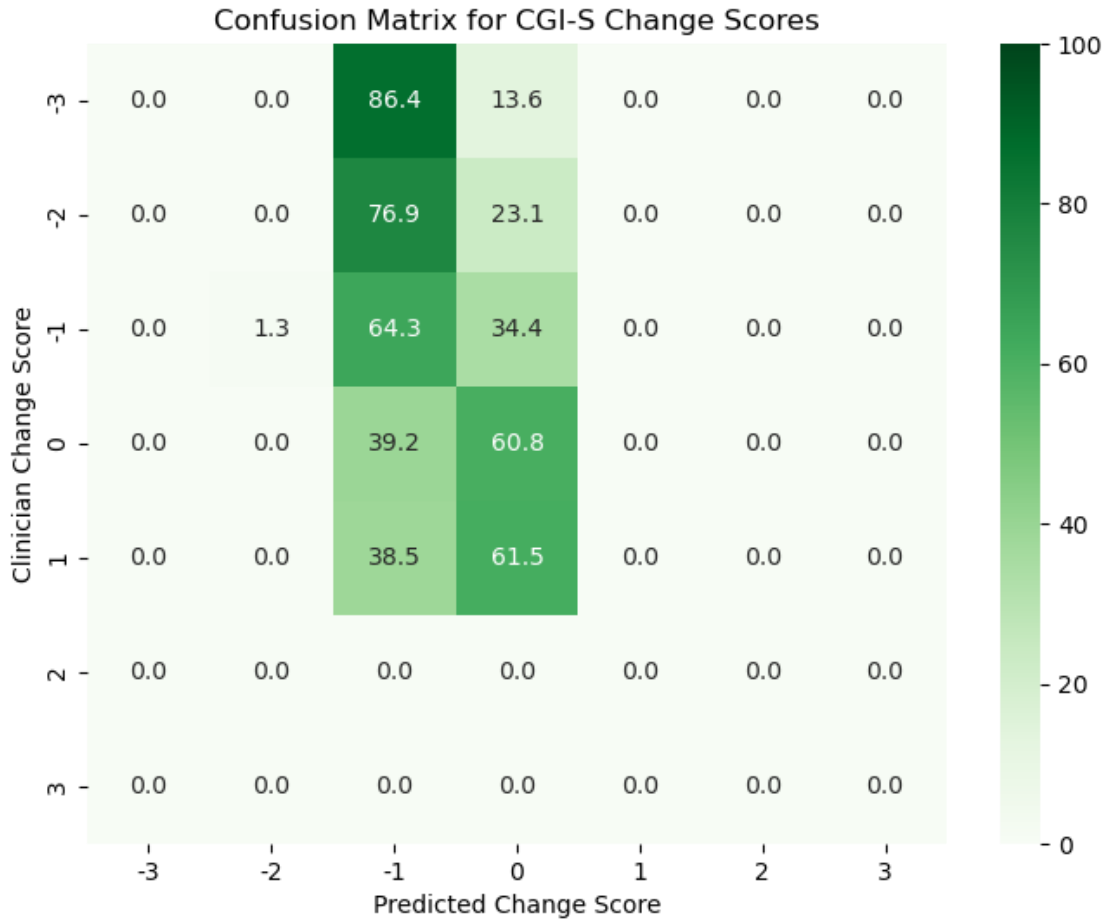
Predicting CGI-S score via regression on embedding features



- Poor performance
- Biased by unbalanced dataset in which majority of CGI-S ratings are either 4 or 5
- Embedding representation does not appear to help for this use-case

Results: Embedding-based prediction of ΔCGI-S

Predicting ΔCGI-S score via regression on embedding features



- Poor performance
- Unbalanced dataset (0 and -1 are most common ΔCGI-S values)

Limitations

- Comparatively narrow range of outcomes – within clinical trials, CGI-S is not uniformly distributed which can bias methods that require training
- Clinical severity is not the only factor that affects transcripts – comparison of two transcripts from one patient may be interpreted differently than comparison of two transcripts across patients

Conclusions

- LLMs can directionally identify clinically meaningful differences but quantitative accuracy is still unclear
- In this case, LLMs do improve upon more classical methods that require training (likely due to the unbalanced nature of the training data)

Use-Case #3

Use-Case	Value
Machine-derived PANSS total score and sub-scores	Improved and efficient rater quality control
Identification of clinically meaningful severity differences from PANSS transcript	Potential for characterization of meaningful, individualized improvement
Extraction of functional outcomes and comorbid symptoms from PANSS transcript	Better characterization of patient-focused outcomes and disease heterogeneity

Value: Linking symptoms to functioning could align clinical and quality-of-life outcomes

