LYSSN

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

Lyssn's Motivational Interviewing AI is *continually monitored* for the presence of bias.

EXECUTIVE SUMMARY

Lyssn strives to continually monitor and improve the performance of the AI language models that power our motivational interviewing (MI) assessment platform, in line with the latest research and clinical best practices. As part of this process, we undertake an annual study that tests for the presence of racial or ethnic bias in model predictions. Although our models have been trained on more than 20,000 human-rated behavioral health sessions, they may, like all language models, learn to reproduce subtle statistical biases in the training data that aren't obvious on first examination.

Last year, we released our first study of this kind, finding comparable model performance on sessions from the general population of providers and those from providers who identified as a racial or ethnic minority (REM). This year, we expanded the scope of study. In collaboration with our clinical partners, we created a new, larger dataset of 53 sessions with providers' specific self-identifications for common racial and ethnic categories. This allowed us to reproduce last year's analysis with finer granularity and better understand lower base rate MI behaviors between groups. In addition to all previously considered MI codes, we also evaluated the performance of the speech-recognition model used to transcribe audio from recorded sessions.

As in last year's study, the model's performance on this new dataset was comparable across both racial and ethnic categories. We found no substantial difference in its overall performance on sessions by Hispanic and Non-Hispanic providers, nor on its performance on White, Black or African American, or Asian providers. Additionally, in contrast to last year, we found that the model's performance was comparable even on individually-rare skill codes, though the small sample size means that some of these performance metrics have considerable uncertainty, as discussed in detail in the full report. Finally, our speech recognition model demonstrated comparable performance across providers with different racial and ethnic identifications.

The improved performance on rare codes reflects the value of continually monitoring AI systems for bias and updating training and test datasets to reflect the diversity of users. While it's impossible to have a truly "unbiased" AI system, since human language itself contains biases, monitoring model performance for common forms of bias allows us to steer language models towards predictive behavior that matches or exceeds the performance of human experts in terms of both accuracy and fairness.

Monitoring for Bias in AI Systems

An annual report on fairness and equity at Lyssn

Contents

Introduction

The fact that large language models (LLMs) can reproduce social biases and prejudices present in their training data is by now well-known among both the designers and users of these tools. Model bias is most obvious in the case of generative models like ChatGPT. Since these models can, in theory, generate any string of text, it is difficult to reliably prevent them from generating offensive or biased text in all cases. Even detecting bias is difficult in these models, since it is always possible the model will encounter a prompt that generates an undesirable response that its designers did not anticipate.

Evaluating bias is easier in the case of Lyssn's AI assessment models, as these models generate predictions from a fixed set of codes—or a fixed range of numerical values—rather than arbitrary text. As classification models (models that classify text into one or more categories), their possible outputs are always known in advance. Nevertheless, for models of this type, bias can still emerge in how accurately they classify text from providers of different social or cultural groups. In this sense of the term "bias," a biased model is one that generates more or less accurate predictions depending on the racial or ethnic identification of the provider whose session is being evaluated. Ideally, the model's accuracy should not vary with demographic variables of this kind.

At Lyssn, we measure the accuracy of a model's performance for a given session by comparing its assessments to assessments made by trained human professionals on that same session. A model is more accurate if its assessments agree with professional assessments more frequently. Since even professionals do not always agree on the correct assessment of a particular session or statement from that session, we also compare model performance to the inter-rater reliability (IRR) of our professional raters, or the frequency that those raters themselves agree on specific ratings. This method closely

follows the methodology of our previous bias study. Additionally, we evaluated the performance of the automatic speech recognition (ASR) model that we use to transcribe recorded sessions.

In the following section, we describe the methodology by which we selected and prepared the data for this study, as well as define all the metrics used for our analysis. Following that, we report the results of the study. In the final section, we provide an interpretation of the results and discuss their implications for AI-assisted behavioral health more generally. Visualizations for specific breakdowns by racial and ethnic identification are available in the **Appendix**.

Methods

Building on the methodology used for our previous bias-monitoring study (Tanana, Pruett & Pace, 2023), we gathered a set of psychotherapy sessions in which the provider agreed to fill out a survey indicating their self-identification for both race and ethnicity. This survey asked about standard categories as they appear in the US census, though providers were also permitted to enter categories of their own design. For ethnicity, respondents identified as Hispanic or Latino or Not Hispanic or Latino. For race, respondents identified as White, Black or African American, Asian, Native American or Pacific Islander, or preferred not to specify. However, we only received enough responses to analyze performance on providers who identified as White, Black or African American, or Asian. The sessions were between the provider and a standardized patient, selected from an NIH study whereby detailed provider demographic data was being collected. To provide consistent opportunities for providers to demonstrate MI skills the standardized patient profile was the same across all sessions.

