Mind the Gap:
Static and Interactive Evaluations of Large Audio Models

Minzhi Li[*νσ§] **William Held**[*γσ] **Michael J. Ryan**[σ]
Kunat Pipatanakul[ω] **Potsawee Manakul**[ω] **Hao Zhu**[σ] **Diyi Yang**[σ]
[γ]Georgia Institute of Technology [ν]National University of Singapore
[ω]SCB 10X, SCBX Group [§]Institute for Infocomm Research (I^{2}R), A*STAR
[σ]Stanford University

Abstract

As AI chatbots become ubiquitous, voice interaction presents a compelling way to enable rapid, high-bandwidth communication for both semantic and social signals. This has driven research into Large Audio Models (LAMs) to power voice-native experiences. However, aligning LAM development with user goals requires a clear understanding of user needs and preferences to establish reliable progress metrics. This study addresses these challenges by introducing an interactive approach to evaluate LAMs and collecting 7,500 LAM interactions from 484 participants. Through topic modeling of user queries, we identify primary use cases for audio interfaces. We then analyze user preference rankings and qualitative feedback to determine which models best align with user needs. Finally, we evaluate how static benchmarks predict interactive performance - our analysis reveals no individual benchmark strongly correlates with interactive results ($\tau \leq 0.33$ for all benchmarks). While combining multiple coarse-grained features yields modest predictive power (R^2=0.30), only two out of twenty datasets on spoken question answering and age prediction show significantly positive correlations. This suggests a clear need to develop LAM evaluations that better correlate with user preferences.

1 Introduction

Compared to text, speech enables faster, more efficient interaction (Ruan et al., 2016) and further enables communication of paralinguistic information (Sutton et al., 2019). These dual motivations make speech interaction a promising step towards more ubiquitous computing (Wei and Landay, 2018). Following this vision, researchers have developed large language models (LLMs) that directly accept audio inputs (Latif et al., 2023; Tang et al., 2023; Chu et al., 2024, 2023; Hurst et al., 2024; Held et al., 2024), known as Large Audio Models (LAMs) (Latif et al., 2023).

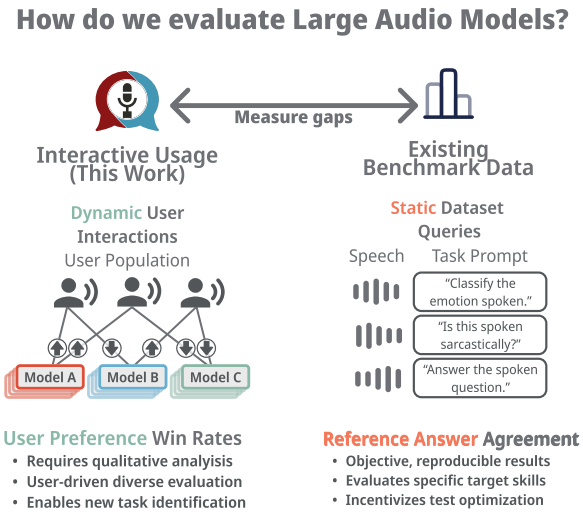

Figure 1: Comparison of static and interactive ways of evaluating Large Audio Models. In this work, we perform interactive evaluations to understand how LAMs are likely to be used and how they can be benchmarked.

Recent work has focused on evaluating these systems by aggregating existing static benchmarks (Yang et al., 2024; Wang et al., 2024). Similar to traditional LLM evaluation approaches (Brown et al., 2020), these frameworks examine LAMs in zero-shot and few-shot settings using tasks originally designed for finetuned models. The evaluations cover both speech interaction capabilities where natural text counterparts exist (Wu et al., 2023; Wang et al., 2020a) and paralinguistic feature recognition (Busso et al., 2008; Hasan et al., 2019), reflecting the dual motivations driving LAM development.

While static benchmarks are ideal for measuring progress on specific capabilities, their value diminishes when gaps exist between measured capabilities and user expectations (Lee et al., 2022). For this reason, benchmarks for text-only LLMs typi-

[*]Equal Contribution. Work completed at Stanford University. Contacts: li.minzhi@u.nus.edu, held@stanford.edu, diyiy@stanford.edu.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8749–8766

cally aim to correlate with real-world user preferences captured through interactive evaluations like Chatbot Arena (Chiang et al., 2024). Indeed, popular text-only benchmarks such as ARC-C (Clark et al., 2018), MMLU (Hendrycks et al., 2021), and AlpacaEval (Dubois et al., 2024b) all show high correlation ($\rho > 0.8$) with interactive user evaluation results (Dubois et al., 2024a).

However, it remains unclear whether existing LAM benchmarks correlate similarly with user evaluation, as no prior work has collected LAM preference data. This knowledge gap leaves LAM researchers without clear direction for model development that better aligns with user needs. In light of this, we collect 7,500 interactions between 484 paid participants (all with prior commercial LLM chat experience) and 6 speech-in text-out systems, including 5 top-performing LAMs and a baseline system combining ASR with a text-only LLM. Compared to static evaluation, it leads to a more user-driven diverse evaluation, and enables new task identification (see Figure 1). This preference data allows us to address questions beyond the scope of existing benchmark-based or text-only interactive evaluations:

- **What tasks do users expect LAMs to perform?** Our interface provides users with a simple audio interface and prompts them to test capabilities they would expect from an AI voice assistant. Following Tamkin et al. (2024), we use LLMs to cluster and categorize these queries. We find that 77% of usage falls into categories where speech primarily serves efficiency purposes (e.g., task execution) rather than conveying information unique to the audio modality.
- **What models are best at these tasks and why?** Users receive text responses from randomly sampled systems and provide pairwise preferences with qualitative feedback. Surprisingly, we find that a pipeline system that combines Whisper (Radford et al., 2022) and Llama (AI@Meta, 2024) is preferred the most. This likely stems from two factors: most user queries rely primarily on text semantics, and three of the five most common feedback types focus on style of the textual output.
- **Which benchmarks are the best proxies for user preferences?** Our evaluation of LAMs across twenty existing speech benchmarks reveals limited predictive power for user preferences. No single benchmark shows strong correlation with human evaluations, and even aggregated benchmarks explain only 30% of preference variance ($R^2 = 0.30$). Only two metrics show positive correlation: speech comprehension ability (measured via Public-SG-Speech) and reduced systematic errors (captured by an age prediction task where all evaluated models perform worse than random chance). This starkly contrasts text-only LLMs, where static evaluations and interactive assessments show high correlation, suggesting the need for new static audio evaluations, which our interaction data can inform.

2 Related Work

Large Audio Models. Large-scale self-supervised audio models have been used to learn generalized audio representations from extensive unlabeled datasets. Early successful approaches such as wav2vec (Schneider et al., 2019) and HuBERT (Hsu et al., 2021) learned audio representations from scratch, achieving robust performance across many tasks when finetuned. Focused primarily on scaling data and training time, recent efforts such as Whisper (Radford et al., 2022) and OWSM (Peng et al., 2023) have led to extremely effective models both for transcription and speech understanding.

