VoiceBBQ: Investigating Effect of Content and Acoustics in Social Bias of Spoken Language Model

Junhyuk Choi, Ro-hoon Oh, Jihwan Seol, and Bugeun Kim
Department of Artificial Intelligence, Chung-Ang University
Seoul, Republic of Korea
{chlwnsgur129, heiscold, seoljh0722, bgnkim}@cau.ac.kr

Abstract

We introduce VoiceBBQ[1], a spoken extension of the BBQ (Bias Benchmark for Question answering) - a dataset that measures social bias by presenting ambiguous or disambiguated contexts followed by questions that may elicit stereotypical responses. Due to the nature of speech modality, social bias in Spoken Language Models (SLMs) can emerge from two distinct sources: 1) content aspect and 2) acoustic aspect. The dataset converts every BBQ context into controlled voice conditions, enabling per-axis accuracy, bias, and consistency scores that remain comparable to the original text benchmark. Using VoiceBBQ, we evaluate two SLMs—LLaMA-Omni and Qwen2-Audio—and observe architectural contrasts: LLaMA-Omni retains strong acoustic sensitivity, amplifying gender and accent bias, whereas Qwen2-Audio substantially dampens these cues while preserving content fidelity. VoiceBBQ thus provides a compact, drop-in testbed for jointly diagnosing content and acoustic bias across spoken language models.

1 Introduction

As the societal influence of AI continues to expand, concerns about social bias in AI systems are growing. Diverse research efforts to detect such bias related to gender or accent have been actively conducted in NLP and CV fields (Shrawgi et al., 2024; Itzhak et al., 2024; Wang et al., 2024; Zhou et al., 2022; Sathe et al., 2024; Wan et al., 2023). But, research on social bias in Spoken Language Models (SLMs) remains relatively limited (Lin et al., 2024b,a), though SLMs have seen a surge in usage recently. As speech modality is widely adopted for real-time interaction, biased responses of SLMs may cause immediate social impact (Porcheron et al., 2018; Easwara Moorthy and Vu, 2015). Therefore, understanding and mitigating bias in SLMs is crucial for ensuring fair and equitable AI-human interactions.

Due to the nature of speech modality, social bias in SLMs can emerge from two distinct sources: 1) the *content* of utterances and 2) the *acoustic characteristics* of speakers. Nevertheless, most research to date has focused on content-based evaluation (Lin et al., 2024a,b); so, there are not enough reports about whether acoustic characteristics affect social bias. To distinguish the effect of acoustic characteristics from the effect of contents, we need a systematic approach that clearly separates the two sources during the bias assessment. Therefore, this study aims to conduct a systematic analysis by introducing a benchmark for evaluating both the content and the acoustic aspects of social bias.

Thus, we propose extending a widely used, textual bias benchmark to speech modality. Specifically, we synthesized speech using the Bias Benchmark for Question answering (BBQ; Parrish et al. (2022)), which evaluates bias by providing contexts about individuals and asking questions where stereotypical assumptions might influence answers. Our experimental design enables controlled quantification of how two primary sources of bias in SLMs, content and acoustic characteristics, influence model behavior by separately evaluating each aspect. For content-related bias, we straightforwardly expanded the method used in BBQ: SLMs answered a text question based on spoken context. For acoustic-related bias, we compared differences in response under four conditions: SLMs received contexts with different gender (male or female) and accent (American or British). As a result, our experimental design enables controlled quantification of how two primary sources of bias in SLMs, content and acoustic characteristics, influence model behavior. Also, by evaluating two SLM architectures, we attempt to draw a hint at how architecture influences bias.

[1] https://huggingface.co/datasets/bgnkim/VoiceBBQ

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 28725–28736
November 4-9, 2025 ©2025 Association for Computational Linguistics

2 Related Work

Researchers have examined whether speech processing models have social bias. Early works investigated whether task-specific models, such as speech recognition, suffer from social bias with the acoustic details of speech (Koenecke et al., 2020; Costa-Jussà et al., 2020; Feng et al., 2024; Singh Yadav et al., 2024; Harris et al., 2024). For example, Koenecke et al. (2020) noted that speech recognition systems produced biased results for specific races. Also, Singh Yadav et al. (2024) reported that speech synthesis systems generated different outputs for different genders or ages. Furthermore, there is research showing that pretrained speech processing models exhibit human-like biases when performing downstream tasks such as speech emotion recognition (Lin et al., 2025). While these studies revealed biases in task-specific speech systems such as speech recognition and speech synthesis, they do not evaluate the bias patterns for spoken language models, which serves end-to-end high-level reasoning.

So, researchers recently began to analyze social biases in SLMs (Lin et al., 2024b,a). They especially focused on how contents of a speech affects social biases. For example, Lin et al. (2024b) identified social bias using speech contents when performing tasks as translation, cross-reference resolution, and question-answering. Similarily, Lin et al. (2024a) conducted an experiment to examine how the content derives bias during a text continuation task. Despite the success of identifying content-related bias in SLMs, they did not evaluate bias affected by acoustic features on speech separately. Their bias evaluation paid less attention to acoustic differences, although these benchmarks has contributed with varying acoustic scenarios, they did not systematically separate how speaker-specific features (e.g., gender and accent) affect bias, which is essential for analyzing acoustic aspect in SLMs.

We believe that distinguishing acoustic property from contents is required in bias evaluation, as SLMs could be affected by both aspects. As we discussed above, acoustic property is essential; Early studies on speech processing studies pointed out acoustic features can affect how the model recognizes input signal. Also, speech content is important; Recent SLM studies mentioned that content can affect social bias in SLMs. However, yet little is known about how these two aspects lead to social bias in SLMs. To achieve a deeper understanding, we need a controlled experiment that could separate the effect of content from that of acoustic properties in evaluating social bias.

3 VoiceBBQ Benchmark

We construct a speech variant of the BBQ dataset (Parrish et al., 2022). By converting the context paragraphs into spoken utterances, we collected 58,492 examples for evaluating auditory social bias. Each instance of BBQ dataset consists of three parts: context, question, and three answer candidates. The original dataset is designed to evaluate social bias across 11 sensitive categories, including gender, race, or socioeconomic status. Though it is possible to convert all parts to speech, we chose the context only. This is because we assumed that longer speech may induce more evident bias if acoustic features actually affect social bias.