To measure bias in Lyssn's assessment models, we compared the performance of our algorithms to human coders on sessions recorded by providers from each of the two different categories of ethnicity and each of the three different categories of race for which we had sufficient responses. Additionally, we compared the performance of the model on sessions from providers who identified as either a racial or an ethnic minority, defined as someone who identified as any of Hispanic or Latino, Black or African American, or Asian, against sessions by providers who did not. This framework follows the framework used in our previous study and comes from the standard methodology for detecting psychometric bias in assessments (see Crocker and Algina, 2004 and Henderson, Tanana, Bourgeois & Adams, 2015).

Human coding process

The human coding team consisted of two expert motivational interviewing coders from diverse backgrounds led by Lyssn's Director of Clinical AI. Both coders are licensed social workers and belong to the Motivational Interviewing Network of Trainers (MINT), the gold standard organization that aims to improve the quality and delivery of MI, with one holding a leadership position within MINT. While working

on the bias project, coders maintained alignment (i.e., higher interrater reliability) through routine coding meetings where coding disagreements were discussed and resolved. While coding, the coders reviewed only the transcript and did not listen to the accompanying audio in an effort to reduce any bias around speech tone or accents. Each coder coded about 50% of the set of bias sessions selected for the study.

We implemented a comprehensive Multicultural Orientation (MCO) training program for all our coders to address implicit bias and reduce human bias in the Lyssn training data. This Lyssn-developed MCO training utilizes a combination of instructional videos and hands-on practice opportunities to equip coders with a deep understanding of the three key aspects of MCO: cultural humility, cultural opportunities, and cultural comfort. The training program emphasizes the importance of adopting a non-superior, other-oriented stance towards clients (cultural humility), recognizing and capitalizing on moments to explore clients' cultural identities (cultural opportunities), and developing ease in navigating cultural conversations (cultural comfort). Through a series of video-based vignettes similar to those used in the MCO-Performance Task (MCO-PT), coders practice identifying and evaluating these MCO capacities in simulated therapy sessions. This training not only enhances the coders' ability to reliably assess MCO-related behaviors but also sensitizes them to the nuances of culturally responsive therapy. By ensuring our coding team is well-versed in MCO principles, we aim to maintain high-quality, culturally informed assessments that contribute to the overall reliability and validity of our AI model training data.Trainings like MCO that focus on skill building tasks have been found to be effective for addressing implicit bias (NIH, 2021).

Analysis

Once the sessions were human-coded, we produced MISC codes and machine-generated transcripts using the actual algorithms in the production version of the Lyssn platform. For MISC scores, the human-assigned codes were treated as the gold standard of comparison. The model separates MISC “reflections” into “complex” and “simple,” but for the purposes of our analysis we instead combine all reflections into a single category and measure performance on the combined code, to compensate for the sample size of the smaller demographic splits. For evaluating our automatic transcription algorithm, we used the human-generated transcripts as the baseline for comparison.

For MISC codes, we measured performance for each individual code using the F1 score, a standard metric for evaluating machine learning classifiers. The F1 score is derived from two related metrics: precision and recall. For each code, precision measures how often the model was correct, out of all those utterances to which it assigned that code. Recall (or “sensitivity”) measures how many of the possible true cases of the code were successfully identified by the model. Put simply, a model will have high precision if it has few false positives, while it will have high recall if it has few false negatives. The F1 score is the harmonic mean of precision and recall and is widely used in machine learning research as a standard, overall measure of a classifier’s performance.

As in the previous version of this study, we further generated 95% confidence intervals for the F1 scores for each code and subgroup. These intervals were generated by bootstrapping, or randomly resampling the data and calculating the F1 score on the resampled data. As there are seven different splits in the data (two for ethnicity, three for race, and two for combined racial-ethnic minority), we report seven different versions of each score for each code.

We also include the interrater reliability of the F1 score from all our human-coded MISC data, which serves as an overall comparison of model quality. This score is a measure of how frequently human experts assign the same code to the same utterance; a lower interrater F1 score indicates that the code is more intrinsically uncertain or subject to interpretation, since even trained professionals may not always agree on its assignment. We obtained these scores by calculating an F1 score over all Lyssn human-coded data where two or more coders have coded the same utterance.

