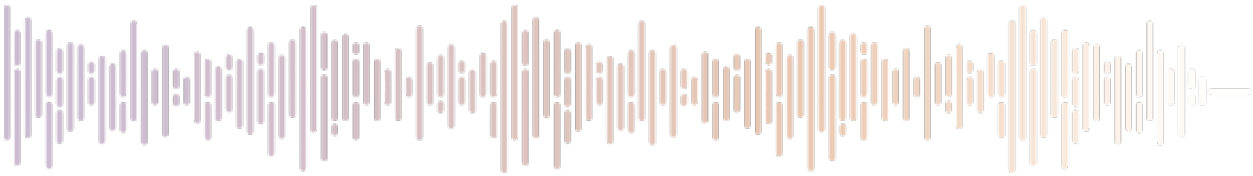
LYSSN

THE NEW STANDARD OF PRACTICE IN
HEALTH AND HUMAN SERVICES.™



Lyssn's Annual AI Bias Report

September 2023

Michael J Tanana PhD, Jordan Pruett PhD, Brian Pace PhD

EXECUTIVE SUMMARY

Background: Lyssn's motivational interviewing (MI) training platform is powered by an AI language model, trained on more than 20,000 human-evaluated sessions from counseling, health coaching, crisis counseling, caseworkers, and other behavioral health and human services settings. This training data is both extensive and has been evaluated by licensed professionals trained in MI. At the same time, the scale and complexity of modern AI models leave the door open to potential biases, and the risk that AI tools learn to reproduce harmful social biases is a concern for the designers and users of such tools.

To assess and address this risk, Lyssn undertook a study that tested for the presence of measurement bias in the MI fidelity algorithms of our platform. In collaboration with clinical partners, the present report assesses 27 therapy sessions in which the provider identifies as a racial or ethnic minority (REM). These sessions were manually transcribed and then evaluated for MI fidelity by Lyssn's trained human coding team. We examined the accuracy of our production AI models for MI fidelity in this subset of REM providers, compared to our standard test set of sessions, featuring a general population of providers.

The present report assesses the measurement bias of our AI models. Each of the two datasets – the general population set and the REM subgroup – were processed separately by Lyssn's MI fidelity algorithms. The model's ratings on each dataset were scored based on how accurately they reflected expert, human fidelity ratings on those same datasets, where accuracy was measured by standard metrics for assessing machine learning classifiers.

Results: The study found no bias between the model's overall performance on the two datasets. There was no significant difference in AI-based predictive fidelity over all MI codes comparing the general provider group and the REM provider subgroup. However, results did show that the model made slightly less accurate predictions on the REM subgroup for two MI skills: Advice-Giving and Giving Information. These skills are relatively rare overall, so they do not appreciably impact the overall model performance. And a simple way to mitigate the performance discrepancy on these codes is to increase the diversity and frequency of these codes in the training data.

This study will be repeated annually, as a regular part of continual quality improvement at Lyssn. The results of this study and its later iterations will inform the composition of future training datasets. Continued monitoring of model performance is essential for ensuring that models generate fair feedback, as well as for maintaining trust with the providers that use the Lyssn platform.

Monitoring for Bias in AI Systems

A report on the potential psychometrics bias of the Lyssn platform

Contents

Introduction

As large language models (LLMs) have become more accessible, AI-powered natural language processing tools have transformed many industries, including behavioral health and human services. However, the scale and complexity of these models makes their behavior difficult to assess, even for experts (Yang et al, 2020). Since LLMs are trained in part on large corpora of text taken from uncurated sources such as online forums (e.g., Reddit, Twitter), they may learn to reproduce latent biases in the training data not intended by their designers (Bender et al, 2021; Dixon et al, 2018). Understandably, the risk that language models might reproduce harmful social biases around race, gender, and other cultural and social identities is unacceptable, could cause harm to vulnerable individuals seeking care, and ultimately may detract from the potential gains that AI technologies can bring.

The AI algorithms of the Lyssn platform are trained on more than 20,000[1] real, human-evaluated psychotherapy sessions from counseling, health coaching, crisis counseling, caseworkers, and other behavioral health and human services settings. All of our training data has been validated by licensed professionals with training in fidelity for evidence-based practices (e.g., treatment interventions, emotion, risk assessment). But even with this level of curation, real-world behavioral health and human services data are messy and complex. AI models might respond to subtle patterns in the data that may not be immediately apparent even to professionals, which could affect the quality of predictions (see Dixon et al, 2018 and Bolukbasi et al, 2016).

