SpeechIQ: Speech Intelligence Quotient Across Cognitive Levels in Voice Understanding Large Language Models

Zhen Wan[1]   **Chao-Han Huck Yang**[2]   **Yahan Yu**[1]   **Jinchuan Tian**[3]
Sheng Li[4]   **Ke Hu**[2]   **Zhehuai Chen**[2]   **Shinji Watanabe**[3]   **Fei Cheng**[1]
Chenhui Chu[1]   **Sadao Kurohashi**[1]

[1]Kyoto University   [2]Nvidia
[3]Carnegie Mellon University   [4]Institute of Science Tokyo
{zhenwan}@nlp.ist.i.kyoto-u.ac.jp
{hucky}@nvidia.com

Abstract

We introduce Speech Intelligence Quotient (SpeechIQ) as a new form of human cognition-inspired evaluation pipeline for voice understanding large language models (LLM_Voice), designed to assess their voice understanding ability. Moving beyond popular voice understanding metrics such as word error rate (WER), SpeechIQ examines LLM_Voice across three cognitive levels motivated by Bloom's Taxonomy: (1) Remembering (i.e., WER for verbatim accuracy); (2) Understanding (i.e., similarity of LLM's interpretations); and (3) Application (i.e., QA accuracy for simulating downstream tasks). We demonstrate that SpeechIQ not only quantifies voice understanding abilities but also provides unified comparisons between cascaded methods (e.g., ASR-LLM) and end-to-end models, identifies annotation errors in existing benchmarks, and detects hallucinations in LLM_Voice. Our framework represents a first-of-its-kind intelligence examination that bridges cognitive principles with voice-oriented benchmarks, while exposing overlooked challenges in multi-modal training. Our code is here.[1]

1 Introduction

The rapid rise of voice understanding Large Language Models (LLM_Voice) that can process voice input via the speech portal has ushered in a new paradigm of human-machine interaction (Reddy, 1988), where LLM_Voice process spoken instructions, infer semantic intent, and execute downstream tasks (Jelinek et al., 1991; Jurafsky et al., 1995; Zue and Glass, 2000; Tur et al., 2005; Mesnil et al., 2014; Kawahara, 2019; Huang et al., 2024; Mahmood et al., 2025). LLM_Voice bridge *speech* and *language intelligence* (Sparks et al., 1996), enabling applications from voice assistants to interactive robots. A fundamental prerequisite for deploying robust LLM_Voice lies in establishing reliable
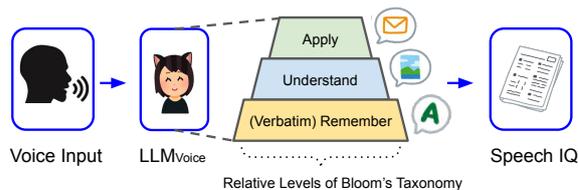

Figure 1: An overview of cognitive levels in voice understanding large language model systems related to the first three foundational hierarchies of Bloom's Taxonomy. We design a corresponding examination to measure SpeechIQ as a detailed pipeline in Figure 2.

evaluation metrics for voice understanding. These metrics ensure that the system accurately comprehends voice inputs. Automatic speech recognition (ASR) is the central component of existing cascaded[2] LLM_Voice systems, transcribing voice into text for subsequent LLM processing. Consequently, the research community primarily assesses voice understanding using transcription error metrics (Hunt, 1990; Klakow and Peters, 2002), with Word Error Rate (WER) standing as the de facto standard.

However, WER primarily measures the lexical recall of ASR, its transcribed quality does **not** fully capture the semantic understanding (He et al., 2011; Chen et al., 2018; Szymański et al., 2020) and task completion (Van Strien et al., 2020) capabilities of LLMs via their speech portal. For instance, WER treats all lexical errors equally, ignoring how transcription inaccuracies propagate to higher-level comprehension (*i.e.*, semantic understanding) and executions (*i.e.*, downstream tasks).

Meanwhile, a growing trend in LLM_Voice research is the shift toward end-to-end multi-modal models (Rubenstein et al., 2023; Radhakrishnan et al., 2023; Maiti et al., 2024; Hu et al., 2024; Dé-

---

[1]https://github.com/ku-nlp/SpeechIQ

[2]We refer to two-pass based neural ASR-LM systems (Hoffmeister et al., 2016; Sainath et al., 2019; Huang et al., 2020), which have been widely adopted in the current voice assistants, including Siri, Alexa, and Google Assistant.

fossez et al., 2024; Chen et al., 2024), which bypass explicit ASR transcription and directly map audio inputs to task-specific outputs (e.g., answers, actions). While this paradigm simplifies pipelines and reduces error cascades, its end-to-end voice service obviates the need for generating an intermediate transcript, rendering traditional WER-based evaluation obsolete (Kuo et al., 2020; Tüske et al., 2021). This creates a critical methodological gap: without a reliable and potentially unified metric, comparing voice understanding across architectural choices, such as the modularity of "*cascaded neural systems*" and the efficiency of "*end-to-end mutlimodal models*," remains insufficiently characterized.