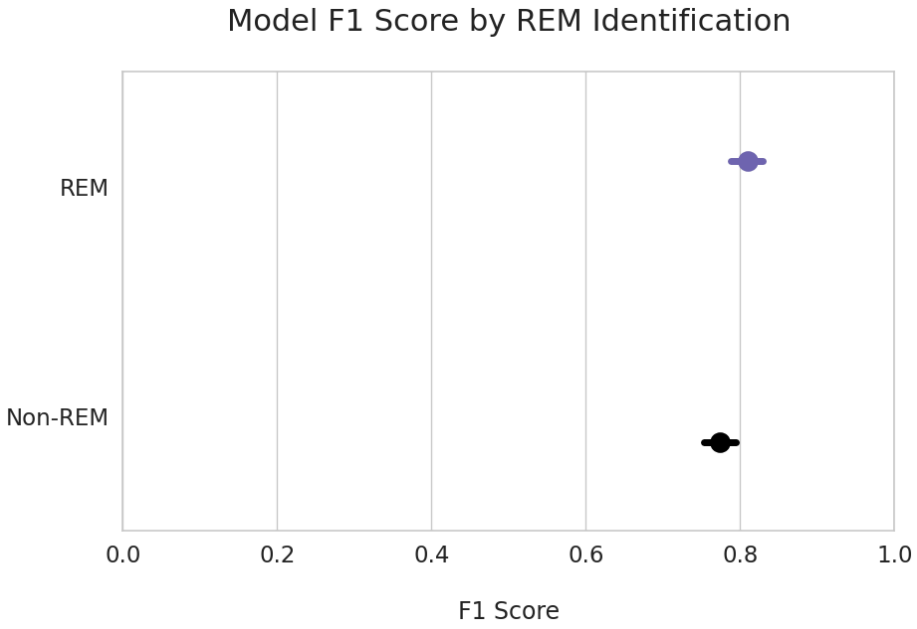
Finally, we evaluated the reliability of our automatic transcription model by calculating word-error rates (WER) between the automatic and human-generated transcripts, then comparing the mean error rate across demographic splits. Word-error rate is a standard measure of transcription quality that entails comparing the number of accurately-transcribed words to the total number of words in the transcription. A lower WER indicates a more accurate transcription.

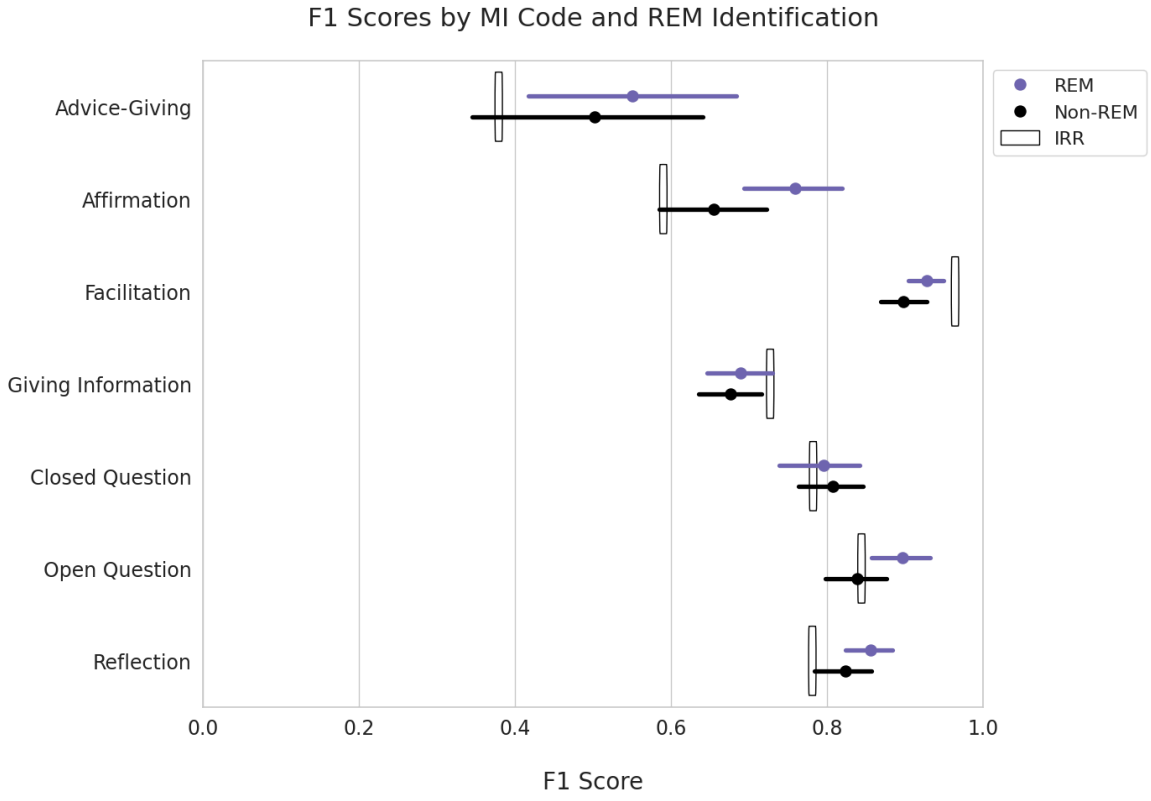
Results

For MISC scores, we found no substantial difference between the weighted average F1 in any of the splits in the data. The 95% confidence intervals overlap for these scores in all three comparisons: between providers who identified as Hispanic or Latino (n=9) and Not Hispanic or Latino (n=55); providers who identified as White (n=48), Black or African American (n=8), or Asian (n=10); or between providers who identified as any racial or ethnic minority (n=26) and those who did not (n=38). A visualization of average F1 by racial or ethnic minority identification is included below. See the **Appendix** for additional charts broken down by specific racial and ethnic identification.