Functional scales are collected but usually not tied to symptom severity assessment

Patient-Focused Drug Development: Incorporating Clinical Outcome Assessments Into Endpoints For Regulatory Decision-Making

A. Factors Affecting the Interpretability of COA Scores

To determine whether a medical product has a positive, meaningful effect on how a patient feels or functions (i.e., a treatment benefit[23]), FDA recommends that sponsors measure how a patient's status on a COA-based endpoint corresponds to the way they feel and/or function in their daily life. For example, if a treatment is shown to reduce scores on a performance outcome measure

A disease might manifest in multiple ways, in which case it is important to consider how or whether a medical product affects different aspects of health. Some aspects of health might be

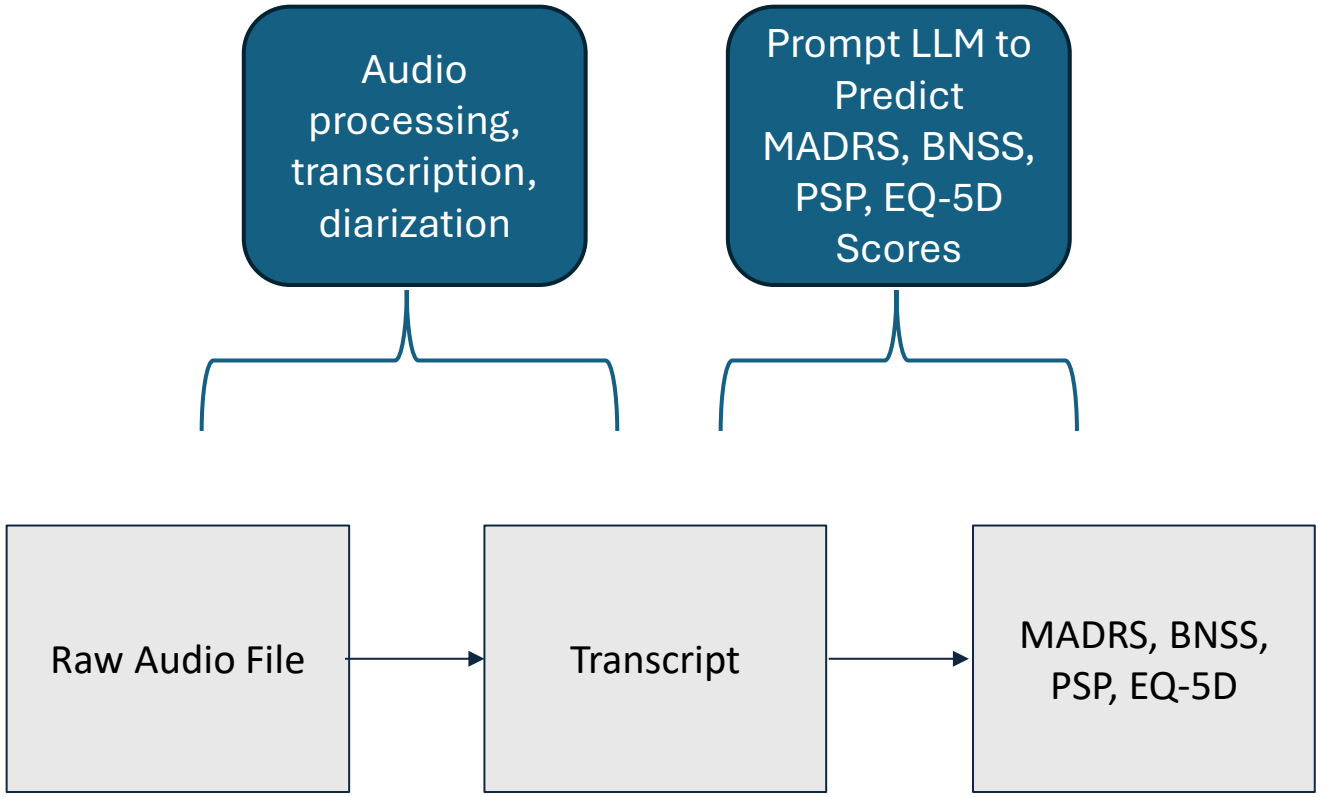
sample size to ensure sufficient statistical power. Finally, if patients differ from one another in their symptoms or functional impacts due to the disease, then the treatment effect estimated for any one endpoint will be diluted by the patients for whom the endpoint is not relevant (e.g., patients who never had a given symptom cannot improve with treatment). Consult the guidance

Schizophrenia symptoms manifest heterogeneously with varying impact on functional outcomes

