Using Large Language Models for Endpoint Oversight

Adam Kolar[1,3], Miguel Amável Pinheiro[1], Alexander Deschamps[1], Todd M Solomon[1], Matus Hajduk[1], Martin Majernik[1], Daniel R Karlin[1,2]

[1]Mind Medicine, [2]Department of Psychiatry, Tufts University School of Medicine, [3]Department of Computer Science, Faculty of Informatics, Masaryk University, Brno, Czech Republic

Key Words: Clinician reported outcomes, large language models, artificial intelligence, central raters, psychiatry

Methodological Issue Being Addressed

In psychiatric drug development, clinician reported outcomes (ClinROs) obtained from participant interviews are considered gold standard primary efficacy endpoints in indications such as depression, schizophrenia and anxiety. However, ClinROs have significant limitations, including clinician bias, inter- and intra-rater variability, poor sensitivity and the inherent subjectivity involved with psychological evaluation (1, 2, 3). To help mitigate these issues, sponsors have deployed methodologies such as rater training and certification, use of central raters, blinded data analytics and third-party review of endpoints to help reduce the risks associated with ClinRO assessments. However, these methodologies add operational complexity, burden and expense to studies and have not been proven to be entirely effective in protecting the reliability and validity of these endpoints (4).

Introduction

Large Language Models (LLMs) have the potential to provide oversight of ClinRO endpoints due to their ability to process text and, through extensive training, learn to accurately infer and assign a score to language-based sentiment. We created Hammy- a system of LLMs which transcribe ClinRO interviews, parse out individual ClinRO items, and provide associated scoring. Here we outline the methodology used to develop and train Hammy and discuss using it to perform a post-hoc data quality check on Hamilton Anxiety Rating Scales (HAM-A) from a recent Phase 2b clinical trial. Finally, we discuss the implications of deploying this technology for data monitoring in ongoing clinical trials.

Hammy Pipeline



Audio recording	Transcription	Parse HAM-A items	Score items
HAM-A interviews between participants and central raters are recorded	Whisper, an open-source transcription model, is prompt engineered to more accurately transcribe ClinRO interviews and is deployed on the recordings	Find the beginning of each of the 14 HAM-A items using regex or fine-tuned model from LLama 3.1 7B	Classify the severity of each item using a model fine-tuned from LLama 3.1 7B