Recent advancements in audio models have integrated learned audio representations with text-based LLMs, enabling native audio understanding while leveraging knowledge and stylistic insights from textual resources. This has led to the emergence of Large Audio Models (LAMs) (Latif et al., 2023). Such models include SpeechGPT (Zhang et al., 2023) which leverages HuBERT (Hsu et al., 2021) for extracting continuous speech as discrete units, LLaMA (Touvron et al., 2023) as the text-LLM foundation, and HiFi-GAN (Kong et al., 2020) as the unit vocoder; LTU (Gong et al., 2023) which consists of an audio spectrogram transformer, LLaMA and a Low-rank Adapter; Qwen-Audio series (Chu et al., 2023, 2024) with Whisper-large-v2 and Whisper-large-v3 as the audio encoder and Qwen-7B as the LLM, and many other Large Audio Models (Borsos et al., 2023; Liu et al., 2023; Held et al., 2024). In our work, we evaluated nine different LAMs that are publicly available on static benchmarks and tested five best-performing ones in the interactive setting.

Evaluation of Large Audio Models. To evaluate the audio processing capability of different models, prior research has constructed a variety of audio benchmarks, targeting particular abilities. For automatic speech recognition, benchmarks such as Librispeech (Panayotov et al., 2015) and Commonvoice (Ardila et al., 2019) are widely used, with metrics like word error rate (WER) and character error rate (CER). For speech translation tasks, there are datasets like Covost (Wang et al., 2020a), Covost2 (Wang et al., 2021), and CVSS (Jia et al., 2022) with evaluation metrics such as BLEU scores. For emotion detection, benchmarks include MELD (Poria et al., 2018) and IEMOCAP (Busso et al., 2008) with speech data labeled with different emotions. In the domain of Speech Question Answering, there are SDQA (Faisal et al., 2021), Social IQ 2.0 (Wilf et al., 2023), and HeySquad (Wu et al., 2023).

However, one problem regarding the evaluation of LAMs is that they have reported evaluation results on different sets of benchmarks, resulting in inconsistent evaluation and difficulty in comparison (Wang et al., 2024). Therefore, there are commendable efforts to aggregate audio datasets together to evaluate LAMs in a holistic way such as AIRBench (Yang et al., 2024), AudioBench (Wang et al., 2024), and VoiceBench (Chen et al., 2024). However, they still utilize static reference-based metrics like WER and accuracy. In contrast, we interactively evaluate LAMs using user preferences.

Interactive Evaluation of LLMs Interactive evaluation can overcome many limitations in using static datasets to evaluate models. One limitation is model overfitting (Ying, 2019) where models are over-optimized for specific datasets and tasks, limiting their generalization capability. Moreover, static benchmarks may have data contamination (Magar and Schwartz, 2022) issues where LLMs have been trained on the data. Furthermore, static evaluation may lack the ability to incorporate real-world scenarios (Lin et al., 2024) and align with human preferences (Oren et al., 2023). Moreover, data drift (Mallick et al., 2022) can happen when the environment generating the data evolves, causing a mismatch between the static datasets and the data in real-world scenarios. Thus, static datasets can fail to keep track of long-term model performance over time. These limitations strongly suggest the need for interactive evaluation of models.

As such, there are many research efforts on creating live NLP benchmarks. For example, DynaBench (Kiela et al., 2021) builds an open platform for dynamic data curation, and Chatbot Arena (Chiang et al., 2024) benchmarks models through chat with LLMs from a larger user base. In a similar line, there are works extending to other modalities and use cases like Wildvision-Arena (Lu et al., 2024) for vision-language models, Long Code Arena (Bogomolov et al., 2024) for coding, and Web-Arena (Zhou et al., 2023) for web-related tasks. To our knowledge, there is no similar interactive evaluation of audio-language models to investigate the gap between static benchmarks and user interactions.

3 Interactive Evaluation

To capture real-world use cases of LAMs, we collect user preferences on an open platform[1]. We then convert pairwise votes to model ranks that reflect the interaction capability of different models.

3.1 Interface

Our interface (see Figure A.2 in Appendix) is built using the Gradio platform (Abid et al., 2019). This allows us to serve a simple web-based user experience with integration to both locally hosted and API-accessible LAMs.

User Interaction and Input Upon arriving on the platform, users are instructed to interact with the model for any use cases they would expect from a voice-based AI assistant. By providing no concrete example tasks, our goal was to capture diverse desired use cases with minimal bias based on our preconceptions of "interesting" or "challenging" use cases. Each query is streamed to the user character by character to avoid users being able to learn a mapping between tokenizations and specific model identities.

Pairwise Model Comparison. After submitting a query, users receive responses from two anonymous models, which are randomly selected and ordered in order to avoid personal and positional bias in their preferences. For assessment, users provide a simple **pairwise preference ranking**—choosing the better response or indicating no preference between the two. This method provides a relative ranking rather than an absolute performance score, allowing the user to make a simple decision and avoiding performance ceilings, which may be induced by reference-driven evaluation.

[1] https://talkarena.org/

User Feedback and Justification. One shortcoming of preference ratings is that they offer minimal insights into the factors that drive user preferences. While this allows for very open-ended user values to be integrated into final model ratings, it also makes the data less directly valuable for deriving insights about what needs improvement in existing models. To gain deeper insights, users can optionally justify their choices via text or speech, following prior work showing that users provide dramatically more detail when given a speech interface (Deitke et al., 2024). Even without requiring the completion of this field, we find that 44.9% of users opted to provide qualitative rationale.

3.2 Data Collection

This research has been approved by the Institutional Review Board (IRB) at the authors' institution. We selected Prolific as the survey platform due to its large participant pool and high average data quality (Eyal et al., 2021; Douglas et al., 2023). Since the average crowd worker is likely not a user of AI voice chat products, we further used the platform's pre-screening to select participants who reported actively having used LLM chatbots such as ChatGPT and Gemini previously. The only other limiting requirement was that participants had access to a working microphone to record and submit their voice.

For each pair of models we evaluated, we recruited 50 participants, and each participant can contribute 10 votes. This allows us to obtain a sufficient pool of votes that can illustrate user preference between model pairs in a statistically significant manner. In total, 7500 votes were collected from a diverse pool of around 484 unique participants. We also apply the selection criteria to ensure the pool of participants is gender-balanced to allow a fair representation of user preference. Each user was paid $2.50 for 10 votes, with a minimum of $15 per hour ensured.

3.3 Model Rank

To convert the collected pairwise preference data to model ranking, we apply the Bradley-Terry model (Bradley and Terry, 1952) to compute scores for each model for its statistical rigor and good interpretability. The Bradley-Terry model provides a principled way to infer latent winning ability to estimate the probability of one entity winning over another. The model defines an exponential score function p_i as e^{β_i} for model i where β is the Bradley-Terry coefficients. For a model pair of model i and model j, the probability of model i being preferred over model j is given by Equation (1):

$$\Pr(i > j) = \frac{e^{\beta_i}}{e^{\beta_i} + e^{\beta_j}} \quad (1)$$

The Bradley-Terry coefficients β are computed by maximizing the log-likelihood of the observed pairwise preferences $\mathcal{D}$, given by

$$\mathcal{L}(\beta) = \sum_{(i,j,y)\in\mathcal{D}} [y \log(\sigma(\beta_i - \beta_j)) + (1-y)\log(\sigma(\beta_j - \beta_i))] \quad (2)$$

where y is 1 when i is preferred over j, 0.5 for ties, and 0 when j is preferred; σ is the sigmoid function. The optimization is performed using the L-BFGS algorithm[2]. We then compute $p_i = e^{\beta_i}$.