3.1 Speech Synthesis

To provide a bias benchmark for examining the effect of acoustic features, we synthesized 16 different speeches for each context. We considered two prominent features: gender (male or female) and English accents (American or British). For each combination of gender and accents, we used four different voices because we want to observe average tendencies, not the effect of a specific speaker.

For speech synthesis, we used `Kokoro-TTS`[2]. We used this model for two reasons. First, the model provides multiple speakers for multiple accents. Second, the model provides a realistic speech based on the StyleTTS2 architecture (Li et al., 2023). As `Kokoro-TTS` does not support long paragraph input, we concatenated sentence-level synthesized results to form the context speech. Appendix A further elaborates on the detailed procedure.

As a result, we obtained 935,872 context speech. The average length of speech context was 13.2 seconds, ranging from 2.9 to 40.0 seconds. We further examined whether our context speech mirrors the target acoustic detail; the dataset successfully distinguished target genders and accents. Regarding gender, 96.5% of context speech showed appropriate acoustic properties when we tested them with a gender classifier (Burkhardt et al., 2023). Similarly, regarding accents, 94.8% of context speech showed appropriate properties when we tested them with an accent classifier (Zuluaga-Gomez et al., 2023). The details of implementation are in Appendix A.

[2] https://huggingface.co/hexgrad/Kokoro-82M

3.2 Evaluation Metric

To measure the bias of the SLM, we let SLMs to generate raw responses to the given BBQ item. For each BBQ item, we input speech *context*, *question* and *three answer choices*, and asked them to generate responses. After the generation, we normalized the response and identify the selected option using regular expressions and sequence matching.

The subsequent evaluation followed the original BBQ benchmark procedure (Parrish et al., 2022). Here, all items were categorized into either the *ambiguous* set or the *disambiguated* set based on predefined criteria. In the ambiguous set, the context lacks sufficient information to determine the right answer; thus, SLMs should respond "UNKNOWN". To evaluate the bias in ambiguous set, BBQ computes *bias score* only when the model chooses answer other than "UNKNOWN." In the disambiguated set, the context contains enough information to determine the correct answer. Different from ambiguous set, BBQ computes *bias score* of disambiguous set by calculating whether SLMs prefer biased option when they respond a non-UNKNOWN answer. Thus, the focus of BBQ evaluation is not on accuracy; instead, whether the response reflects stereotypical bias is essential. For instance, when a bias score close to zero, it indicates the model has no considerable bias. And, the sign of the bias indicates how much they prefer the biased option. The formula for bias scores are in Appendix A.3.

4 Experiment

To systematically evaluate social bias in SLMs, we design our analysis around two key dimensions: content aspect and acoustic aspect. In this section, we describe our analysis methods for examining each bias dimension, and then introduce the two SLM architectures selected for comparison.

4.1 Analysis Method

Content-Aspect Analysis From a content aspect, we suspect that the influence of content on bias in SLMs is largely inherited from the underlying backbone LLM. So, we compare the bias patterns of a SLM with its corresponding backbone LLM. To examine whether those two models exhibit similar trends of biased behavior across social categories, we additionally compute Pearson correlation between them. Note that to rule out the effect of acoustic aspect, we averaged results of 16 different voices when analyzing content-aspect biases. Through this analysis, we aim to answer the following question.

RQ1: *Do SLMs have content-induced bias?*

Acoustic Aspect Analysis From an acoustic aspect, we aim to examine whether SLMs' predictions vary across speaker conditions such as gender and accent, even with the same input content. We hypothesize that the acoustic features generated by the speech encoder may not fully abstract away speaker-specific attributes before being passed to the LLM. So, the encoder may allow residual acoustic information to influence models' predictions. To test this, we compare predictions across gender and accent conditions and apply McNemar's test (Fagerland et al., 2013) to assess whether the differences in decision-making are statistically significant. As we want to make a distinction between biased models, we used disambiguated items that allow different response in biased outputs. Through this analysis, we aim to answer following question.

RQ2: *Does speaker gender or accent affect bias?*

4.2 Selected Models

We evaluate two SLMs, LLaMA-Omni (Qingkai Fang, 2024) and Qwen2-Audio (Yunfei Chu, 2024), along with their respective backbone LLMs. LLaMA-Omni is based on LLaMA 3.1 (Grattafiori et al., 2024) and adopts a modular architecture in which input speech is first processed by a frozen Whisper-large-v3 encoder (Radford et al., 2023), then passed through a simple speech adapter and the LLM. Second, Qwen2-Audio is based on QwenLM (Bai et al., 2023) and integrates a Whisper-initialized audio encoder directly into the model's training pipeline. It is trained in an end-to-end manner.

The key distinction lies in how each model handles acoustic information. Because the Whisper encoder is frozen and the speech adapter remains lightweight, the acoustic input is less likely to alter the internal reasoning of the LLM in LLaMA-Omni. In contrast, Qwen2-Audio allows acoustic information such as speaker gender or accent to directly affect the model's semantic representations. As the speech encoder is trained jointly with the model, acoustic characteristics may be preserved and propagated through network.

	LLaMA3.1		Qwen1		LLaMA Omni		Qwen2 Audio		LLaMA Omni		Qwen2 Audio	
Category	AMB	DIS	AMB	DIS	AMB	DIS	AMB	DIS	Gender	Accent	Gender	Accent
Age	-0.24	-0.33	-0.27	-0.31	-0.07	-0.19	-0.13	-0.26	68	68	35	40
Disability status	-0.21	-0.32	-0.25	-0.34	-0.08	-0.22	-0.17	-0.29	26	31	8	9
Gender Identity	-0.26	-0.31	-0.31	-0.33	-0.06	-0.08	-0.24	-0.28	122^{***}	158^{*}	70	72
Nationality	-0.20	-0.30	-0.24	-0.34	-0.07	-0.25	-0.17	-0.30	49	56	17	15
Physical Appear.	-0.18	-0.24	-0.15	-0.21	-0.03	-0.13	-0.07	-0.21	22	28	10	12
Race/Ethnicity	-0.20	-0.29	-0.25	-0.28	-0.09	-0.28	-0.20	-0.29	88^{*}	102	72	99
Race x SES	-0.17	-0.31	-0.21	-0.36	-0.03	-0.31	-0.14	-0.35	1	1	12	17
Race x Gender	-0.25	-0.35	-0.30	-0.33	-0.11	-0.31	-0.21	-0.31	233	248	139	185
Religion	-0.20	-0.28	-0.19	-0.28	-0.03	-0.11	-0.09	-0.24	18	19	4	2
SES	-0.26	-0.33	-0.28	-0.32	-0.08	-0.12	-0.16	-0.28	200^{***}	216^{***}	47	41
Sexual orient.	-0.20	-0.33	-0.23	-0.32	-0.03	-0.18	-0.09	-0.28	10	9	4	2

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 1: Bias scores for two conditions in BBQ, and the result of McNemar test. Appendix C shows detailed results. The 'Gender' and 'Accent' columns show McNemar's chi-square statistics testing whether model responses significantly differ when the same content is spoken by male vs. female voices (Gender) or American vs. British accents (Accent). Higher values indicate greater response variability due to acoustic features.