Fine-tuned from 1500 sessions from Phase 2b – 21,000 symptom ratings

Method

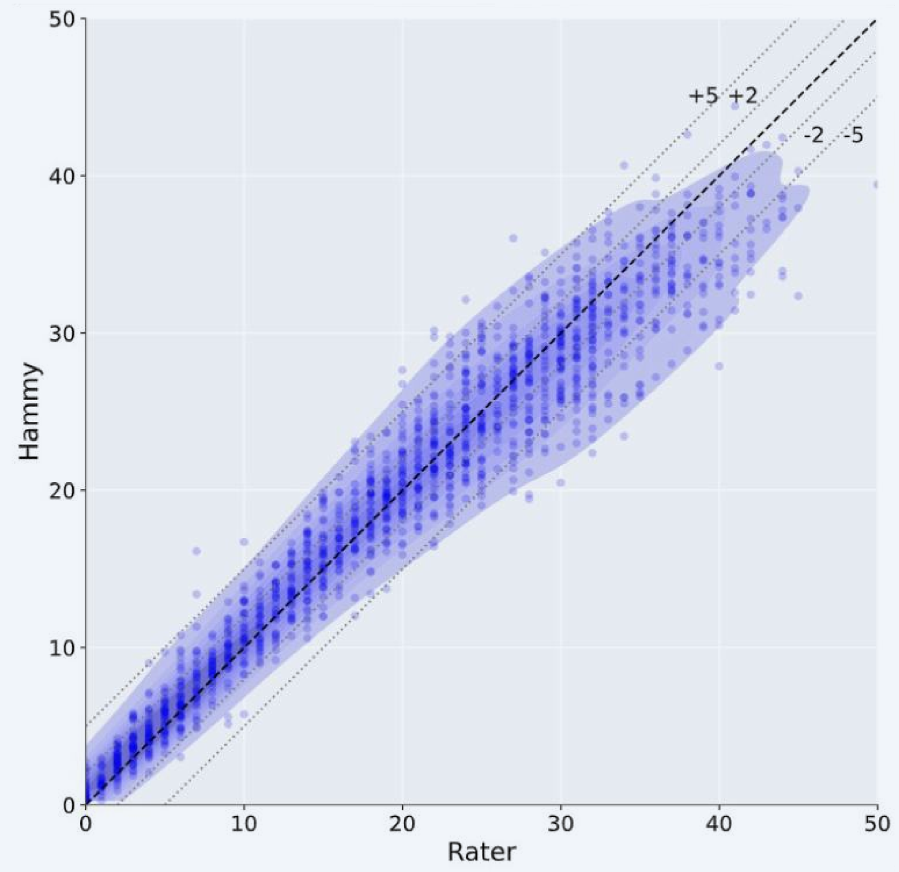
Hammy consists of multiple concurrent models that ingest audio recordings of ClinRO interviews and produce associated scores for each item of that interview. The first step utilizes Whisper, an open-source transcription model which we prompt engineered to more accurately transcribe ClinRO interviews (5). After audio recordings are transcribed by the model, a second model parses the transcripts into item-level segments. A third model analyzes this version of the transcript and produces a score for each item. These models are based on Llama 3.1 and were trained on data from a recent Phase 2 clinical trial run by Mind Medicine, Inc. (MindMed), where over 1500 HAM-A interviews were audio recorded (6). Once trained, a cross-validation technique was used to create different versions of Hammy in order to appropriately test performance on the Phase 2 dataset as well as approximate scoring confidence in other datasets. As a measure of performance, Hammy scored three training interviews used to certify human raters. Finally, we re-analyzed the results of the Phase 2 study using Hammy's scoring in place of the original central rater scoring.

Results

On training interviews, Hammy matched the "answer key" scoring for all 14 items (100%) on the first 2 recordings and for 12 of the 14 items (85.7%) for the third recording. When deployed on the Phase 2b data, Hammy's scores differed on average 1.57 (+- 1.39) points from the central raters' scores, with Pearson's r = 0.98, indicating a very strong correlation between the two sets of scores. Hammy's scoring reaffirmed the topline results of the Phase 2b trial. Notably, Hammy's scoring would have resulted in 21 participants who were originally excluded from the study based on HAM-A scoring below to be included, and 7 participants who were originally included to be excluded. However, exclusion of these 7 participants would not have significantly changed results.

Testing Hammy On:

Phase 2 Data


50
40
30
20
10
0
Hammy
+5 +2
-2 -5
0
10
20
30
40
50
Rater

A graph with each dot representing a single HAM-A interview from the Phase 2b trial, plotted based on Hammy's total score (y-axis) and the central rater's total score (x-axis). The bold dashed line represents perfect agreement between the scores, with the grey dashed lines representing +/- 2 or +/- 5 points between the total scores.

Training Interviews

Training Interview 1

Answer Key Score: 36-38
Hammy Score: 37

Training Interview 2

Answer Key Score: 8-11
Hammy Score: 8

Training Interview 3

Answer Key Score: 19-22
Hammy Score: 19

Hammy achieved passing scores on all 3 training interviews as compared to an answer key created by rating experts.

Discussion

We evaluated Hammy's ability to be used as a tool for ClinRO oversight by deploying Hammy on three central rater training interviews and on real data from a Phase 2b clinical trial. Hammy's scores on the training interviews would have been sufficient to pass training, suggesting scoring capabilities equivalent with trained central raters. When deployed on data from the Phase 2b trial, Hammy's scores were strongly correlated with the central rater's scores, with a small average absolute difference well within normal levels of inter-rater variability (7). Given the inherent variation in psychological assessment, some degree of difference between a single rater's scores and those produced by a group evaluating the same interviews is not surprising. At scale, this difference resulted in some changes to the specific participants who would have been included and excluded in the trial. This does not diminish our confidence in Hammy's feasibility as an oversight tool, but may be important to consider if LLMs take on a more expanded role in clinical trials.