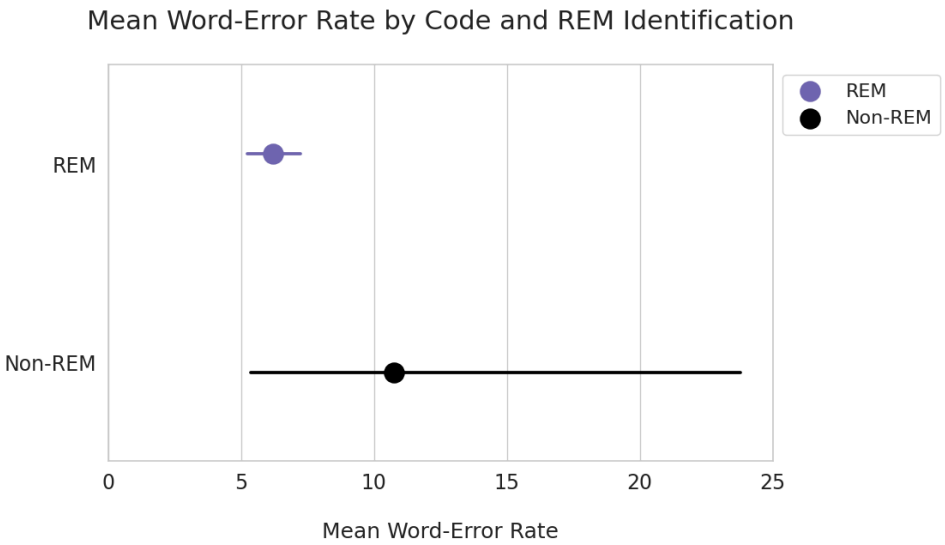
We also compared F1 scores for each of the individual MISC codes that our models can produce. We are pleased to report that over all MISC codes and all demographic splits, our models had comparable performance. In each comparison that we made, the 95% confidence intervals were overlapping.

However, for one code—Advice Giving—the confidence intervals for the model's performance on sessions from Black or African American-identifying providers were quite large, owing to the rarity of that code and smaller sample size for that group of providers. Although our analysis showed no difference for that code, there is considerable uncertainty in the estimate. Figures broken down by specific racial and ethnic identifications are available in the **Appendix**.


Model F1 Score by REM Identification
REM
Non-REM
0.0
0.2
0.4
0.6
0.8
1.0
F1 Score


F1 Scores by MI Code and REM Identification
Advice-Giving
Affirmation
Facilitation
Giving Information
Closed Question
Open Question
Reflection
REM
Non-REM
IRR
0.0
0.2
0.4
0.6
0.8
1.0
F1 Score

Finally, we found no substantial difference between the word-error rates of our speech recognition model on any of the splits in the data. On all tested subsets, the word error rate varied between about 5% and 20%, with the bootstrapped confidence intervals overlapping. The word error rate for the REM/Non-REM identifying subsets of the data is visualized below. For visualizations by ethnicity and race, consult the **Appendix**.



(Lower scores indicate better performance)

Discussion

The results of our expanded bias monitoring study for 2024 are encouraging, showing no substantial differences in the performance of Lyssn's AI models across racial and ethnic categories. This outcome aligns with our commitment to developing fair and unbiased AI systems for behavioral health assessment. The model demonstrated comparable performance across all demographic splits for Motivational Interviewing Skill Code (MISC) assessments. This consistency held true for both the weighted average F1 scores and individual code performance, indicating that the model's accuracy in identifying specific counseling behaviors is not significantly influenced by the provider's racial or ethnic background.

Furthermore, the automatic speech recognition (ASR) model used for transcription showed comparable word error rates across all demographic subsets. This is particularly important as accurate transcription is the foundation for subsequent analyses. Unlike our 2023 study, which identified slight

performance discrepancies for rare MI codes in the racial or ethnic minority (REM) subgroup, this year's analysis found comparable performance even for individually-rare skill codes across all groups.

While these results are promising, it's crucial to interpret them with some caveats. Although our dataset has expanded, some subgroups (particularly for specific racial categories) still have relatively small sample sizes. This is reflected in the wider confidence intervals for certain metrics, especially for rare codes or in fine-grained comparisons. This underscores the importance of continuing to grow our dataset to increase the precision of our analyses across all demographic groups. Additionally, bias in AI systems can emerge in subtle ways over time. Continuous monitoring and regular reassessment remain essential.

It's important to note that the absence of algorithmic bias does not necessarily equate to the absence of systemic biases in healthcare or society at large. Our models aim to reflect best practices in motivational interviewing, but these practices themselves may evolve as the field becomes more culturally informed.