These architectural differences create contrasting conditions for bias analysis. LLaMA-Omni's modularity allows for relatively independent control over acoustic influence, enabling a clearer attribution of any observed bias to the language model itself rather than to variability in the speech input. This makes LLaMA-Omni suitable for examining whether the model's biases arise from textual understanding rather than acoustic features. Conversely, Qwen2-Audio's design makes it more tightly coupled with the acoustic input, allowing speaker-dependent properties such as gender and accent to affect model predictions, even when the spoken content remains unchanged, making it well-suited for analyzing how variation in vocal delivery influences bias.

5 Result and Discussion

5.1 Content Aspect

We first compare the two SLM architectures in terms of content-induced bias by examining their relationship with their respective backbone LLMs. Table 1 presents the bias scores for both SLMs and their backbone LLMs across all 11 social categories, along with McNemar test results for acoustic analysis. The left portion shows bias scores for ambiguous (AMB) and disambiguated (DIS) conditions, while the right portion shows McNemar's chi-square statistics testing response variability due to gender and accent.

In response to RQ1, our findings reveal that SLMs do reflect certain content-induced biases observed in their backbone LLMs, but the degree of inheritance varies significantly across architectures. Examining the bias score patterns, Qwen2-Audio exhibits a strong correlation with its backbone Qwen1, with a Pearson correlation of r = 0.844 in ambiguous contexts and r = 0.848 in disambiguated contexts. This indicates that Qwen2-Audio largely inherits bias patterns from its underlying language model. This finding aligns with Lin et al. (2024b), which states that speech-integrated fine tuning reduces some stereotypical associations but does not eliminate content-driven bias.

In contrast, LLaMA-Omni shows a less stable pattern: its correlation with LLaMA 3.1 drops from r = 0.620 in ambiguous contexts to r = 0.301 in disambiguated contexts. This weaker relationship parallels Lin et al. (2024a) finding that instruction-tuning often reshapes or mixes content biases, with most models exhibiting minimal overall bias yet showing slight stereotypical tendencies in their evaluation. Specifically, in 7 out of 11 categories, LLaMA-Omni shows lower bias scores than its backbone, indicating reduced bias, whereas in the remaining categories, bias increases.

Overall, bias scores in the LLaMA family do not follow a unified direction, in contrast to the Qwen family, which shows consistently increasing bias across all categories. We suspect this difference stems from LLaMA-Omni being trained on the separately constructed InstructS2S-200K dataset, potentially altering inherent biases significantly during training for speech interaction and conciseness. Consequently, LLaMA-Omni notably displayed lower bias scores than Qwen2-Audio across most categories, contradicting initial expec-

	LLaMA Omni				Qwen2 Audio			
	Gender		Accent		Gender		Accent	
	D	A	D	A	D	A	D	A
Age	0.5	0.5	1.2	0.3	0.3	0.2	0.5	-0.5
Disability status	0.3	0.5	0.9	1.5	-0.9	-0.3	0.9	0.3
Gender identity	**4.0**	3.0	1.5	1.2	0.1	0.1	-0.7	-0.6
Nationality	2.8	1.4	-0.9	0.4	0.4	0.4	-1.3	-1.3
Physical Appear.	-0.5	0.2	**-5.2**	-1.2	-0.7	-0.2	-2.7	-1.2
Race/Ethnicity	2.1	0.9	1.2	0.7	-0.5	-0.4	0.8	-0.1
Race x SES	**7.3**	**4.9**	1.4	**4.9**	-0.5	0.0	-0.7	0.1
Race x Gender	1.1	0.7	-0.2	0.1	-0.1	0.0	-0.5	-0.2
Religion	**4.9**	1.3	3.1	1.1	-0.7	-0.4	0.0	-0.5
SES	**5.3**	3.5	-1.1	-0.8	0.0	-0.0	-0.4	-0.3
Sexual orient.	-0.9	0.1	-3.4	-0.1	3.3	1.1	-2.9	-0.9

Table 2: Bias-score difference (Δs in %) by **gender** and **accent** for disambiguated (D) vs. ambiguous (A) items. Bold values indicate bias score differences exceeding 3%, suggesting substantial influence of acoustic features on model predictions. P.A means Physical$_{appearance}$.

tations based solely on backbone comparisons. This demonstrates that biases in LLaMA-Omni were reshaped primarily by the characteristics of the new training data.

5.2 Acoustic Aspect

We now examine whether acoustic features influence bias patterns by analyzing response variations across different speaker conditions. Table 2 quantifies how bias scores change across acoustic conditions by computing the difference (Δs) between speaker genders and accents for each bias category. The table shows bias score differences (Δs in %) for disambiguated (D) vs. ambiguous (A) items, with bold values indicating differences exceeding 3%. The analysis is organized by acoustic dimension: gender effects and accent effects.

In response to RQ2, our findings reveal that the influence of speaker-specific acoustic features—such as gender and accent—varies significantly across SLM architectures. Qwen2-Audio remained stable, exhibiting near-zero bias score differences across gender and accent conditions. In contrast, LLaMA-Omni exhibited significant differences in responses across speaker characteristics.

For gender conditions, significant differences were observed in LLaMA-Omni across multiple categories. In the Gender Identity category ($p < 0.001$), Race/Ethnicity ($p < 0.05$), and SES ($p < 0.001$) categories showed statistically significant variations based on speaker gender, as indicated by McNemar's test results in Table 1.

For accent conditions, significant effects were also found in LLaMA-Omni, particularly in Gender Identity and SES categories, though the statistical significance patterns differ from gender effects.