Conclusion

The present work indicates that LLMs have the potential to be deployed as a consistent and relatively inexpensive method of providing oversight of clinical outcomes ratings in ongoing clinical trials. As LLMs deploy a uniform scoring methodology and can rapidly score every interview conducted, they can provide a level of data quality oversight that has previously been prohibitively expensive or logistically impossible to obtain via human raters. Hammy is currently being deployed in ongoing Phase 3 trials to augment central rater oversight, and to date has analyzed over 1,300 HAM-A interviews. Immediate future directions for this work include expanding the variety of scales Hammy can score, improving performance through retraining on bigger datasets, and incorporating explainability features to provide further granularity and insight into rater performance.

Applying Hammy to Ongoing Trials

HAM-A Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Total Score
Rater	3	4	0	3	3	0	0	0	0	0	2	4	3	3	25
		●●			●							●●●●		●	
Hammy	3	2	0	3	2	0	0	0	0	0	2	0	3	2	17


15
10
5
0
-5
Absolute Difference
Rater 1
Rater 2
Rater 3
Rater 4
Rater 5
Rater 6
Rater 7

Ongoing deployment of Hammy involves comparing scores from Hammy and raters on both on a per interview basis (above) and at an aggregated level across raters (below), with the latter comparing each rater's average absolute differences between their scores and Hammy's.

References

1. Heneghan, C., Goldacre, B., & Mahtani, K. R. (2017). Why clinical trial outcomes fail to translate into benefits for patients. Trials, 18(1), 122. https://doi.org/10.1186/s13063-017-1870-2
2. Dal-Ré, R., Bobes, J., & Cuijpers, P. (2017). Why prudence is needed when interpreting articles reporting clinical trial results in mental health. Trials, 18(1), 143. https://doi.org/10.1186/s13063-017-1899-2
3. McNamara, C., Engelhardt, N., Potter, W., Yavorsky, C., Masotti, M., & Di Clemente, G. (2019). Risk-Based Data Monitoring: Quality Control in Central Nervous System (CNS) Clinical Trials. Therapeutic Innovation & Regulatory Science, 53(2), 176–182. https://doi.org/10.1177/2168479018774325
4. Barnes, B., Stansbury, N., Brown, D., Garson, L., Gerard, G., Piccoli, N., Jendrasek, D., May, N., Castillo, V., Adelfio, A., Ramirez, N., McSweeney, A., Berlien, R., & Butler, P. J. (2021). Risk-Based Monitoring in Clinical Trials: Past, Present, and Future. Therapeutic Innovation & Regulatory Science, 55(4), 899–906. https://doi.org/10.1007/s43441-021-00295-8
5. Radford, A., Kim, J. W., Xu, T., Brockman, G., Mcleavey, C., & Sutskever, I. (2023). Robust Speech Recognition via Large-Scale Weak Supervision. Proceedings of the 40th International Conference on Machine Learning, 28492–28518. https://proceedings.mlr.press/v202/radford23a.html
6. Robison, R., Barrow, R., Conant, C., Foster, E., Freedman, J. M., Jacobsen, P. L., Jemison, J., Karas, S. M., Karlin, D. R., Solomon, T. M., Halperin Wernli, M., & Fava, M. (2025). Single Treatment With MM120 (Lysergide) in Generalized Anxiety Disorder: A Randomized Clinical Trial. JAMA. https://doi.org/10.1001/jama.2025.13481
7. Trajković, G., Starčević, V., Latas, M., Leštarević, M., Ille, T., Bukumirić, Z., & Marinković, J. (2011). Reliability of the Hamilton Rating Scale for Depression: A meta-analysis over a period of 49 years. Psychiatry Research, 189(1), 1–9. https://doi.org/10.1016/j.psychres.2010.12.007