Value

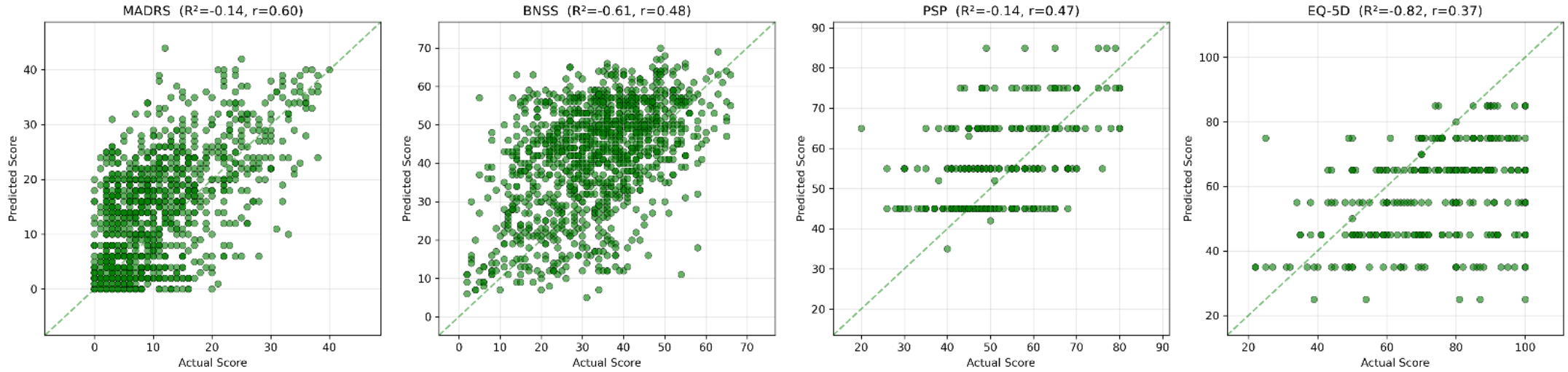
- Potential insights into symptom contribution to disease and functional burden
- Could facilitate more precise understanding and treatment of heterogeneous diagnoses

Approach: LLM predicts functional and comorbid symptom scores from full transcripts



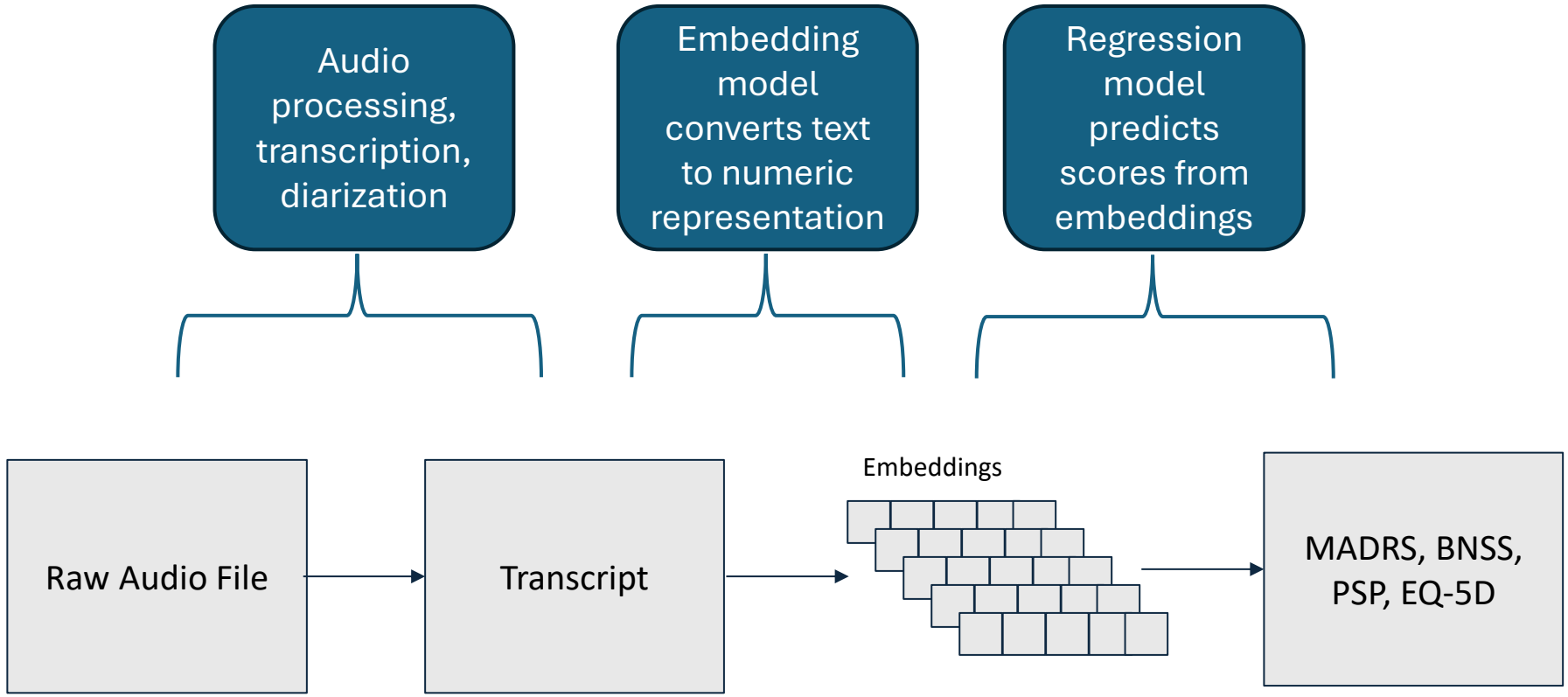
Results: LLM predicts MADRS/BNSS/PSP/EQ-5D