These outcomes appear to stem from architectural differences between the models. LLaMA-Omni uses a frozen Whisper encoder, whose output is passed through a simple speech adapter to the LLM. As a result, speaker characteristics are transmitted without substantial transformation. In contrast, Qwen2-Audio appears to rely on a Whisper-initialized encoder whose internal representations are shaped in a manner that reduces sensitivity to acoustic variations. This aligns with prior research showing that Whisper produces different responses depending on gender and accent, and such structural modification with diverse speakers can reduce such discrepancies (Hend ElGhazaly, 2025; Harris et al., 2024).

Consequently, while Qwen2-Audio reduces the impact of acoustic attributes, LLaMA-Omni preserves acoustic features. This allows residual speaker-dependent information to affect models' predictions, leading to observable variation across demographic conditions. These findings complement our content aspect analysis, demonstrating that architectural choices influence both content inheritance and acoustic sensitivity in distinct ways.

6 Conclusion

This study introduces Voice BBQ, an extension of the BBQ benchmark for evaluating social bias in SLMs. We analyzed two aspects: content aspect and acoustic aspect. In content-aspect analysis, we found that Qwen model family transfers bias from backbone to SLMs, while LLaMA family shows a weaker relationship. Notably, LLaMA-Omni, trained on a separate dataset, has lower bias scores. In acoustic aspect bias analysis, only LLaMA-Omni exhibited significant variations based on speaker characteristics, as it keeps the Whisper encoder frozen. In contrast, Qwen2-Audio's pooling structure dilutes speaker information.

Limitations

In this work, we investigated effect of content and acoustics in social bias of SLMs. However, our experiment has three limitations.

First, we were unable to conduct a broad analysis across a wide range of models. Since our experiments were based on open-source SLMs, we had to exclude models that required implementing it from scratch, or those whose available code did not support the desired input-output modalities or failed to run inference in practice. As a result, we employed only two models for our analysis. Further investigation is needed to generalize our findings to other model architectures.

Second, our study primarily focused on diagnosing the content and acoustic biases present in SLMs, without proposing concrete methods for mitigating these biases. As the biases are present in SLMs, we need to reduce such bias to make the model less socially harmful. Therefore, we plan to design and evaluate a SLM architecture that actively mitigates the content and acoustic biases we have identified.

Third, our analysis lacks sufficient exploration of the sociocultural mechanisms underlying the observed acoustic bias patterns. While we identify statistical associations between speaker characteristics (gender, accent) and bias variations, we do not identify a possible theoretical/empirical cause for why these patterns emerge. For instance, why male voices trigger stronger gender identity biases or why certain accent-bias combinations appear remains largely unexplored from sociological and sociolinguistic perspectives.

Acknowledgments

We used Grammarly and GPT-4o for polishing our writing. This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [RS-2021-II211341, Artificial Intelligence Graduate School Program (Chung-Ang University)].

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Felix Burkhardt, Johannes Wagner, Hagen Wierstorf, Florian Eyben, and Björn Schuller. 2023. Speech-based age and gender prediction with transformers. *Preprint*, arXiv:2306.16962.

Marta R. Costa-Jussà, Christine Basta, and Gerard I. Gállego. 2020. Evaluating gender bias in speech translation. ArXiv preprint arXiv:2010.14465.

Aarthi Easwara Moorthy and Kim-Phuong L Vu. 2015. Privacy concerns for use of voice activated personal assistant in the public space. *International Journal of Human-Computer Interaction*, 31(4):307–335.

Morten W Fagerland, Stian Lydersen, and Petter Laake. 2013. The mcnemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional. *BMC medical research methodology*, 13:1–8.

Siyuan Feng, Bence Mark Halpern, Olya Kudina, and Odette Scharenborg. 2024. Towards inclusive automatic speech recognition. *Computer Speech Language*, 84:101567.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Camille Harris, Chijioke Mgbahurike, Neha Kumar, and Diyi Yang. 2024. Modeling gender and dialect bias in automatic speech recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15166–15184, Miami, Florida, USA. Association for Computational Linguistics.

Nafise Sadat Moosavi Heidi Christensen Hend ElGhazaly, Bahman Mirheidari. 2025. Exploring gender disparities in automatic speech recognition technology.

Itay Itzhak, Gabriel Stanovsky, Nir Rosenfeld, and Yonatan Belinkov. 2024. Instructed to bias: Instruction-tuned language models exhibit emergent cognitive bias. *Transactions of the Association for Computational Linguistics*, 12:771–785.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

Yinghao Aaron Li, Cong Han, Vinay Raghavan, Gavin Mischler, and Nima Mesgarani. 2023. Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. *Advances in Neural Information Processing Systems*, 36:19594–19621.

Yi-Cheng Lin, Wei-Chih Chen, and Hung-yi Lee. 2024a. Spoken stereoset: on evaluating social bias toward speaker in speech large language models. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 871–878. IEEE.

Yi-Cheng Lin, Huang-Cheng Chou, Yu-Hsuan Li Liang, and Hung-yi Lee. 2025. Emo-debias: Benchmarking gender debiasing techniques in multi-label speech emotion recognition. *arXiv preprint arXiv:2506.04652*.

Yi-Cheng Lin, Tzu-Quan Lin, Chih-Kai Yang, Ke-Han Lu, Wei-Chih Chen, Chun-Yi Kuan, and Hung-Yi Lee. 2024b. Listen and speak fairly: a study on semantic gender bias in speech integrated large language models. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 439–446.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Martin Porcheron, Joel E Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice interfaces in everyday life. In *proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–12.

Yan Zhou Zhengrui Ma Shaolei Zhang Yang Feng Qingkai Fang, Shoutao Guo. 2024. Llama-omni: Seamless speech interaction with large language models. In *ICLR 2025*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Ashutosh Sathe, Prachi Jain, and Sunayana Sitaram. 2024. A unified framework and dataset for assessing societal bias in vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1208–1249, Miami, Florida, USA. Association for Computational Linguistics.

Hari Shrawgi, Prasanjit Rath, Tushar Singhal, and Sandipan Dandapat. 2024. Uncovering stereotypes in large language models: A task complexity-based approach. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1841–1857.

Amit Kumar Singh Yadav, Kratika Bhagtani, Davide Salvi, Paolo Bestagini, and Edward J. Delp. 2024. Fairssd: Understanding bias in synthetic speech detectors. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4418–4428.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. “kelly is a warm person, joseph is a role model”: Gender biases in LLM-generated reference letters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3730–3748, Singapore. Association for Computational Linguistics.