[1] 20,000 sessions includes sessions used to train all Lyssn models, including both speech and behavioral labels. Many of these models use transfer learning, so that information learned for one set of labels can assist other types of predictions.

Lyssn's mission is to use AI technologies to improve the quality of care and services clients and patients receive. At Lyssn, we take the risk of measurement bias related to provider identities seriously. We strive to continually monitor and improve our AI systems, in line with the most current research and industry best practices. As such, we recently evaluated our AI motivational interviewing assessment tool for the presence of bias related to the racial or ethnic identity of providers. This study is the inaugural version of what will be an annual evaluation of our training data and AI systems.

Although there are many different kinds of bias that might affect an AI system, the present study evaluated potential bias relating to the measurement bias of our AI assessment models (Crocker & Algina, 2002). Our aim was to test whether our models made more or less accurate assessments depending on the racial or ethnic identity of the provider being assessed. We consider a model's assessments to be more accurate if they more closely correspond to the human assessments made by our trained MI fidelity team.

In the following section, we describe the methodology of the study and define the metrics used to make our evaluations. Following that, we report the results of the study, including the model performance metrics on both test groups and an interrater baseline for comparison. Finally, in the discussion section we offer some interpretation of the results and thoughts about directions for future analyses. All metrics referenced in the paper and in any visualizations are available in the Appendix.

Methods

In collaboration with our clinical partners, we identified a set of anonymized behavioral health and human service sessions where the provider identifies as a racial or ethnic minority (REM) via customers that volunteered therapist demographic data. The REM sample included Latino, African American, and Asian American providers. To measure bias in the assessment portion of the Lyssn platform, we compared the performance of our algorithms on a general dataset of behavioral health and human service sessions to this subgroup of REM providers. This framework comes from the standard methodology for detecting psychometric bias in assessments (see Crocker and Algina, 2004; Henderson, Tanana, Bourgeois & Adams, 2015).

The first step to assess potential bias was to develop an expert, human evaluated dataset from REM provider sessions. These sessions were first transcribed by professional, medical transcriptionists and coded for motivational interviewing fidelity metrics by the Lyssn Clinical AI team. Fidelity was defined by the Motivational Interviewing Skills Code manual (see Miller et al, 2003), rating both global constructs of MI (e.g., empathy) and utterance-level codes (e.g. open questions). The team consisted of three human coders, two identified as female, one male, and two identified as REM. All coders were clinically

trained (social work, licensed counselor, psychologist) and were previously trained and experienced in MISC and MITI coding, achieved baseline reliability, and coded every utterance in each session. Lyssn coders worked only from transcripts; they had no access to session audio, meaning they had no explicit information about provider voice or accent. Coding disagreements and discussions continued throughout the coding process to maintain the highest human-to-human agreement possible across codes.

After all sessions were evaluated for MI fidelity by the Clinical AI team, we generated utterance-level MISC codes for these same sessions using Lyssn's AI production algorithms. We treated the human-assigned codes as the gold-standard: the more often that the model predicted the same code for an utterance as the one assigned by a human coder, the better its performance.

We measured model performance on each MISC code separately, using three standard metrics for evaluating the accuracy of a machine learning classifier: a) precision, b) recall, and c) the F1 score. For a given MISC code, precision measures how often the model was correct, out of all those cases where it assigned that code to an utterance. Recall (also called "sensitivity"), meanwhile, measures how many of the possible true cases of that code were correctly identified by the model. A model has high precision if it has few false positives, whereas a model has high recall if it has few false negatives. The F1 score combines precision and recall in a single measure and is often used as an overall measure of the model's performance. Each measure is on a 0-1 scale, with higher scores indicating better performance.

In the results section, we report the precision, recall, and F1 score for each MISC code for the general population dataset and the REM subgroup. For F1 scores, we further report 95% bootstrapped confidence intervals for each code and test group.

Finally, for F1 scores, we also report the interrater reliability of the F1 score from all our human-coded MISC data, as a baseline comparison. Interrater reliability captures how often the individual members of Lyssn's Clinical AI team agree on MI fidelity metrics. Since the AI models are trained on human-assigned codes, a lower interrater F1 score may reflect that a code is more intrinsically uncertain or subject to interpretation; even two highly trained professionals will not always agree on the appropriate code for a given utterance due to inherent ambiguity in transcripts of real-world sessions.