Predicting MADRS, BNSS, PSP, EQ-5D scores via LLM with zero-shot prompting



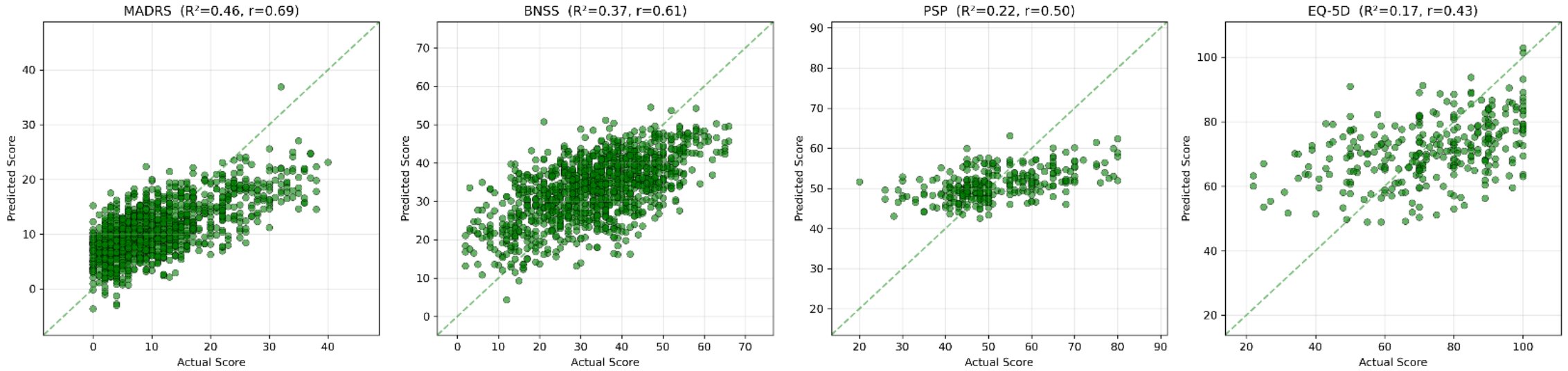
- Poor performance on functional measures (PSP, EQ-5D)
- Better, but relatively noisy performance on MADRS, BNSS

Approach: Predict functional and comorbid symptom scores from embedded representation