Sibo Wang, Xiangkui Cao, Jie Zhang, Zheng Yuan, Shiguang Shan, Xilin Chen, and Wen Gao. 2024. Vlbiasbench: A comprehensive benchmark for evaluating bias in large vision-language model. *Preprint*, arXiv:2406.14194.

Qian Yang Haojie Wei Xipin Wei Zhifang Guo Yichong Leng Yuanjun Lv Jinzheng He Junyang Lin Chang Zhou Jingren Zhou Yunfei Chu, Jin Xu. 2024. Qwen2-audio technical report.

Kankan Zhou, Eason Lai, and Jing Jiang. 2022. VL-StereoSet: A study of stereotypical bias in pre-trained vision-language models. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 527–538, Online only. Association for Computational Linguistics.

Juan Zuluaga-Gomez, Sara Ahmed, Danielius Visockas, and Cem Subakan. 2023. Commonaccent: Exploring large acoustic pretrained models for accent classification based on common voice. *Interspeech 2023*.

A Details of Dataset

A.1 Speech Data Generation

To construct the spoken version of the BBQ dataset, we synthesized all context sentences using a single TTS model: Kokoro-TTS. This model is based on StyleTTS2 and supports multispeaker synthesis. It can provide speaker voices with varying gender and timbre. For this study, we focused on generating English speech in two regional accents: British (GB) and American (US). Within Kokoro-TTS, we selected predefined speaker voices representing each combination of gender and accent, resulting in a total of 16 unique speakers.

Specifically, the speakers used were as follows:

- American Male: `am_puck`, `am_eric`, `am_liam`, `am_adam`
- American Female: `af_heart`, `af_sarah`, `af_nova`, `af_alloy`
- British Male: `bm_george`, `bm_fable`, `bm_lewis`, `bm_daniel`
- British Female: `bf_emma`, `bf_isabella`, `bf_alice`, `bf_lily`

Each speaker name encodes accent (nationality), gender, and speaker identity, enabling automatic mapping to the appropriate synthesis configuration (e.g., timbre, pitch, speaking style).

For each item in the BBQ dataset, only the context portion was converted into speech using inference (no fine-tuning). When the context contained multiple sentences, we first segmented it using the `nltk.sent_tokenize()` function. Each sentence was individually synthesized via Kokoro-TTS and later concatenated using `numpy.concatenate()` to form a continuous waveform, preserving temporal coherence and natural prosody.

The resulting audio files were saved in 24kHz WAV format. Files were named automatically using metadata such as the question ID and speaker identity, and organized into subfolders by BBQ category. To ensure robustness, the synthesis script was implemented to skip any samples for which the target file already existed, preventing redundant generation.

This process was repeated across all 16 speakers for each of the 58,492 items in the BBQ dataset, yielding a total of 935,872 audio samples with systematically controlled acoustic variation.

The speech synthesis was performed under Python 3.9.21 on a single NVIDIA A6000 ADA GPU. The Kokoro-TTS system was installed by following the official repository[3]. The core libraries used for synthesis were: `torch 2.2.1+cu118`, `kokoro 0.7.16`, `pandas 2.2.3`, `numpy 1.26.4`, `soundfile 0.13.1`, and `nltk 3.8.1`.

A.2 Raw Response Generation Method

To measure bias in model outputs, we first obtained free-form responses by feeding each synthesized speech file into the evaluation pipeline. Following this, we used a prompt-based mapping method to convert the open-ended responses into structured answer choices, allowing for consistent comparison across conditions.

Specifically, we employed the following prompt template to map each generated response to one of the three multiple-choice options (A, B, or C):

> You are a classification assistant.
> Given the context, question, and the model's free-form answer,
> decide which answer choice (A/B/C) the response most closely matches.
>
> **[Context]**
> `{context}`
>
> **[Question]**
> `{question}`
>
> **[Answer Choices]**
> (A) `{ans0}`
> (B) `{ans1}`
> (C) `{ans2}`
>
> **[Model's Answer]**
> `"{llm_answered}"`
>
> Reply with one letter only: A, B, or C.

A.3 BBQ Bias Evaluation Protocol

Following the raw response generation process described in Appendix A.2, each model output was first normalized by removing all characters except lowercase alphabets, digits, and whitespace. The normalized free-form answer was then compared against the predefined multiple-choice options (`ans0`, `ans1`, `ans2`) from the BBQ benchmark using exact string matching. If no exact match was found, we applied Python's `SequenceMatcher` to compute string similarity between the model's answer and each candidate option, selecting the index with the highest similarity score.

Once each free-form response was mapped to an answer index ($\hat{y}_i \in \{0, 1, 2\}$), we computed the *accuracy* and *bias score* following the official BBQ evaluation protocol. The formulas are given below:

$$\text{Accuracy} = \frac{1}{|\mathcal{D}_{\text{dis}}|} \sum_{i \in \mathcal{D}_{\text{dis}}} \mathbf{1}[\hat{y}_i = y_i]$$

$$\text{Bias} = \frac{\sum_{i \in \mathcal{D}_{\text{und}}} b_i \, \mathbf{1}[\hat{y}_i \neq 2]}{\sum_{i \in \mathcal{D}_{\text{und}}} \mathbf{1}[\hat{y}_i \neq 2]}$$

Here, $\mathcal{D}_{\text{dis}}$ denotes the set of disambiguated items and $\mathcal{D}_{\text{und}}$ denotes the set of ambiguous items. y_i is the ground-truth label, b_i is the bias indicator (i.e., which choice reflects a stereotyped response), and $\hat{y}_i$ is the model's predicted choice index. These definitions follow the official BBQ benchmark metrics, allowing direct comparability with prior studies (Parrish et al., 2022).

A.4 Data Validation Process

The speaker metadata validation step was conducted under Python 3.10.16 using a single NVIDIA A6000 ADA GPU. The setup followed the configuration guidelines provided on the Hugging Face model pages[4][5].