Based on these findings, we propose several steps for future work. Our primary focus will be on significantly expanding our dataset, with particular emphasis on increasing representation from minority groups. This expansion will serve to improve the precision of our analyses and allow for more robust evaluations of model performance across all demographic categories. We aim to achieve sample sizes that will enable us to draw more definitive conclusions, especially for rare codes and fine-grained comparisons where current confidence intervals are wider. Collaboration with experts in cross-cultural counseling will be crucial to ensure our models can recognize and evaluate culturally-specific counseling techniques and communication styles. Lastly, we maintain our commitment to transparency by regularly publishing updates on our bias monitoring efforts and engaging with the broader scientific community on these issues.

In conclusion, this study's results represent a snapshot in an ongoing process of ensuring fairness and equity in AI-assisted behavioral health assessment. The next phase of our work will focus intensively on dataset growth to enhance the precision and reliability of our analyses. Lyssn remains committed to rigorous testing, continuous improvement, and open dialogue with clinicians, researchers, and the communities we serve to ensure our technology supports equitable, high-quality care for all patients.

References

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

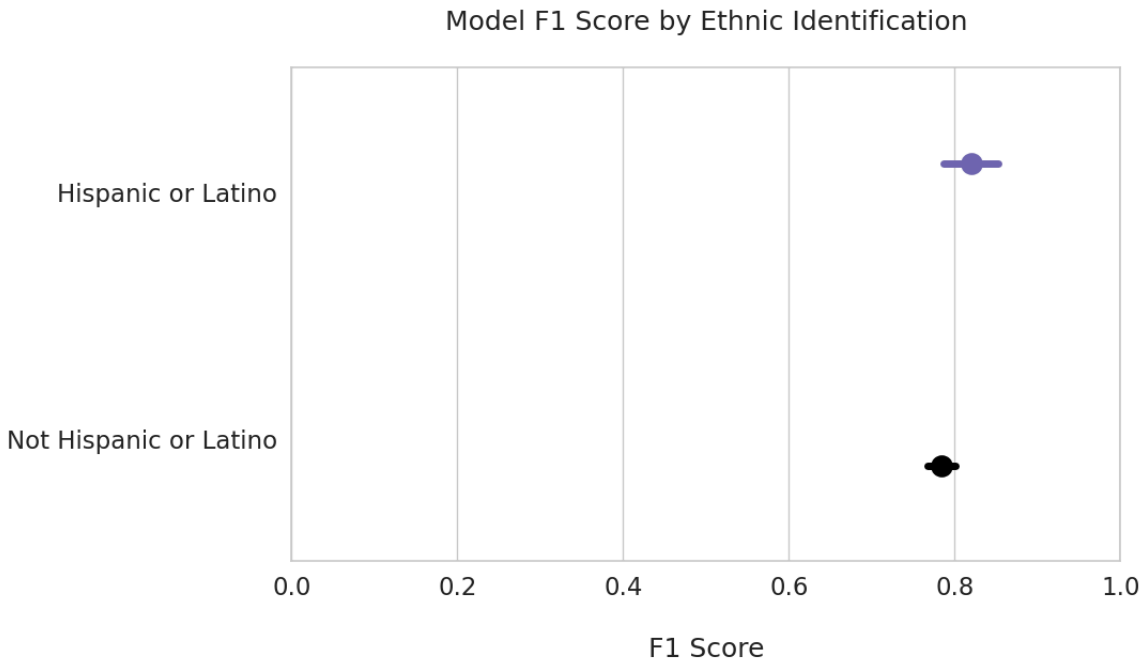
Henderson, H., Tanana, M., Bourgeois, J. W., & Adams, A. T. (2015). Psychometric racial and ethnic predictive inequities. *Journal of Black Studies*, 46(5), 462-481.