With quality scores $\mathbf{p}$ estimated, the models can be ranked in a descending order based on the scores. A higher score indicates a higher likelihood of being preferred by users.

4 What Tasks Do Users Expect LAMs to Perform?

We adopt the Clio analysis flow (Tamkin et al., 2024) on transcripted user queries to explore topics in users' queries. We first apply the K-Means clustering algorithm on BERT embeddings of summaries of 1000 randomly sampled queries and identify **task execution**, **knowledge expansion**, **chat**, and **advice seeking** as four initial clusters through the Elbow Method (Bholowalia and Kumar, 2014) and merging similar clusters after manual inspection. We then discover hierarchical clusters through recursive application of clustering algorithm as shown in Figure 2.

To better understand the topic distribution, we manually listen to 100 randomly sampled recordings of user queries and classify them into one of the four main topics discovered. We found most users ask about knowledge-related questions (50%) (e.g. *"What is galaxy?"*), followed by advice seeking (17%) (e.g. *"I'm thinking of getting some brine shrimp. What should I know before I get them?"*), chat (16%) (e.g. *"Good morning, how are you?"*), and task execution (10%) (e.g. *"Summarize Volume 1 of Lord of the Mysteries."*). These dominant uses suggest about areas LAMs could work

[2]We follow the updated methodology used by Chatbot Arena without scaling or normalization.

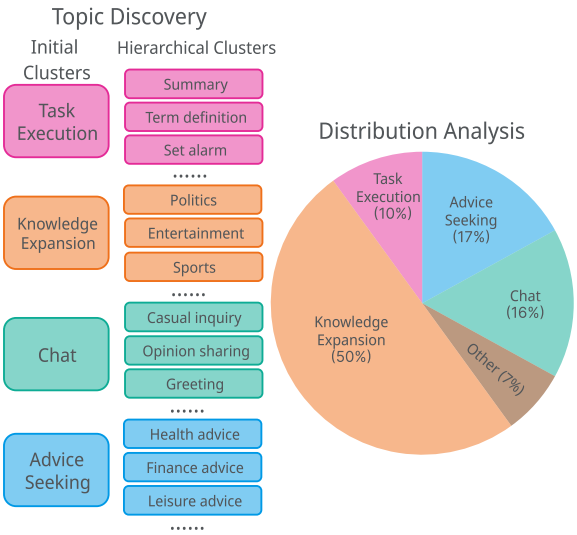

Figure 2: We identify four main topics in user queries — task execution, knowledge expansion, chat, advice seeking as well as sub-topics under each category through hierarchical clustering (left) and analyze the relative proportions of each query type (right).

towards to better user experience: it is important for LAMs to be equipped with up-to-date and comprehensive factual knowledge, potentially through methods like retrieval augmentation, to address the key use of knowledge query. Moreover, emotional and contextual understanding is crucial in advice seeking and chat situations. Furthermore, accurate intent detection is needed for task execution tasks. Moreover, 7% of the sampled recordings include background noises such as music and voices of other people who is not the speaker. This introduces an additional factor beyond textual content to be considered in improving LAMs' interactive capability during real-world interactions.

Compared to top user queries during interactions with text-LLMs (Zheng et al., 2023a), participants rarely talk about mathematical concepts and coding problems. This is expected as math and coding require precise syntax, symbols, and formatting, which can be difficult to dictate naturally. This suggests that when mapping text queries to audio inputs, we should consider distributional differences between written and spoken language for data curation.

5 What Models are Best at these Tasks and Why?

User Preferences In Figure 3, although close-sourced models like GPT4o demonstrates the best performance on static benchmarks (Table 1), it is not the most preferred model in voice-in text-out interactions. Instead, the ASR pipeline of whisper-v2.0 and Llama3-8B-Instruct is most preferred among all six model settings, followed by DiVA which is trained by distilling a text LLM. This is because most of the user queries do not rely on nuanced speech understanding (Figure 2) as we place minimal constraints on users' queries without asking them to submit challenging samples. It also suggests that the most effective way to improve current models' interactive capability for general single-turn use cases is to **leverage a powerful text language model's interactive capabilities**.

Reasons for Preferences We manually analyze 100 randomly sampled user explanations for non-tie votes and summarize five most commonly observed reasons for users' preferences. 31% of user explanations mention about **(1) level of details** in the text responses during interactions. In general, users find a model more preferable if they can generate *specific, concrete and in-depth* responses (e.g. *"I think both models here were able to answer me a bit, but certainly model two answered me more in-depth, so we're going to go with that one."*). Another important factor in determining users' preferences is **(2) helpfulness** (24%) of the response. Users will not prefer a model which refuses to respond to their questions or fail to address their questions (e.g. *"Model 2 is completely irrelevant and refuses to answer the question I asked because it claims it is political, which it is not."*). On the other hand, they prefer models which understand their needs and provide *useful* feedback which *adheres to their instructions*.

Moreover, **(3) language appropriateness** (12%) of the response can affect users' interaction experience with the models. Some models generated responses in a different language from the user query, and users find such responses *illegible* and less preferred. This also observed through relatively similar model rank in interactive setting and model rank on the language detection benchmark. On top of that, users find **(4) accuracy** (11%) of responses key. They prefer models which provide numerically and factually correct answers (e.g. *"Model 1 is the factually accurate and most detailed and preferable response."*).

Furthermore, **(5) human-likeness** was mentioned in 11% of user explanations sampled. Interestingly, some prefer responses that are not human-

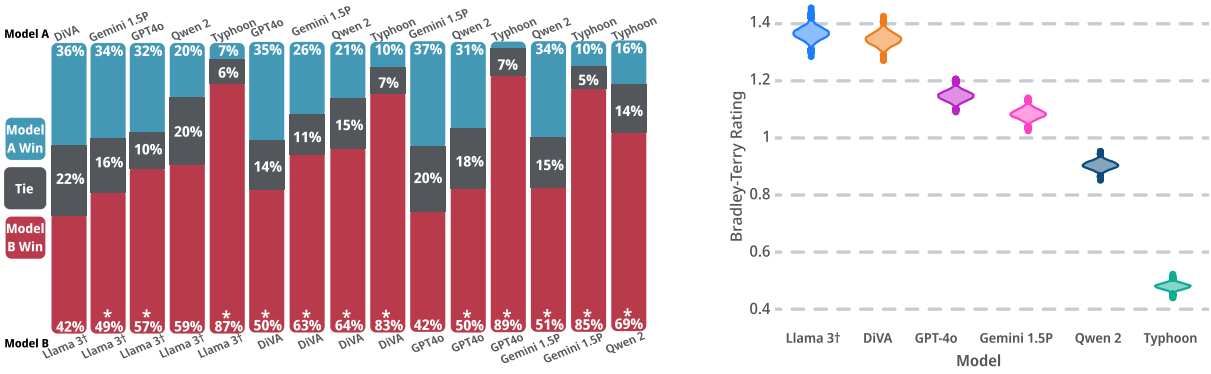

Figure 3: Head-to-head model comparisons (left) and Bradley-Terry (right) Scores from our evaluation. For win rates, * indicates the difference between preferences is significant (P<0.05) by a pairwise bootstrap test. For Bradley-Terry scores, distributions are shown shown for 10,000 bootstraps. † denotes an ASR + LLM pipeline.

like (e.g. *"I like the AI that admits it cannot have opinions."*) while others prefer the model which possesses human traits (e.g. *"Model1 was friendly and inquisitive"*). This shows the degree of human likeness is an important dimension that could affect interaction experience but needs adaptations to different users' preferences.