We employed two pretrained audio classification models based on different architectures. The first, `wav2vec2-large-robust-24-ft-age-gender`, is a wav2vec 2.0–based model fine-tuned for

[3] https://github.com/hexgrad/kokoro

[4] https://huggingface.co/audeering/wav2vec2-large-robust-24-ft-age-gender

[5] https://huggingface.co/Jzuluaga/accent-id-commonaccent_ecapa

Ground-Truth	Total Samples	Accuracy (%)
Gender: Female	467,936	100.00
Gender: Male	467,936	93.16
Region: GB	467,936	99.49
Region: US	467,936	90.15

Table 3: Prediction accuracy by ground-truth category (GB and US denote Great Britain and United States).

Comparison model	Context	r	p-value
LLaMA-based	Ambiguous	0.620	0.042
LLaMA-based	Disambiguated	0.301	0.369
Qwen-based	Ambiguous	0.844	p < 0.001
Qwen-based	Disambiguated	0.848	p < 0.001

Table 4: Pearson correlation analysis for LLaMA family and Qwen family for ambiguous and disambiguated context.

age and gender classification after pretraining on large-scale datasets such as VoxCeleb. According to its official documentation, it achieves over 80% balanced accuracy on gender classification tasks. The second, `accent-id-commonaccent_ecapa`, is built upon the ECAPA-TDNN architecture and was trained on the CommonVoice dataset to identify English regional accents, reporting classification accuracy above 90%.

These models were chosen due to their verified performance on downstream tasks relevant to our study—namely, speaker gender and accent (region) classification—which made them suitable for validating the integrity of the synthesized speech data. Each model was used to infer gender or region from the input audio waveform under `eval()` mode with batch size 1.

The major libraries used for inference were: `torch 2.5.1+cu124`, `transformers 4.51.3`, `numpy 1.26.4`, and `pandas 2.2.3`.

B Experimental Environment

In this section provides a concise overview of the hardware configurations, software setups, and library dependencies used in our Qwen2-Audio and LLaMA-Omni experiments.

B.1 Qwen2-Audio

All Qwen2-Audio experiments were conducted on a single NVIDIA A6000 GPU under Python 3.9.21. The environment was configured following the model's Hugging Face page[6]. Inference was performed using the Hugging Face `AutoProcessor` and `Qwen2AudioForConditionalGeneration`, jointly processing text and audio inputs and generating outputs via: `model.generate(max_length=1024)` To ensure comparability across runs, the maximum token length was fixed at 1024, and all experiments were executed with batch size 1 in evaluation mode `eval()`. Major dependencies included `torch 2.2.1+cu118`, `transformers 4.52.0`, `numpy 2.2.5`, `pandas 2.2.3`, and `soundfile 0.13.1`.

B.2 LLaMA-Omni

LLaMA-Omni experiments were carried out on a single A6000 GPU under Python 3.10.17. The setup was based on OmniMMI's OpenOmniNexus framework[7] and the official LLaMA-Omni repository[8]. Model and tokenizer were loaded using : `load_pretrained_model(model_path, None, s2s=False)` Inference was performed with batch size 1 in evaluation mode, using `do_sample=False`, `num_beams=1`, `top_p=None`, and `max_new_tokens=1024`. For fairness, the maximum token count was fixed at 1024. Major dependencies included `torch 2.5.0+cu118`, `transformers 4.44.0`, `flash-attn 2.6.3`, `fairseq 0.12.2`, `deepspeed 0.14.5`, and `numpy 1.26.4`.

B.3 Backbone LLM Model Setting

The BBQ benchmark evaluation of the backbone LLMs was performed under Python 3.10.16 using a single NVIDIA A6000 ADA GPU. The experimental setup followed the official Hugging Face configurations for each model: `Qwen/Qwen-7B`[9] and `meta-llama/Llama-3.1-8B-Instruct`[10].

Inference was conducted using `AutoTokenizer` and `AutoModelForCausalLM`. Text inputs were passed to the models and generation was performed using model.generate(max_new_tokens=1024). To ensure consistency across runs, the maximum number of tokens was fixed at 1024. All inference runs were executed in evaluation mode (`eval()`) with batch size set to 1.

The main libraries used for this process included: `torch 2.5.1+cu124`, `transformers 4.51.3`, `numpy 1.26.4`, and `pandas 2.2.3`.

[6] https://huggingface.co/Qwen/Qwen2-Audio-7B-Instruct

[7] https://github.com/OmniMMI/OpenOmniNexus)

[8] https://github.com/ictnlp/LLaMA-Omni

[9] https://huggingface.co/Qwen/Qwen-7B

[10] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

	LLaMA3.1				Qwen1				LLaMA Omni				Qwen2 Audio			
Category	acc_A	acc_D	s_A	s_D	acc_A	acc_D	s_A	s_D	acc_A	acc_D	s_A	s_D	acc_A	acc_D	s_A	s_D
Age	0.258	0.793	-0.242	-0.326	0.131	0.852	-0.266	-0.306	0.655	0.536	-0.065	-0.187	0.494	0.803	-0.132	-0.260
Disability status	0.337	0.867	-0.21	-0.317	0.275	0.942	-0.247	-0.341	0.646	0.644	-0.076	-0.216	0.423	0.899	-0.169	-0.293
Gender Identity	0.17	0.634	-0.255	-0.308	0.081	0.823	-0.307	-0.334	0.251	0.558	-0.061	-0.081	0.133	0.824	-0.241	-0.278
Nationality	0.352	0.909	-0.197	-0.303	0.286	0.903	-0.244	-0.342	0.735	0.554	-0.065	-0.246	0.428	0.758	-0.168	-0.294
Physical Appearance	0.271	0.738	-0.177	-0.243	0.299	0.782	-0.15	-0.214	0.744	0.545	-0.033	-0.128	0.653	0.718	-0.074	-0.213
Race/Ethnicity	0.322	0.722	-0.196	-0.289	0.096	0.793	-0.252	-0.279	0.684	0.534	-0.089	-0.281	0.321	0.812	-0.197	-0.291
Race x SES	0.448	0.828	-0.17	-0.309	0.434	0.993	-0.206	-0.363	0.918	0.622	-0.025	-0.305	0.619	0.946	-0.135	-0.354
Race x gender	0.271	0.743	-0.254	-0.348	0.1	0.844	-0.297	-0.33	0.642	0.655	-0.110	-0.306	0.331	0.874	-0.206	-0.308
Religion	0.287	0.837	-0.202	-0.283	0.314	0.860	-0.190	-0.277	0.776	0.553	-0.025	-0.112	0.636	0.927	-0.087	-0.240
SES	0.207	0.871	-0.263	-0.331	0.102	0.952	-0.283	-0.315	0.368	0.554	-0.075	-0.119	0.438	0.912	-0.155	-0.276
Sexual orientation	0.415	0.826	-0.195	-0.333	0.288	0.832	-0.225	-0.315	0.839	0.584	-0.029	-0.181	0.676	0.854	-0.091	-0.281

Table 5: Accuracy and Bias for disambiguated (acc_D, s_D) vs. ambiguous (acc_A, s_A) items. a_D, a_A denote accuracy, and s_D, s_A denote bias score.