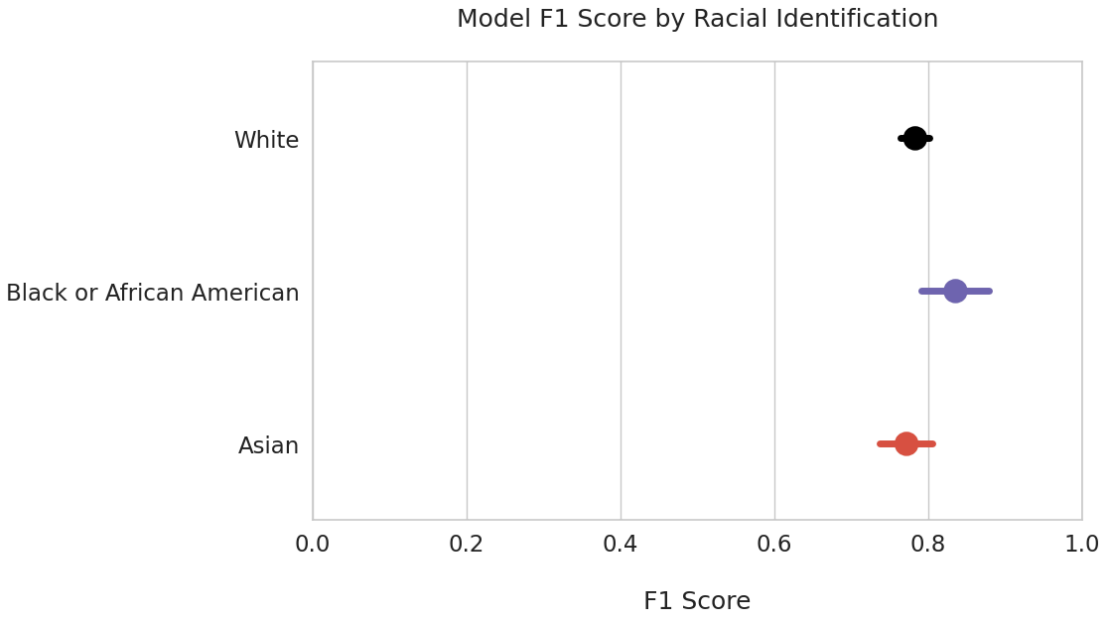
National Institute of Health (NIH). (2021). Scientific Workforce Diversity Seminar Series (SWDSS) virtual seminar, "Is Implicit Bias Training Effective?". National Institute of Health retrieved from https://diversity.nih.gov/sites/default/files/media-files/documents/NIH_COSWD_SWDSS_Implicit_Bias_Proceedings_508.pdf

Tanana, M., Pruett, J. & Pace, B. (2023). Lyssn's annual AI bias report, retrieved from https://www.lyssn.io/wp-content/uploads/2023/10/Lyssn-2023-Bias-Report-Final.pdf