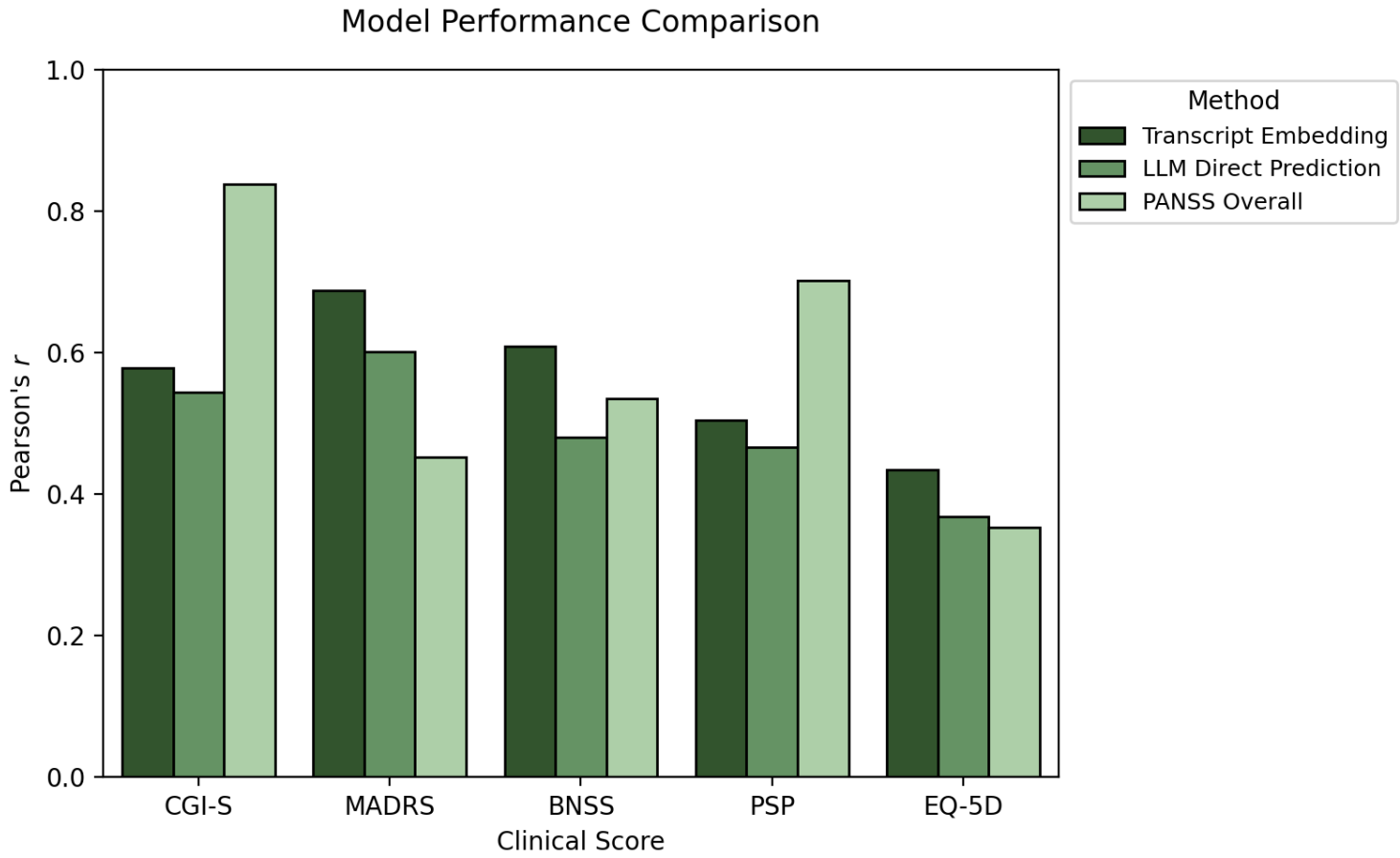
Results: Embedding model predicts MADRS/BNSS/PSP/EQ-5D

Predicting MADRS, BNSS, PSP, EQ-5D scores via regression on embedding features



- Predictions improved, but still poor for PSP/EQ-5D
- Same bias towards intermediate values observed with PANSS prediction model

Results: Comparison of LLM and Embedding Models



- In this case, embedding models slightly outperform LLMs
- CGI-S & PSP are actually better predicted by direct regression against Total PANSS score – indicates that this modeling is not particularly informative for those scales

Limitations

- Recorded transcripts are from clinical outcome assessments – not intended to interrogate functioning, which could limit availability of meaningful information
- Heterogeneity of symptoms and questions across full transcript may limit ability to isolate specific symptoms or functional impairment

Conclusions

- Some promise for extracting comorbid or overlapping symptom severity from PANSS transcript
- A different modeling approach would likely be required to required to extract any functional information from PANSS transcripts

Overall Conclusions

- In addition to automated scoring, LLMs could have broader utility with respect to clinical outcomes assessments and assist with our overall goal of more precise drug development
- **Transcript alone is currently insufficient to quantitatively predict outcomes**
- Approaches are not limited to PANSS but may be applicable to other key assessments
- LLMs tend to perform better when the analyzing text data that is more specifically aligned with the prediction to be made (e.g. LLMs analyzing PANSS interviews predict PANSS better than MADRS; Positive/General sub-scores predicted better than Negative sub-score)
- Extracting tangential information (e.g., MADRS, PSP) may requires a more nuanced approach with more domain knowledge and/or training
- **Still significant unknowns and limitations (due to both nature of LLMs and limitations of collected data) that need to be explored retrospectively in available data prior to any prospective use**

Potential Future Directions

- More robust prompt engineering and/or fine-tuning, e.g., to extract specific symptoms and potential link to functioning
- Incorporation of additional audio-based features to potentially capture more information not available in the transcript
- Exploration with other assessments to isolate methodological limitations from data limitations