6 Which Benchmarks are the Best Proxies for User Preferences?

To understand the degree to which current speech benchmarks reflect the interaction capability of LAMs, we perform static evaluation on a comprehensive set of benchmarks that measure capabilities that may affect users' interaction experience. We run a logistic regression analysis to investigate the predictive power of each static dataset and obtain insights on directions future datasets should consider to align with real-world user experience.

6.1 Dataset Selection

We construct a superset of 20 datasets related to speech understanding and interactions from existing aggregated evaluation sets for LAMs (AudioBench (Wang et al., 2024) and AIR-Bench (Yang et al., 2024)). The datasets cover a wide range of tasks that evaluate models' knowledge of **Speaker Cognitive State**, **Speaker Identity**, and **Speech Content Understanding**. Our goal is to evaluate a comprehensive set of tasks that potentially influence user experience during interactions. This set can then be filtered to identify the most predictive tasks in an unopinionated fashion.

Speaker Cognitive State The ability to understand the cognitive states of speakers can be closely related to the interaction capability of models as effective interactions depend on accurate interpretation of intents and emotions (Tomasello, 2023; Jensen and Pedersen, 2016). For intent detection, we include datasets on pragmatic intent detection (Bastianelli et al., 2020), humorous intent detection (Hasan et al., 2019), and sarcastic intent detection (Castro et al., 2019). For emotion recognition, we include the IEMOCAP (Busso et al., 2008) and MELD (Poria et al., 2018) datasets.

Speaker Identity Understanding speakers' identity is crucial in context-dependent and personalized interactions (Wu and Cai, 2024). We include audio benchmarks with annotated fields on speaker identity (language, accent, demographic information, social relationship). We evaluate models on tasks like language identification (Yang et al., 2024), accent classification (Ardila et al., 2019), gender and age classification (Ardila et al., 2019; Veliche et al., 2024), as well as the relationship classification (Ardila et al., 2020).

Speech Content Understanding Another component fundamental to models' interaction capability is the ability to understand speech content (Greenberg, 1996). Besides traditional automatic speech recognition (ASR) tasks (Ardila et al., 2019; Panayotov et al., 2015), it also involves understanding the entities, events, and user instructions in the speech. We evaluate models' ability for speech grounding (Wang et al., 2024), speech entity recognition (Bastianelli et al., 2020), speech instruction following (Wang et al., 2024), and speech question answering (Wang et al., 2020b) tasks.

Model	Humor	Sarcasm	Intent	Emotion	Relation	Gender	Age	Accent	Grounding	Language	Entity	QA	Instruction	ASR
Commercial LAMs														
GPT4o	†44.6	†**53.6**	†89.2	†**29.1**	†**59.7**	13.6	†12.2	†**35.3**	†22.2	†**73.3**	35.8	†**65.2**	†64.0	0.19
Gemini	35.7	36.0	†**91.4**	†27.2	35.9	†43.9	7.9	†24.5	†**25.9**	†68.8	23.6	†64.2	59.2	†0.17
Open-Weights LAMs														
Qwen2-audio	34.9	†41.5	†81.1	23.2	17.3	†**69.1**	†**12.3**	5.4	10.0	†66.5	†**43.7**	†62.3	62.6	†0.16
Typhoon	†44.6	†48.8	45.3	21.5	†44.2	†55.3	5.0	7.9	†22.1	36.4	†38.5	52.6	†**68.3**	0.50
DiVA	†**46.2**	38.3	61.5	†25.2	34.9	30.5	10.4	13.0	17.3	46.5	18.8	50.5	†66.6	0.83
Qwen-audio	39.9	30.8	69.1	16.4	30.9	45.5	8.4	5.0	5.0	58.1	†38.7	60.3	45.6	†**0.07**
NExTGPT	26.6	16.9	12.7	8.6	27.4	24.1	8.5	6.8	8.7	26.4	12.2	37.9	6.4	2.37
PandaGPT	42.6	33.4	13.9	11.1	†44.2	42.5	†11.7	4.0	8.7	33.5	17.6	39.5	25.7	3.34
Baselines														
ASR Pipeline	37.8	32.8	64.8	24.0	22.8	31.4	9.7	†13.9	20.4	50.4	16.5	56.5	54.4	0.25
Random Baseline	50.0	50.0	25.0	25.0	25.0	50.0	14.3	20.0	25.0	25.0	25.0	-	-	-

Table 1: Average performance of LAMs and random baseline on 20 different benchmarks across 14 different speech understanding tasks. Top model performance is **bolded**, and top three model performances are marked with †.

6.2 Experiment Setup

Models We evaluate 9 different LAMs that are publicly available with coverage of both open and close sourced models. Due to budget, the models we tested in interactive evaluation are the six best-performing model settings (based on frequency of ranking as top five) which ensures a decent interactive capability for interactive evaluation: Qwen2-Audio (Chu et al., 2024), DiVA-8B (Held et al., 2024), Typhoon-1.5 (Pipatanakul et al., 2023), Gemini-1.5-pro (Team et al., 2024), GPT4o (Hurst et al., 2024), and ASR pipeline setting with whisper-large-v2 (Radford et al., 2022) and Llama3-8B (AI@Meta, 2024).

Metrics For classification tasks, we report macro F1 scores to account for the importance of different classes due to class imbalance in some datasets. We compute PEDANTS score (Li et al., 2024) for tasks requiring a short text response using the questions and reference answers. For ASR tasks, we report the Word Error Rate (WER).

Prompt Previous work (Wang et al., 2024) shows that some models like Qwen-Audio are prompt-sensitive. Therefore, we elicit models' responses using three different variations of text instruction prompts (see Appendix A). We take the average score of responses to different text prompts to get a more robust reflection of the models' capability.

6.3 Model Performance on Static Benchmarks

In general, close-sourced models generally top the leaderboard (Table 1): GPT4o has the highest frequency of ranking first among all tested models (6 out of 14) and emerges as one of the top three for most tasks (11 out of 14). Gemini-1.5-pro also ranks among the top three models on more than half of the tasks tested (8 out of 14). It demonstrates strong performance in tasks related to speaker identity such as classification of accent (average F1 score of 24.5) and language (average F1 score of 68.8) as well as emotion classification tasks (average F1 score of 27.2).

Among the open-sourced models, Qwen2-Audio and Typhoon-1.5 are the strongest performers based on the frequency of being among the top 3 models (Qwen2-Audio: 8/14; Typhoon-1.5: 7/14). Qwen2-Audio shows outstanding performance on gender (average F1 score of 69.1) and age classification (average F1 score of 12.3) which outperform all other models. Typhoon demonstrates best instruction following capability among all models, exceeding that of closed-models.