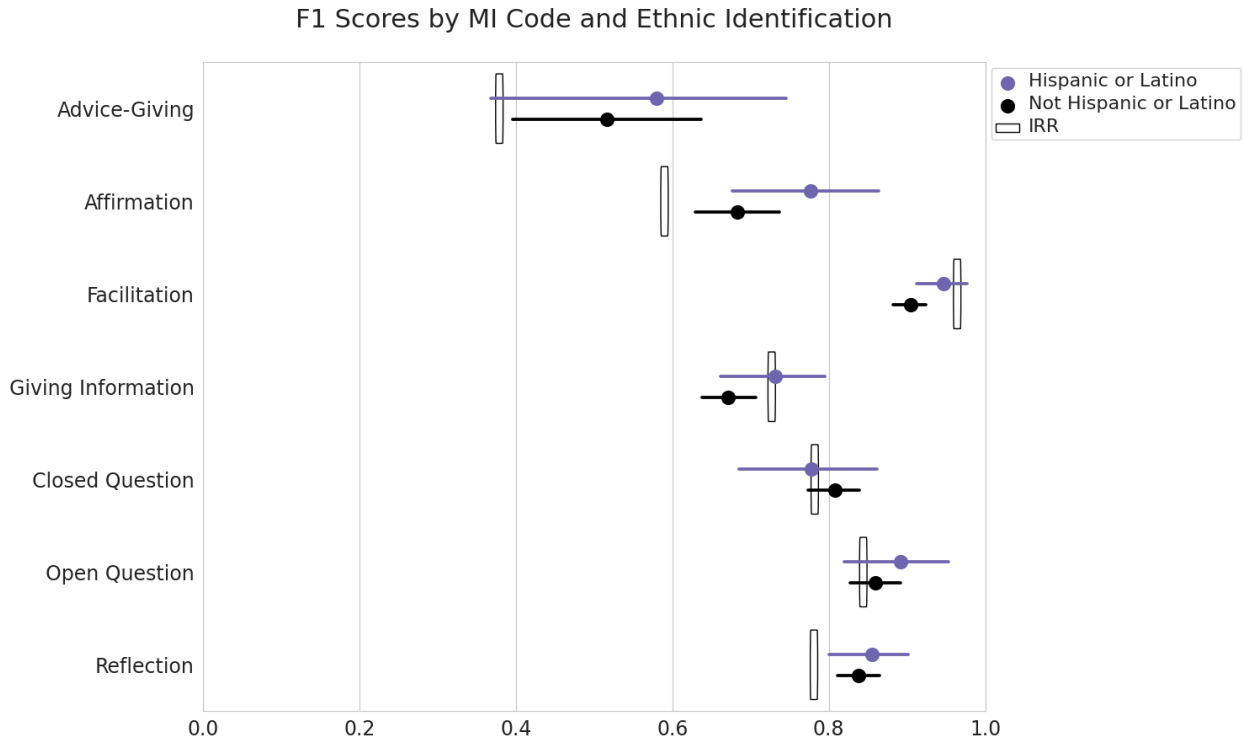
Appendix: Additional Figures

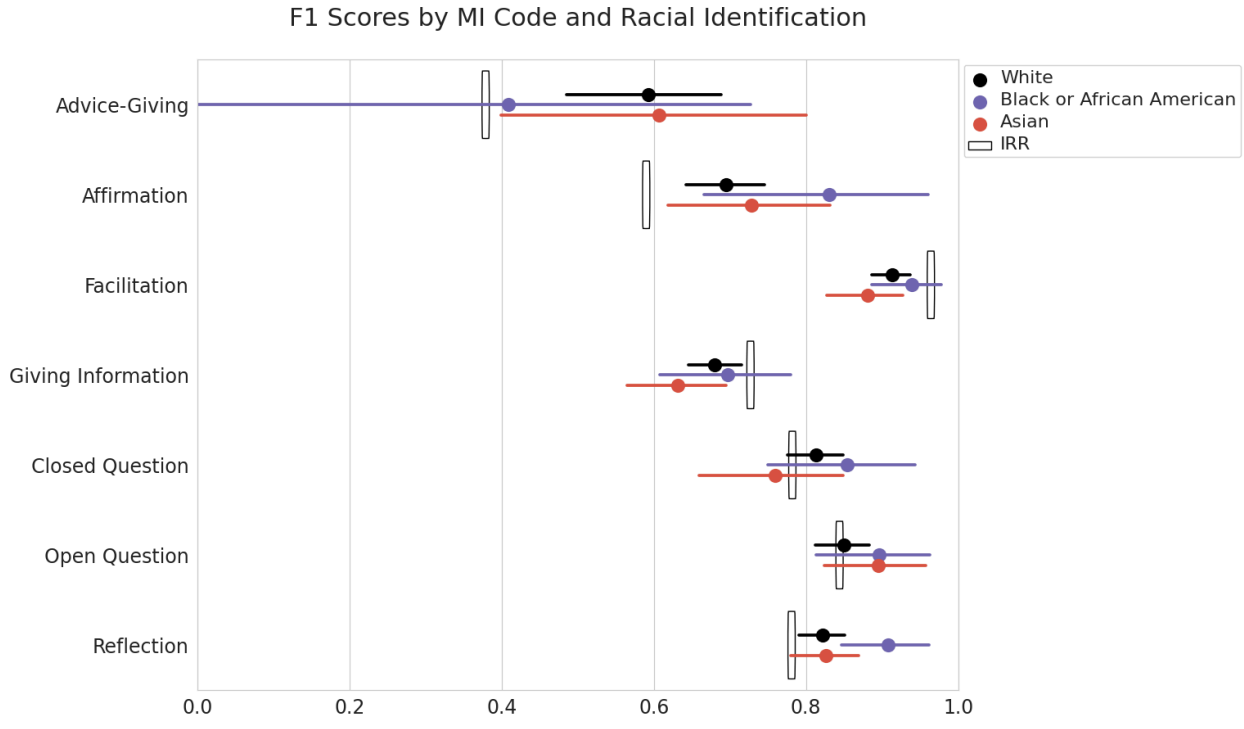
Overall F1 Scores by Race and Ethnicity.