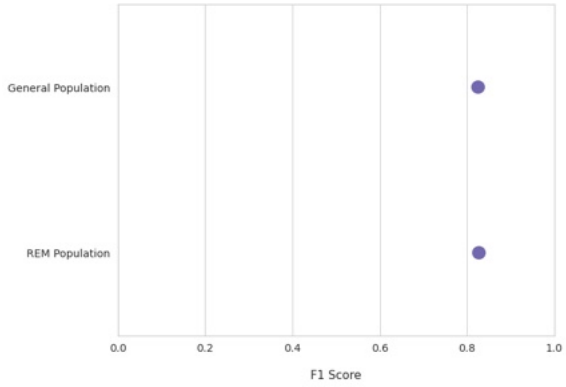
Results

The weighted average F1 scores over all codes for the general population and select REM provider subgroup are .83 and .83, respectively (see **Figure 1**). The 95% confidence intervals for these scores are overlapping, suggesting that there is no substantial difference between the two. See **Table 1** and **Table 2**

in the Appendix for precision, recall, and F1 scores for a full breakdown of all scores by code and test group.

Since the weighted average F1 is derived from the F1 score for each individual code, we also report individual F1 scores. For the majority of the tested codes, there was not a significant difference between the model's F1 score on the general dataset and the model's F1 score on the REM subgroup dataset. Results showed comparable predictions for the two groups for the following MI fidelity codes: affirmation, facilitation, closed question, open question, complex reflection, and simple reflection. For all of these codes, the F1 scores for classification on the racial and ethnic minority group fall well within the 95% confidence interval of the F1 scores on the general population.

Figure 1. **F1 Score by Test Set**[2]

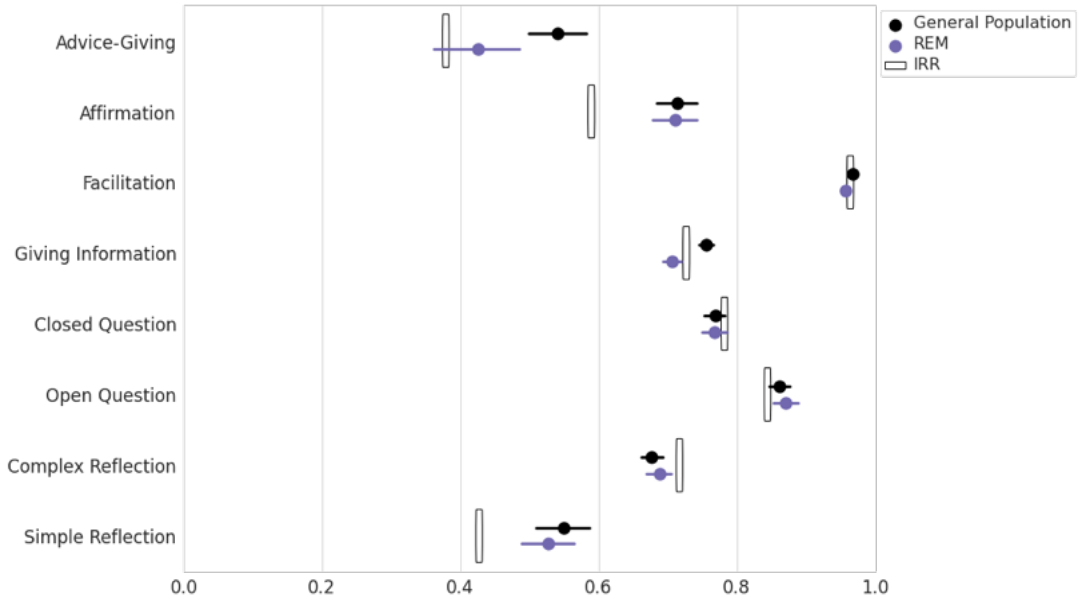


Two codes, however, showed some difference: "advice-giving" and "giving information." The 95% confidence intervals for the subpopulation F1 scores for these codes do not overlap. In both cases, the F1 score on the general population test group is higher: .54 vs .43 for advice-giving, and .76 vs .71 for giving information. F1 scores for all codes in both test sets are reported in **Table 1** and **Table 2** in the Appendix.

As a point of comparison, we also report interrater reliability scores for all codes. **Figure 2** below shows F1 scores for codes across both subgroups paired with their associated interrater reliability scores. In the case of advice-giving, the 95% confidence interval overlaps with the interrater score. However, the same is not true of giving information.

[2] Confidence intervals are not visible at this scale, but are available in **Table 1** and **Table 2** in the Appendix.

Figure 2. **F1 Scores by MI Code and Test Group**



It is worth noting that advice-giving is a reasonably rare code. Human coders assigned it to less than 1% of utterances in the select REM provider subgroup test set and to only about 1.4% of utterances in the general test set. Because of this, the lower score for this code for the REM provider subgroup does not substantially affect the overall, weighted average F1 score for all codes.