	LLaMA-Omni								**Qwen2-Audio**							
	Female				Male				Female				Male			
	acc_D	acc_A	s_D	s_A	acc_D	acc_A	s_D	s_A	acc_D	acc_A	s_D	s_A	acc_D	acc_A	s_D	s_A
Age	0.537	0.647	-0.190	-0.067	0.534	0.663	-0.184	-0.062	0.809	0.493	-0.262	-0.133	0.797	0.495	-0.258	-0.130
Disability_status	0.647	0.636	-0.217	-0.079	0.641	0.657	-0.214	-0.074	0.896	0.418	-0.289	-0.168	0.901	0.427	-0.298	-0.171
Gender_identity	0.564	0.254	-0.101	-0.075	0.552	0.249	-0.061	-0.046	0.823	0.133	-0.278	-0.241	0.824	0.133	-0.277	-0.240
Nationality	0.553	0.721	-0.259	-0.072	0.555	0.749	-0.231	-0.058	0.757	0.425	-0.296	-0.170	0.759	0.431	-0.292	-0.166
Physical_appearance	0.556	0.733	-0.126	-0.034	0.534	0.756	-0.130	-0.032	0.724	0.653	-0.209	-0.073	0.711	0.653	-0.216	-0.075
Race_ethnicity	0.539	0.680	-0.291	-0.093	0.529	0.688	-0.271	-0.084	0.810	0.323	-0.288	-0.195	0.814	0.319	-0.293	-0.199
Race_x_SES	0.632	0.839	-0.308	-0.050	0.376	0.997	-0.235	-0.001	0.945	0.616	-0.352	-0.135	0.946	0.622	-0.357	-0.135
Race_x_gender	0.652	0.636	-0.311	-0.113	0.657	0.647	-0.300	-0.106	0.875	0.329	-0.307	-0.206	0.873	0.333	-0.309	-0.206
Religion	0.557	0.769	-0.137	-0.032	0.549	0.783	-0.087	-0.019	0.926	0.639	-0.237	-0.086	0.928	0.633	-0.243	-0.089
SES	0.564	0.360	-0.145	-0.093	0.544	0.376	-0.092	-0.057	0.909	0.439	-0.276	-0.155	0.915	0.437	-0.276	-0.155
Sexual_orientation	0.585	0.833	-0.176	-0.029	0.584	0.845	-0.185	-0.029	0.858	0.676	-0.297	-0.096	0.851	0.676	-0.264	-0.085

Table 6: Accuracy and Bias-score by speaker **Gender** for disambiguated (acc_D, s_D) vs. ambiguous (acc_A, acc_A) items. acc_D, acc_A denote accuracy, and s_D, s_A denote bias score.

	LLaMA-Omni								**Qwen2-Audio**							
	US				UK				US				UK			
	acc_D	acc_A	s_D	s_A	acc_D	acc_A	s_D	s_A	acc_D	acc_A	s_D	s_A	acc_D	acc_A	s_D	s_A
Age	0.539	0.652	-0.181	-0.062	0.532	0.657	-0.193	-0.066	0.807	0.479	-0.258	-0.134	0.799	0.509	-0.262	-0.129
Disability_status	0.635	0.673	-0.211	-0.069	0.653	0.620	-0.220	-0.084	0.894	0.420	-0.289	-0.168	0.903	0.426	-0.298	-0.171
Gender_identity	0.556	0.254	-0.073	-0.055	0.561	0.249	-0.088	-0.066	0.826	0.133	-0.281	-0.244	0.822	0.134	-0.274	-0.238
Nationality	0.553	0.747	-0.250	-0.063	0.556	0.722	-0.242	-0.067	0.751	0.419	-0.301	-0.175	0.765	0.437	-0.287	-0.162
Physical_appearance	0.547	0.748	-0.154	-0.039	0.544	0.741	-0.102	-0.027	0.724	0.647	-0.226	-0.080	0.712	0.659	-0.199	-0.068
Race/Ethnicity	0.534	0.691	-0.275	-0.085	0.535	0.678	-0.287	-0.093	0.812	0.311	-0.287	-0.198	0.813	0.331	-0.294	-0.197
Race_x_SES	0.389	0.997	-0.291	-0.001	0.634	0.839	-0.306	-0.049	0.947	0.624	-0.358	-0.134	0.944	0.613	-0.351	-0.136
Race_x_gender	0.652	0.645	-0.307	-0.109	0.657	0.639	-0.305	-0.110	0.874	0.334	-0.311	-0.207	0.874	0.329	-0.305	-0.205
Religion	0.560	0.795	-0.097	-0.020	0.546	0.757	-0.127	-0.031	0.926	0.625	-0.240	-0.090	0.929	0.647	-0.240	-0.085
SES	0.558	0.362	-0.124	-0.079	0.550	0.373	-0.113	-0.071	0.913	0.437	-0.278	-0.157	0.911	0.439	-0.274	-0.154
Sexual_orientation	0.580	0.850	-0.197	-0.030	0.588	0.828	-0.163	-0.028	0.861	0.677	-0.295	-0.095	0.848	0.676	-0.266	-0.086

Table 7: Accuracy and Bias-score by speaker **Accent** for disambiguated (acc_D, s_D) vs. ambiguous (acc_A, s_A) items. acc_D, acc_A denote accuracy, and s_D, s_A denote bias score.

C Detail Result

Tables from 5 to 7 shows the detailed result of Bias Socre, and table 4 show the detailed result of correlation result and std value.