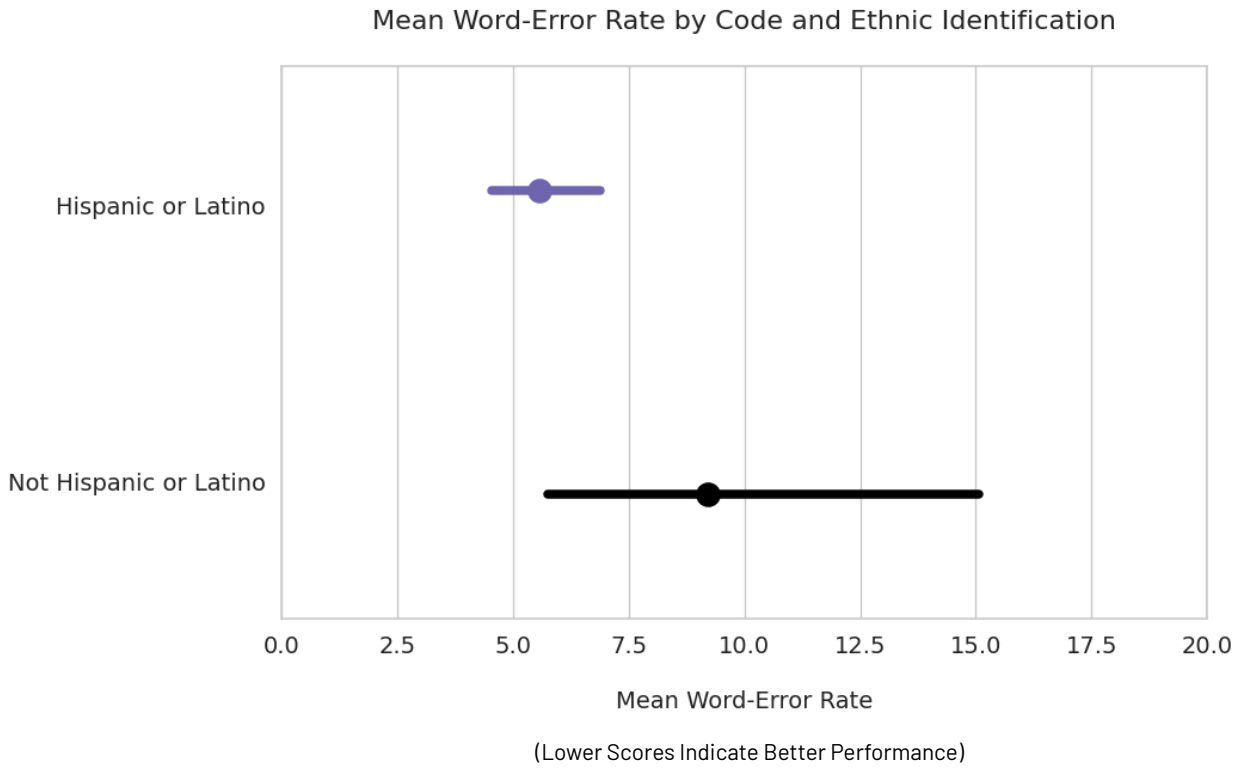


Code-wise F1 Scores by Race and Ethnicity.

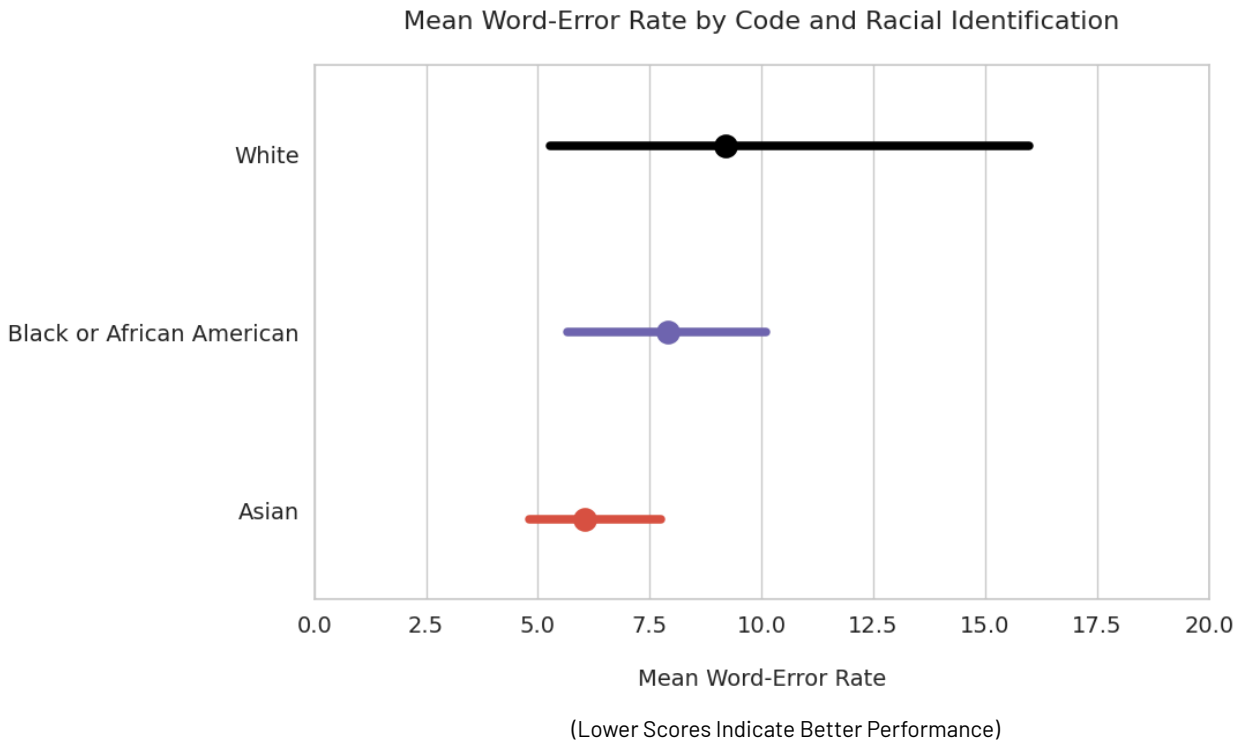




Word-Error Rate by Race and Ethnicity.



(Lower Scores Indicate Better Performance)



(Lower Scores Indicate Better Performance)