Discussion

Any psychometric tool, whether completed by a human or a machine, should be assessed for differential performance with different subpopulations (Crocker and Algina, 2002). In this report, we have analyzed the production Lyssn MISC identification algorithms for differential item functioning with our general test set compared to a test set composed of racial-ethnic minority providers. The goals of this report were to 1) identify any areas for improvement in our algorithms and their cross-cultural performance, and 2) transparently communicate detailed psychometric information to our customers and end-users. Overall, we found that the Lyssn MISC algorithm can predict Motivational Interviewing Skills Codes similar to an expert human rater. In addition, these models have statistically similar performance for both a general population of providers as well as an REM subpopulation.

The two areas where the algorithm could most improve for REM providers are the prediction of advice-giving and giving information. Both of these codes are typically considered undesirable in Motivational Interviewing practice. To get a better sense of how these errors affect the model's overall ratings, we did a closer analysis of what predictions the model is making when it gets these codes wrong. We found that in the cases when our model is incorrectly identifying advice-giving, it is most often incorrect because it is mistaking it for giving information (see appendix Figures 3 and 4). From an end-user perspective, this kind of mistake would not represent a change in the overall rating of MI adherence for the individual provider, since both codes are an MI non-adherent behavior. And importantly, though advice giving is rare, the Lyssn algorithm still outperforms human reliability on both the REM providers and general test sets.

One limitation of the current dataset is that our clinical partners assessed demographic information in somewhat different formats; thus, we were not able to compare individual racial and ethnic groups in the current analyses. In future years we plan to develop better systems for combining race and ethnicity information from providers into a similar format. We also acknowledge that we have collected a finite population of racial ethnic backgrounds that were collected by current Lyssn customers. As we expand into new communities, the validity and generalizability of our algorithms may change. For this reason, we will continue to sample more data to add to our test set of REM providers.

There are other potential sources of bias that can contribute to an assessment tool like the Lyssn platform. We do not consider, for example, whether MISC standards themselves are equally applicable for all cultures or provider communication styles, though we do note that the Motivational Interviewing Network of Trainers is an international group, and MI is used widely around the world and across different cultures. This report is currently focused on the contribution of the natural language processing algorithms on the accuracy of the MISC predictions, as this is the source of bias most relevant to our users.

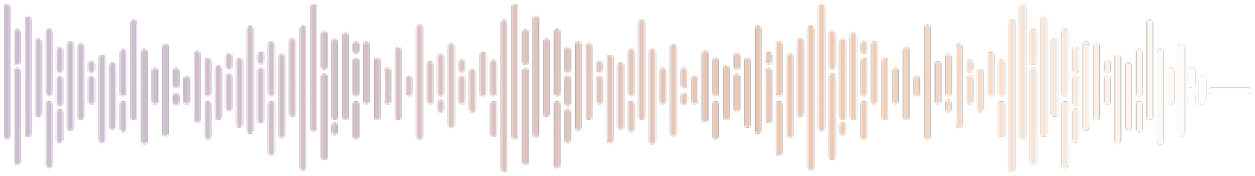
Future Directions

One of the most straightforward solutions to lower model performance of rare codes is to overselect for sessions that are likely to contain more of these behaviors. As such, over the next year, we plan to increase our human coding of sessions with REM providers and high levels of advice-giving to improve our model performance in this area. In addition, we plan to explore changes to our system for weighting model errors during training. One way we adjust for rare codes across all of the Lyssn algorithms is to increase the penalty for model errors for these codes. It is possible that some of the small bias observed in rare codes may be a byproduct of the specific weighting system that we use.

In future years, we also plan to examine the relative word error rates in the Lyssn speech recognition system for different accents. As the Lyssn platform is used in more diverse settings, we will increasingly encounter English dialects that were not previously seen in our speech training data.

There is almost always some increase in word error rates with a new English dialect, but it is important to quantify this change and report that information to our customers. Moreover, this will allow us to add these sessions to our training to reduce these errors in the future. This shows the two major advantages of the speech algorithm we use at Lyssn. First, that it is developed internally and optimized specifically for the type of speech that our customers encounter. And second, that it uses an end-to-end model that has no explicit notion of phonemes, which allows us to easily adapt it to new English dialects and accents by simply adding new training data.

Finally, we are encouraged that our initial results show that the Lyssn AI system shows virtually the same overall performance between REM providers and non-REM providers. Moving forward we at Lyssn plan to continue our work to monitor and improve our system and ensure that it functions in an equitable way for all users.



References

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '21). Association for Computing Machinery, New York, NY, USA, 610–623.