We also perform an evaluation for the sequential ASR pipeline of Whisper-large-v2 and Llama3-8B-Instruct. It shows relatively good performance on benchmarks like CN-College-Listen (average F1 score of 62.6), IEMOCAP (average F1 score of 25.2), and MELD (average F1 score of 22.8), which means information in some of the data instances in those benchmarks can be inferred from textual content only. However, for every task there are

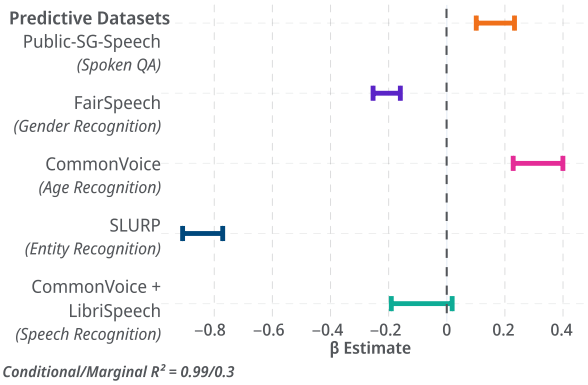

Figure 4: Mixed-effect regression of benchmark performance differences as a predictor of user preferences across models. 15 other features were pre-screened using VIF thresholding (threshold=10.0) with ties removed. Model fitting was performed fixed effects for benchmarks and random effects for model identity. The model achieved conditional/marginal R^2 of 0.99/0.30.

end-to-end LAMs outperforming the ASR pipeline setting. This highlights that elements such as emotion, relationships, and sarcasm can be conveyed through vocal cues, necessitating **speech comprehension that goes beyond textual information**.

6.4 Predictive Power of Static Benchmarks

The speed of iteration offered by static evaluations is invaluable during model development. As such, static benchmarks which correlate strongly with interactive evaluations, such as MMLU (Hendrycks et al., 2021), MTBench (Zheng et al., 2023b), and AlpacaEval (Dubois et al., 2024b), are often used as proxies for text-based LLMs. Here, we test *whether* this holds for speech evaluation and *which* speech benchmarks are the best proxies.

First, using simple correlation checks, we find that no single metric strongly correlates with win rates in our evaluation ($\rho_s \leq 0.49$, $\tau \leq 0.33$). Furthermore, the model-capability matrix on static benchmarks is low-rank, with 95% of the variance across our 20 tasks explained by five principle components. Similar to findings in Ruan et al. (2024), this suggests that despite a large number of benchmarks, only a few core axes of model performance are evaluated by these benchmarks.

While individual benchmarks are not clearly correlated, aggregated benchmarks could potentially have a smaller gap. To test this, we model head-to-head comparisons using logistic regression, where the dependent variable represents whether users preferred Model B over Model A. To perform the analysis shown in Figure 4, we computed the normalized performance difference as $\frac{s_B^i - s_A^i}{|s_A^i| + |s_B^i|}$ for each benchmark i, where s_A^i and s_B^i are the scores of Model A and Model B, respectively. To account for benchmark-specific effects while controlling for other underlying model features, such as output style, we use a mixed-effects model, treating benchmarks as fixed effects and model identity as a random effect. Our regression model gives a marginal R^2 of 0.3. This indicates that, unlike the text-only setting, current speech benchmarks have limited predictive power for user preferences.

However, some benchmarks did show significant positive correlations with user preferences. *CommonVoice - Age* showed the strongest positive association ($\beta = 0.314$). However, this correlation is particularly notable as all models performed below random chance on this benchmark, suggesting it may be **capturing systematic biases in model behavior** rather than meaningful capability differences. *Public-SG-Speech*, a speech question answering task (Wang et al., 2024), also shows a moderate positive effect ($\beta = 0.167$). Notably, this task is, by construction, solvable using solely textual transcripts of the input speech since the questions were created based on the transcripts. This aligns with our observations on the overall strength of the pipeline model and the range of tasks found in Section 4.

7 Conclusion

Speech as an input modality has clear advantages for users when interacting with AI. It offers faster communication speed and enables the use of paralinguistic information. For users to benefit from these advantages, model developers must evaluate LAMs in ways aligned with user preferences. To test benchmark alignment, we collect over 7500 user voice queries and preferences for six different model settings, allowing us to analyze the expected use cases of LAMs and the models that users prefer the most during their interactions. Our results suggest that future benchmarks should focus more on testing models' ability to interact for efficiency purposes like knowledge expansion and task execution. With users' free-text explanations, we also identify key dimensions LAMs could work towards for better interactive capability: users still value the *pragmatic value* (e.g. helpfulness, level of details, accuracy) (Garza et al., 2021) and *degree of adaptation* (e.g. language appropriateness, human-likeness) (Zargham et al., 2022).

While this work establishes key findings for speech-in text-out interaction, speech-in *speech*-out models are a natural next step for LAMs. Exploring how our insights extend to such rich, real-time audio interactions presents key challenges to the existing norm of pairwise preferences, but represents an exciting direction for future work.

Limitations

The current platform we use to evaluate LAMs' performances only supports single-turn interactions, and users are paid to interact with models and contribute their votes. Furthermore, since crowdworkers were given minimal constraints to better reflect their expectations, our evaluation focuses only on top-of-mind use cases. We expect that usage is likely to shift through long-term interaction with LAM systems as users become more familiar with the models' capabilities and become more comfortable with interacting naturally and emotively. These biases influence our current model ranks and likely negatively influence the ranks of commercial LAMs such as GPT-4o and Gemini which are designed for long-term uses.

Similarly, while our data collection did not require users to only utilize English, we only recruited participants who live in the United States. Therefore, our evaluation primarily assesses English language capabilities. In particular, this punishes models which aim for multilingual support, such as Typhoon and the Qwen models, which respectively include Thai and Chinese training data and occasionally respond to users in those languages rather than in English. However, this may not influence the ranks of Gemini and GPT-4o, which are also multilingual LAMs.

Finally, our setting is restricted to speech-in text-out format. As our analysis highlights, much of the qualitative user feedback focuses on the text output style rather than the capabilities of content. LAMs that are tuned for speech-in speech-out interaction, which only describes GPT-4o in our evaluations, likely have output styles that are more biased towards preferences for speech outputs and are likely penalized for this in our model ranks. While this is a limitation, we also think this highlights that model developers should likely tune their models to understand style preferences dependent on output modality instead of using a uniform treatment.

Ethical Statement

Interacting with speech models can be associated with some ethical considerations about privacy and security, as we will have access to users' voice data, which, if mishandled, could lead to unauthorized surveillance or data breaches. This study has been approved by the Institutional Review Board (IRB) at the researchers' institution, and we obtained participant consent with a standard institutional consent form to record their voices. We anonymously store the data by applying advanced noise-masking techniques to the audio recordings, effectively reducing the recognizability of voices and ensuring that individuals cannot be easily identified. We will release the processed data only upon request and only for research purposes, ensuring strict control over its distribution and use.

Acknowledgment

We appreciate the feedback provided by SALT members and anonymous ACL reviewers. We are thankful for computing support provided by SCB 10X, SCBX Group through Stanford HAI and credit support through the Stanford HAI-GCP Cloud Credit Grants. This work is funded in part by ONR Grant N000142412532. We also would like to thank Qwen team for their help with building a model API endpoint. Minzhi Li is supported by the A*STAR Computing and Information Science (ACIS) Scholarship and Google PhD fellowship.

References

Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. 2019. Gradio: Hassle-free sharing and testing of [ml] models in the wild. *arXiv preprint arXiv:1906.02569*.

AI@Meta. 2024. Llama 3 model card.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. *Preprint*, arXiv:1912.06670.

Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. 2020. SLURP: A spoken lan-

guage understanding resource package. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7252–7262, Online. Association for Computational Linguistics.

Purnima Bholowalia and Arvind Kumar. 2014. Ebk-means: A clustering technique based on elbow method and k-means in wsn. *International Journal of Computer Applications*, 105(9).