Category	s_D **Std**	**acc**$_D$ **Std**	**acc**$_A$ **Std**	s_A **Std**
Age	0.021	0.017	0.035	0.015
Disability_status	0.016	0.010	0.034	0.010
Gender_identity	0.009	0.008	0.011	0.008
Nationality	0.016	0.014	0.025	0.014
Physical_appearance	0.027	0.020	0.030	0.013
Race_ethnicity	0.009	0.009	0.022	0.010
Race_x_SES	0.009	0.008	0.021	0.008
Race_x_gender	0.005	0.005	0.031	0.011
Religion	0.014	0.008	0.029	0.009
SES	0.009	0.008	0.019	0.008
Sexual_orientation	0.039	0.021	0.031	0.018

Table 8: Standard Deviation of Qwen Audio, Global

Category	s_D **Std**		**acc**$_D$ **Std**		**acc**$_A$ **Std**		s_A **Std**	
	female	male	female	male	female	male	female	male
Age	0.027	0.015	0.020	0.013	0.040	0.033	0.019	0.011
Disability_status	0.017	0.014	0.008	0.010	0.034	0.035	0.011	0.009
Gender_identity	0.009	0.008	0.010	0.007	0.013	0.010	0.009	0.008
Nationality	0.016	0.017	0.013	0.015	0.024	0.028	0.015	0.014
Physical_appearance	0.036	0.018	0.017	0.021	0.028	0.035	0.016	0.009
Race_ethnicity	0.009	0.009	0.011	0.006	0.025	0.020	0.010	0.010
Race_x_SES	0.009	0.009	0.004	0.011	0.013	0.027	0.007	0.010
Race_x_gender	0.006	0.004	0.006	0.004	0.034	0.030	0.013	0.010
Religion	0.013	0.014	0.008	0.008	0.027	0.032	0.007	0.010
SES	0.011	0.007	0.007	0.008	0.019	0.021	0.007	0.008
Sexual_orientation	0.045	0.026	0.021	0.022	0.026	0.036	0.021	0.013

Table 9: Standard Deviation of Qwen Audio, Gender

Category	s_D **Std**		**acc**$_D$ **Std**		**acc**$_A$ **Std**		s_A **Std**	
	GB	US	GB	US	GB	US	GB	US
Age	0.019	0.024	0.013	0.021	0.037	0.027	0.015	0.016
Disability_status	0.018	0.013	0.008	0.009	0.038	0.031	0.010	0.009
Gender_identity	0.006	0.010	0.009	0.008	0.010	0.013	0.007	0.009
Nationality	0.015	0.014	0.014	0.009	0.022	0.027	0.014	0.012
Physical_appearance	0.018	0.029	0.017	0.022	0.035	0.026	0.011	0.012
Race_ethnicity	0.008	0.009	0.010	0.008	0.021	0.019	0.011	0.010
Race_x_SES	0.009	0.009	0.006	0.010	0.020	0.022	0.009	0.008
Race_x_gender	0.003	0.004	0.004	0.007	0.031	0.033	0.011	0.012
Religion	0.012	0.016	0.008	0.008	0.025	0.030	0.008	0.010
SES	0.010	0.008	0.006	0.010	0.020	0.021	0.005	0.009
Sexual_orientation	0.028	0.045	0.025	0.015	0.035	0.028	0.014	0.020

Table 10: Standard Deviation of Qwen Audio, Accent

Category	s_D **Std**	**acc**$_D$ **Std**	**acc**$_A$ **Std**	s_A **Std**
Age	0.022	0.009	0.015	0.009
Disability_status	0.037	0.022	0.045	0.014
Gender_identity	0.026	0.010	0.012	0.020
Nationality	0.036	0.016	0.029	0.013
Physical_appearance	0.051	0.022	0.015	0.013
Race_ethnicity	0.017	0.007	0.018	0.008
Race_x_SES	0.134	0.136	0.141	0.044
Race_x_gender	0.011	0.005	0.009	0.006
Religion	0.062	0.033	0.028	0.016
SES	0.046	0.017	0.021	0.031
Sexual_orientation	0.045	0.023	0.022	0.008

Table 11: Standard Deviation of LLaMA-Omni, Global

Category	s_D **Std**		**acc**$_D$ **Std**		**acc**$_A$**Std**		s_A **Std**	
	female	male	female	male	female	male	female	male
Age	0.024	0.022	0.010	0.009	0.016	0.011	0.009	0.009
Disability_status	0.032	0.043	0.028	0.016	0.059	0.024	0.012	0.017
Gender_identity	0.015	0.018	0.009	0.008	0.010	0.014	0.012	0.013
Nationality	0.037	0.031	0.016	0.018	0.030	0.021	0.012	0.011
Physical_appearance	0.049	0.057	0.024	0.014	0.009	0.011	0.013	0.013
Race_ethnicity	0.015	0.011	0.006	0.003	0.020	0.017	0.007	0.007
Race_x_SES	0.104	0.160	0.128	0.098	0.169	0.001	0.053	0.000
Race_x_gender	0.009	0.010	0.004	0.005	0.006	0.008	0.004	0.005
Religion	0.059	0.059	0.036	0.030	0.030	0.026	0.017	0.014
SES	0.031	0.045	0.017	0.011	0.013	0.025	0.021	0.029
Sexual_orientation	0.046	0.047	0.025	0.022	0.029	0.011	0.010	0.007

Table 12: Standard Deviation of LLaMA Omni, Gender

Category	s_D **Std**		**acc**$_D$ **Std**		**acc**$_A$ **Std**		s_A **Std**	
	gb	us	gb	us	gb	us	gb	us
Age	0.023	0.022	0.010	0.008	0.007	0.021	0.009	0.009
Disability_status	0.035	0.040	0.019	0.022	0.050	0.014	0.012	0.014
Gender_identity	0.025	0.027	0.007	0.012	0.010	0.013	0.019	0.020
Nationality	0.029	0.044	0.015	0.018	0.028	0.026	0.011	0.015
Physical_appearance	0.054	0.035	0.015	0.029	0.015	0.016	0.014	0.009
Race_ethnicity	0.014	0.018	0.007	0.007	0.020	0.015	0.008	0.007
Race_x_SES	0.107	0.160	0.177	0.034	0.169	0.001	0.053	0.000
Race_x_gender	0.011	0.011	0.006	0.004	0.007	0.009	0.005	0.007
Religion	0.070	0.053	0.031	0.034	0.022	0.019	0.019	0.012
SES	0.039	0.055	0.011	0.022	0.025	0.015	0.026	0.036
Sexual_orientation	0.049	0.036	0.020	0.026	0.024	0.014	0.011	0.005

Table 13: Standard Deviation of LLaMA Omni, Accent