Bolukbasi, T., Chang, K., Zou, J., Venkatesh, S., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (NIPS '16). Curran Associates Inc., Red Hook, NY, USA, 4356–4364.

Crocker, L., & Algina, J. (2002). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston, 6277 Sea Harbor Drive, Orlando, FL.

Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018). Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (AIES '18). Association for Computing Machinery, New York, NY, USA, 67–73.

Henderson, H., Tanana, M., Bourgeois, J. W., & Adams, A. T. (2015). Psychometric racial and ethnic predictive inequities. *Journal of Black Studies*, 46(5), 462-481.

Miller, W. R., Moyers, T. B., Ernst, D., & Amrhein, P. (2003). *Manual for the motivational interviewing skill code (MISC)*. Unpublished manuscript. Albuquerque: Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico.

Voigt, R., Jurgens, D., Prabhakaran, V., Jurafsky, D., & Tsvetkov, Y. (2018). RtGender: A Corpus for Studying Differential Responses to Gender. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan.

Yang, Q., Steinfeld, A., Rosé, C., & Zimmerman, J. (2020). Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–13.

Appendix

Table 1. **REM Test Set Performance Metrics**

Code	Precision	Recall	F1 Score	F1: 95% CI
Advice-Giving	.43	.43	.43	.36, .49
Affirmation	.68	.74	.71	.68, .74
Facilitation	.95	.97	.96	.95, .96
Giving Information	.75	.67	.71	.69, .72
Closed Question	.74	.80	.77	.75, .79
Open Question	.85	.89	.87	.85, .89
Complex Reflection	.67	.70	.69	.67, .71
Simple Reflection	.70	.42	.53	.49, .57
All Codes	**.83**	**.83**	**.83**	**.82, .83**

Table 2. **General Population Test Set[3] Performance Metrics**

Code	Precision	Recall	F1 Score	F1: 95% CI
Advice-Giving	.56	.53	.54	.50, .58
Affirmation	.71	.72	.71	.69, .74
Facilitation	.96	.98	.97	.96, .97
Giving Information	.76	.75	.76	.75, .77
Closed Question	.76	.78	.77	.75, .78
Open Question	.84	.88	.86	.85, .88
Complex Reflection	.70	.66	.68	.66, .69
Simple Reflection	.58	.52	.55	.51, .59
All Codes	**.83**	**.83**	**.83**	**.82, .83**

[3] This is the test set for just the MISC set of labels, which is a small fraction of the total set of data used for training the full Lyssn platform.

Figure 3. Confusion matrix for general test set errors.

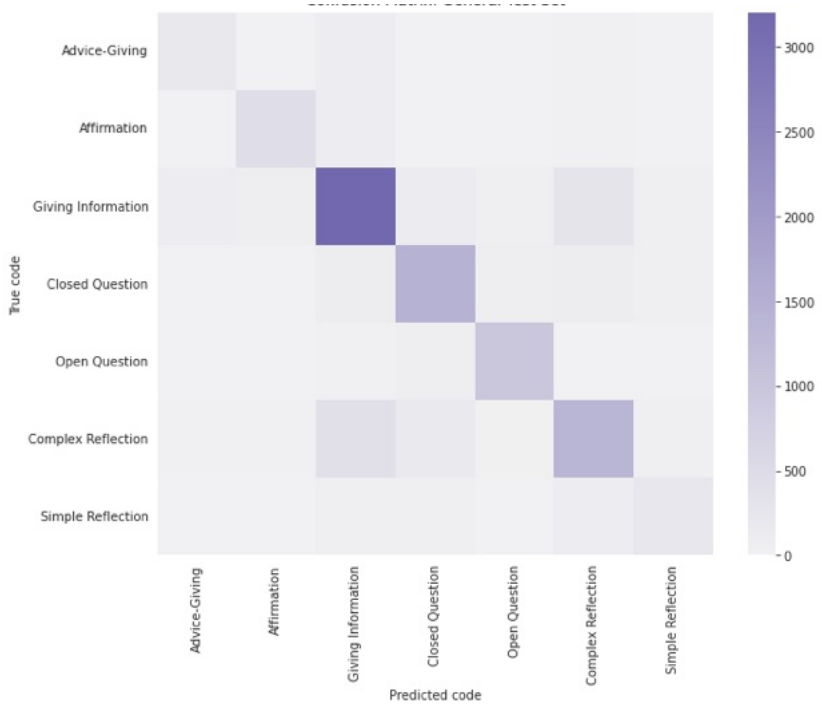


Figure 4. Confusion matrix for REM subgroup test set errors.