Egor Bogomolov, Aleksandra Eliseeva, Timur Galimzyanov, Evgeniy Glukhov, Anton Shapkin, Maria Tigina, Yaroslav Golubev, Alexander Kovrigin, Arie van Deursen, Maliheh Izadi, and 1 others. 2024. Long code arena: a set of benchmarks for long-context code models. *arXiv preprint arXiv:2406.11612*.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and 1 others. 2023. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _Obviously_ perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629, Florence, Italy. Association for Computational Linguistics.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T Tan, and Haizhou Li. 2024. Voicebench: Benchmarking llm-based voice assistants. *arXiv preprint arXiv:2410.17196*.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, and 1 others. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *Preprint*, arXiv:1803.05457.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, and 31 others. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. *Preprint*, arXiv:2409.17146.

Benjamin D Douglas, Patrick J Ewell, and Markus Brauer. 2023. Data quality in online human-subjects research: Comparisons between mturk, prolific, cloudresearch, qualtrics, and sona. *Plos one*, 18(3):e0279720.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2024a. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *Preprint*, arXiv:2404.04475.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2024b. Alpacafarm: A simulation framework for methods that learn from human feedback. *Preprint*, arXiv:2305.14387.

Peer Eyal, Rothschild David, Gordon Andrew, Evernden Zak, and Damer Ekaterina. 2021. Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, pages 1–20.

Fahim Faisal, Sharlina Keshava, Antonios Anastasopoulos, and 1 others. 2021. Sd-qa: Spoken dialectal question answering for the real world. *arXiv preprint arXiv:2109.12072*.

Katelyn Garza, Katrina Henley, and Cameron Long. 2021. Artificial intelligence (ai) assistant helpfulness.

Yuan Gong, Hongyin Luo, Alexander H Liu, Leonid Karlinsky, and James Glass. 2023. Listen, think, and understand. *arXiv preprint arXiv:2305.10790*.

Steven Greenberg. 1996. Understanding speech understanding: Towards a unified theory of speech perception. In *Proceedings of the ESCA Tutorial and Advanced Research Workshop on the Auditory Basis of Speech Perception*, pages 1–8. Keele, England.

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 2019. UR-FUNNY: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2046–2056, Hong Kong, China. Association for Computational Linguistics.

William Held, Ella Li, Michael Ryan, Weiyan Shi, Yanzhe Zhang, and Diyi Yang. 2024. Distilling an end-to-end voice assistant without instruction training data. *arXiv preprint arXiv:2410.02678*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Thomas Wiben Jensen and Sarah Bro Pedersen. 2016. Affect and affordances–the role of action and emotion in social interaction. *Cognitive Semiotics*, 9(1):79–103.

Ye Jia, Michelle Tadmor Ramanovich, Quan Wang, and Heiga Zen. 2022. Cvss corpus and massively multilingual speech-to-speech translation. *arXiv preprint arXiv:2201.03713*.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, and 1 others. 2021. Dynabench: Rethinking benchmarking in nlp. *arXiv preprint arXiv:2104.14337*.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033.

Siddique Latif, Moazzam Shoukat, Fahad Shamshad, Muhammad Usama, Yi Ren, Heriberto Cuayáhuitl, Wenwu Wang, Xulong Zhang, Roberto Togneri, Erik Cambria, and 1 others. 2023. Sparks of large audio models: A survey and outlook. *arXiv preprint arXiv:2308.12792*.

Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, and 1 others. 2022. Evaluating human-language model interaction. *arXiv preprint arXiv:2212.09746*.

Zongxia Li, Ishani Mondal, Huy Nghiem, Yijun Liang, and Jordan Boyd-Graber. 2024. Pedants: Cheap but effective and interpretable answer equivalence. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9373–9398.

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. 2024. Wildbench: Benchmarking llms with challenging tasks from real users in the wild. *arXiv preprint arXiv:2406.04770*.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. 2023. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*.

Yujie Lu, Dongfu Jiang, Wenhu Chen, William Yang Wang, Yejin Choi, and Bill Yuchen Lin. 2024. Wildvision: Evaluating vision-language models in the wild with human preferences. *arXiv preprint arXiv:2406.11069*.

Inbal Magar and Roy Schwartz. 2022. Data contamination: From memorization to exploitation. *arXiv preprint arXiv:2203.08242*.

Ankur Mallick, Kevin Hsieh, Behnaz Arzani, and Gauri Joshi. 2022. Matchmaker: Data drift mitigation in machine learning for large-scale systems. *Proceedings of Machine Learning and Systems*, 4:77–94.

Yonatan Oren, Nicole Meister, Niladri Chatterji, Faisal Ladhak, and Tatsunori B Hashimoto. 2023. Proving test set contamination in black box language models. *arXiv preprint arXiv:2310.17623*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Yifan Peng, Jinchuan Tian, Brian Yan, Dan Berrebbi, Xuankai Chang, Xinjian Li, Jiatong Shi, Siddhant Arora, William Chen, Roshan Sharma, and 1 others. 2023. Reproducing whisper-style training using an open-source toolkit and publicly available data. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Kunat Pipatanakul, Phatrasek Jirabovonvisut, Potsawee Manakul, Sittipong Sripaisarnmongkol, Ruangsak Patomwong, Pathomporn Chokchainant, and Kasima Tharnpipitchai. 2023. Typhoon: Thai large language models. *arXiv preprint arXiv:2312.13951*.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint*.

Sherry Ruan, Jacob O Wobbrock, Kenny Liou, Andrew Ng, and James Landay. 2016. Speech is 3x faster than typing for english and mandarin text entry on mobile devices. *arXiv preprint arXiv:1608.07323*.

Yangjun Ruan, Chris J Maddison, and Tatsunori Hashimoto. 2024. Observational scaling laws and the predictability of language model performance. *arXiv preprint arXiv:2405.10938*.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Selina Jeanne Sutton, Paul Foulkes, David Kirk, and Shaun Lawson. 2019. Voice as a design material: Sociophonetic inspired design strategies in human-computer interaction. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–14.

Alex Tamkin, Miles McCain, Kunal Handa, Esin Durmus, Liane Lovitt, Ankur Rathi, Saffron Huang, Alfred Mountfield, Jerry Hong, Stuart Ritchie, Michael Stern, Brian Clarke, Landon Goldberg, Theodore R. Sumers, Jared Mueller, William McEachen, Wes Mitchell, Shan Carter, Jack Clark, and 2 others. 2024. Clio: Privacy-preserving insights into real-world ai use. *Preprint*, arXiv:2412.13678.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Michael Tomasello. 2023. Having intentions, understanding intentions, and understanding communicative intentions. In *Developing theories of intention*, pages 63–76. Psychology Press.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: open and efficient foundation language models. arxiv. *arXiv preprint arXiv:2302.13971*.

Irina-Elena Veliche, Zhuangqun Huang, Vineeth Ayyat Kochaniyan, Fuchun Peng, Ozlem Kalinli, and Michael L. Seltzer. 2024. Towards measuring fairness in speech recognition: Fair-speech dataset. *Preprint*, arXiv:2408.12734.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F Chen. 2024. Audiobench: A universal benchmark for audio large language models. *arXiv preprint arXiv:2406.16020*.

Changhan Wang, Juan Pino, Anne Wu, and Jiatao Gu. 2020a. Covost: A diverse multilingual speech-to-text translation corpus. *arXiv preprint arXiv:2002.01320*.

Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. 2021. Covost 2 and massively multilingual speech translation. In *Interspeech*, pages 2247–2251.

Changhan Wang, Anne Wu, and Juan Pino. 2020b. Covost 2 and massively multilingual speech-to-text translation. *Preprint*, arXiv:2007.10310.

Zhuxiaona Wei and James A. Landay. 2018. Evaluating speech-based smart devices using new usability heuristics. *IEEE Pervasive Computing*, 17(2):84–96.

Alex Wilf, Leena Mathur, Sheryl Mathew, Claire Ko, Youssouf Kebe, Paul Pu Liang, and Louis-Philippe Morency. 2023. Social-iq 2.0 challenge: Benchmarking multimodal social understanding. `https://github.com/abwilf/Social-IQ-2.0-Challenge`.

Hanlin Wu and Zhenguang G Cai. 2024. Speaker effects in spoken language comprehension. *arXiv preprint arXiv:2412.07238*.

Yijing Wu, SaiKrishna Rallabandi, Ravisutha Srinivasamurthy, Parag Pravin Dakle, Alolika Gon, and Preethi Raghavan. 2023. Heysquad: A spoken question answering dataset. *arXiv preprint arXiv:2304.13689*.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and 1 others. 2024. Air-bench: Benchmarking large audio-language models via generative comprehension. *arXiv preprint arXiv:2402.07729*.

Xue Ying. 2019. An overview of overfitting and its solutions. In *Journal of physics: Conference series*, volume 1168, page 022022. IOP Publishing.

Nima Zargham, Dmitry Alexandrovsky, Jan Erich, Nina Wenig, and Rainer Malaka. 2022. "i want it that

way": Exploring users' customization and personalization preferences for home assistants. In *CHI conference on human factors in computing systems extended abstracts*, pages 1–8.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric. P Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2023a. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *Preprint*, arXiv:2309.11998.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023b. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, and 1 others. 2023. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*.

A Prompt

A.1 Humor Detection

Prompt 1
Is the audio humorous?
A. Yes
B. No
{*input audio*}
Prompt 2
Respond whether the input is intended to be humorous. Answer with a simple yes or no.
{*input audio*}
Prompt 3
Based on the audio, indicate if it is humorous. Please give your answer as yes or no.
{*input audio*}

A.2 Sarcasm Detection

Prompt 1
Is the audio sarcastic?
A. Yes
B. No
{*input audio*}
Prompt 2
Respond whether the input is intended to be sarcastic. Answer with a simple yes or no.
{*input audio*}
Prompt 3
Based on the audio, indicate if it is sarcastic. Please give your answer as yes or no.
{*input audio*}

A.3 Intent Detection

Prompt 1
What is the intent of the speaker?
A. query alarm
B. remove alarm
C. set alarm
D. turn down audio volume
{*input audio*}
Prompt 2
Respond what intent the input exhibits. Answer with "query alarm", "remove alarm", "set alarm", or "turn down audio volume".
{*input audio*}
Prompt 3
Based on the audio, indicate the intent of the speaker. Please give your answer as "query alarm", "remove alarm", "set alarm", or "turn down audio volume".
{*input audio*}

A.4 Emotion Recognition

Prompt 1
What is the emotion state of the speaker?
A. Angry
B. Happy
C. Sad
D. Neutral
{*input audio*}
Prompt 2
Respond in a single word what emotion the input exhibits. Answer with "angry", "happy", "sad", or "neutral".
{*input audio*}
Prompt 3
Based on the audio, indicate the emotion of the speaker. Please give your answer as "angry", "happy", "sad", or "neutral".
{*input audio*}

The prompts above are used for IEMOCAP dataset. For MELD dataset, we apply the same format with the only changes in emotion options.

A.5 Age Classification

Prompt 1
What is the age of the speaker?
A. 18-22
B. 23-30

C. 31-45
D. 46-65
{*input audio*}

Prompt 2

Respond the age of the speaker based on the input. Answer with “18-22”, “23-30”, “31-45”, or “46-65”.
{*input audio*}

Prompt 3

Based on the audio, indicate the age of the speaker. Please give your answer as “18-22”, “23-30”, “31-45”, or “46-65”.
{*input audio*}

The prompts above are used for FairSpeech dataset. For CommonVoice dataset, we apply the same format with the only changes in age options.

A.6 Gender Classification

Prompt 1

What is the gender of the speaker?
A. female
B. male
{*input audio*}

Prompt 2

Respond the gender of the speaker based on the input. Answer with “female” or “male”.
{*input audio*}

Prompt 3

Based on the audio, indicate the gender of the speaker. Please give your answer as “female” or “male”.
{*input audio*}

A.7 Relationship Classification

Prompt 1

Is the relationship between the two speakers more likely to be friend or relative?
A. friend
B. relative
{*input audio*}

Prompt 2

Respond what relationship the two speakers have based on the input. Answer with “friend” or “relative”.
{*input audio*}

Prompt 3

Based on the audio, indicate the relationship between the speakers. Please give your answer as “friend” or “relative”.
{*input audio*}

A.8 Accent Classification

Prompt 1

’What is the accent of the speaker?
A. Australian English
B. Canadian English
C. England English
D. India and South Asia (India, Pakistan, Sri Lanka)
E. United States English {*input audio*}

Prompt 2

Respond the accent of the speaker based on the input. Answer with “Australian English” , “Canadian English”, “England English”, “South Asia (India, Pakistan, Sri Lanka)”, or “United States English”.
{*input audio*}

Prompt 3

Based on the audio, indicate the accent of the speaker. Please give your answer as “Australian English”, “Canadian English”, “England English”, “South Asia (India, Pakistan, Sri Lanka)”, or “United States English”.
{*input audio*}

The prompts above are used for FairSpeech dataset. For CommonVoice dataset, we apply the same format with the only changes in age options.

A.9 Instruction Following

Please follow the instruction in the speech.
{*input audio*}

Prompt 2

Respond to the instruction in the given audio.
{*input audio*}

Prompt 3

Based on the audio instruction, please provide a response following it.
{*input audio*}

A.10 Speech Grounding, Entity Recognition, Language Classification, Question Answering

In the datasets for these tasks, there are *question* and *options* provided so we structured our prompts as the following:

Prompt 1

{*question*}
{*Option A*}
{*Option B*}
{*Option C*}
{*Option D*}
{*input audio*}

Prompt 2

Respond to the question: {*question*}. Answer with "{*Option A*}", "{*Option B*}", "{*Option C*}", or "{*Option D*}".
{*input audio*}

Prompt 3

Based on the audio, {*question*}. Please give your answer as "{*Option A*}", "{*Option B*}", "{*Option C*}", or "{*Option D*}".
{*input audio*}

The prompts above are used for FairSpeech dataset. For CommonVoice dataset, we apply the same format with the only changes in age options.

B Correlation among Static Benchmarks

We compute correlation of LAM performance on 20 static benchmarks (see Figure A.1). During regression analysis, we take the average model performance for benchmarks with a very high correlation as independent variables to remove high multicollinearity.

C Interactive Evaluation Interface

In Figure A.2, we illustrate the gradio interface we used to collect user preferences. Users can submit their voice input, get responses from two randomly sampled LAMs, contribute their votes and reasons for their votes.

D Model Ranking in Static and Interactive Evaluation

To understand the extent to which previous static benchmarks can reflect the relative interactive capability of LAMs, we also compare the result in interactive evaluation to that in static evaluation by computing the top-k Kendall Tau Distance (Bogomolov et al., 2024) between the model rankings (see Figure A.3).

We found that *none of the 20 static benchmarks reflects exactly the same model rank in the interactive evaluation* with non-zero rank distance, suggesting that any single static benchmark is inadequate in reflecting the relative interactive capabilities of audio models and an interactive way of evaluation is essential. Among the 20 static benchmarks we tested, model ranks on Commonvoice age classification (rank distance: 0.20) is most similar to that in the interactive evaluation. On the other hand, model rank on SLURP speech entity recognition task (rank distance: 0.56) is most uncorrelated with that reflected in user preferences. The result is also reflected in our regression analysis.

E Logistic Regression

Besides mixed effect regression, we also perform a regular logistic regression without mixed effects to test the predictive power of static benchmarks with regard to interactive capability (Figure A.4). We obtain similar findings as mixed effect regression where *Public-SG-Speech* and *CommonVoice Age Recognition* demonstrate positive effects. On the other hand, the effect of ASR benchmarks turns positive without taking mixed effects into consideration.

Figure A.1: PCA Analysis of model performance on 20 static benchmarks.

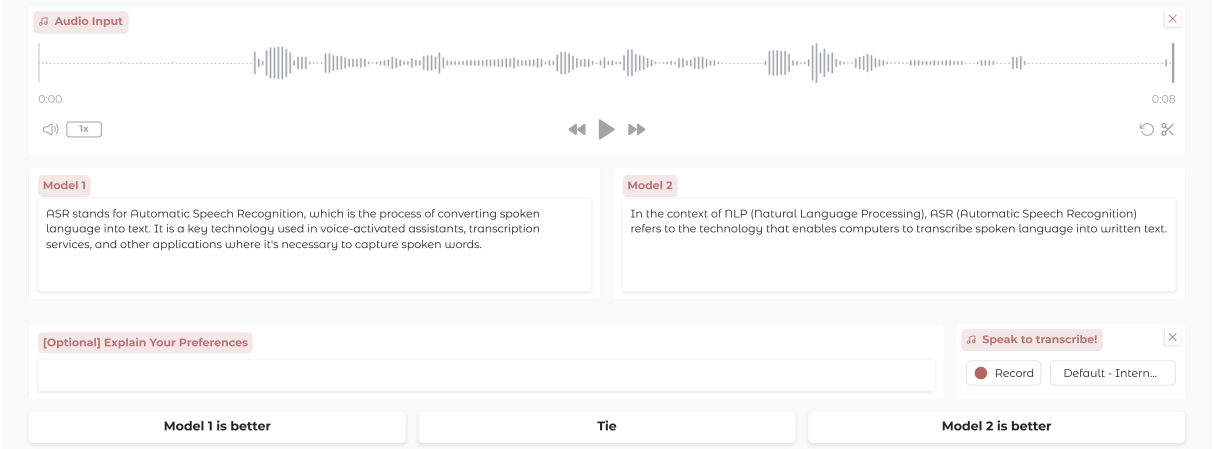

Figure A.2: Gradio interface of interactive evaluation. Users record their own speech and audio without constraints and receive responses from two LAM systems anonymously. They then provide a binary preference between the models, and are provided the option to provide qualitative feedback through either voice or text.

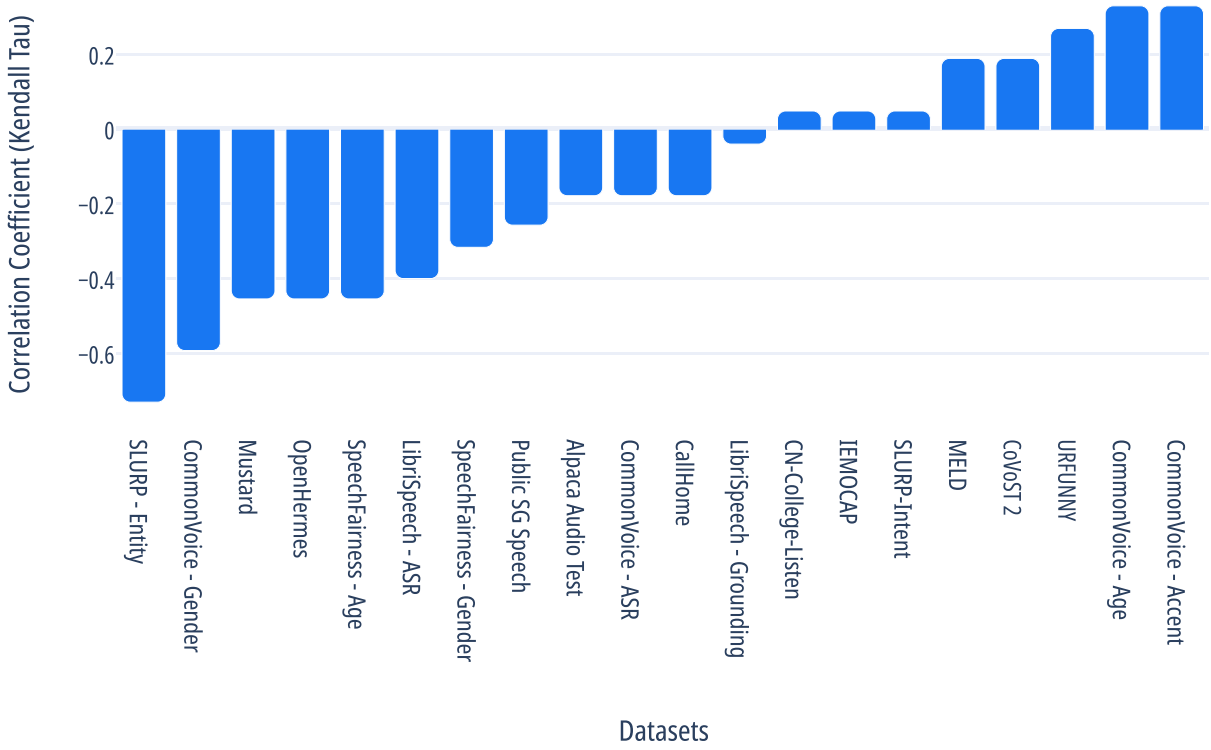

Figure A.3: Kendall tau rank distance between model rank in interactive evaluation and that on different static benchmarks.

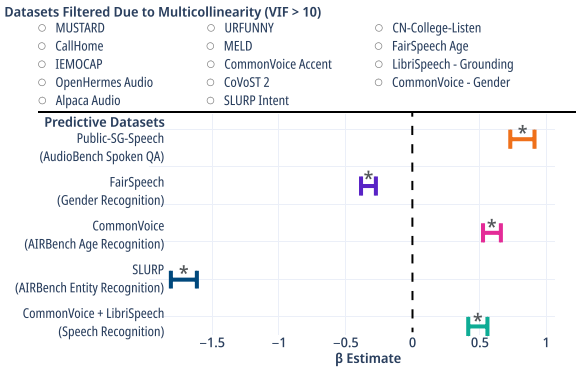

Figure A.4: Forest plot demonstrating the effect sizes of static benchmarks in a logistic regression without mixed-effects for predicting individual user preferences. The results are overall consistent with our mixed effects regression with the only shift being the β for ASR tasks.