The above two limitations mirror a fundamental challenge in assessing artificial intelligence: as Figure 1, human-like intelligence is inherently hierarchical. Drawing on Bloom's Taxonomy (Bloom and Krathwohl, 1969; Bengio et al., 2009), a cognitive framework where skills ascend from basic remembering to advanced creating, we posit that LLM_{Voice} should be evaluated through analogous levels. In this framework, WER represents the lowest level by measuring verbatim recall while neglecting higher-order abilities such as semantic understanding and task-solving. For instance, two transcripts with identical WER may differ drastically in meaning (Szymański et al., 2023), leading to divergent LLM_{Voice} responses or failed instructions.

To address this gap, we propose a hierarchical IQ test for LLM_{Voice} that evaluates intelligence across three fundamental levels (refer to limitations) aligned with Bloom's Cognitive Taxonomy: (1) Remember in verbatim: Lexical accuracy is quantified using WER; (2) Understand: Semantic consistency is assessed by comparing LLM responses to ASR outputs and ground-truth transcripts. LLMs are prompted to infer the background context (*e.g.*, domain, speaker intent) and generate a summary of the speech, then measure hidden states similarity between ASR-derived and ground truth-derived responses; (3) Apply: Task-solving capability is tested via multi-choice question-answering (QA) pairs constructed from ground-truth transcripts. LLM_{Voice} answers questions based on speech inputs, with accuracy reflecting real-world utility. Then, we draw inspiration from Raven's Progressive Matrices (Raven and Court, 1998; John and Raven, 2003) (*i.e.*, a human IQ test) that aggregates performance across various dimensions into a single score.

Beyond overall evaluation, our SpeechIQ framework yields two key insights:

- Cascade systems (ASR+LLM) outperform end-to-end models under similar scaling, revealing modality interference in joint speech-text training.

- Cross-model QA consistency exposes ground-truth errors. By isolating questions most LLM_{Voice}, cannot solve, we create an "unanswerable" set that helps detect hallucinations (Frieske and Shi, 2024; Koenecke et al., 2024) and refine benchmark annotations.

Our work advances LLM_{Voice} evaluation by merging cognitive principles with practical metrics, revealing hidden challenges in multimodal inference (*e.g.*, hallucination inheritance), and offering tools to build more robust systems. Code and data will be publicly available.

2 Related Work

2.1 ASR Evaluation

ASR systems have long been evaluated using lexical similarity metrics such as WER, which quantifies the Levenshtein distance (Navarro, 2001) between ASR outputs and reference transcripts. Metrics like character error rate (MacKenzie and Soukoreff, 2002) (CER), sentence error rate (Juffs and Harrington, 1996) (SER), and translation error rate (TER) extend this framework to the sentence or translation level but share WER's core limitation: treating all errors equally, regardless of semantic impact. While effective for benchmarking transcription fidelity, these metrics do not fully capture how errors propagate to downstream tasks.

Recent work mitigates WER's semantic insensitivity by incorporating sentence embeddings into evaluation. Sentence similarity-based metrics (Kim et al., 2021, 2022) and BERTScore (Zhang et al., 2020) leverage pre-trained language models to compute semantic correspondence between ASR hypotheses and ground-truth references. Others propose task-specific metrics, such as Medconcept WER (Adedeji et al., 2024) which arranges more weights on *medical entities* in WER computation or severity score (Whetten and Kennington, 2023) that involves sentiment similarity in the evaluation. However, these methods are either confined to specific tasks or require **curated labeled data**, limiting their generalizability to open-domain LLM_{Voice}.

Efforts to develop hybrid metrics that combine error rate and semantic similarity, such as H_eval (Sasindran et al., 2023), and Sema (Sasindran et al., 2024), further enrich the evaluation landscape. Concurrently, machine translation inspired metrics like BLEU Score (Papineni et al., 2002) and COMET (Rei et al., 2020), have started to be adapted for ASR to assess fluency and pragmatic adequacy. However, these approaches still emphasize text-to-text alignment between ASR outputs and references, rather than from a view of influencing the responses and actions of downstream LLMs, motivating the proposed SpeechIQ as a multi-dimensional metric.

2.2 LLM$_{\text{Voice}}$ Understanding Systems

In general, speech-based LLMs encompass both the understanding of voice signals and the generation (Agostinelli et al., 2023; Borsos et al., 2023a; Yang et al., 2024c) of vocal or general audio outputs. In this work, we focus exclusively on **voice understanding tasks** with text outputs as one first step to examine voice intelligence (Sparks et al., 1996) or its potential voice world model (Ha and Schmidhuber, 2018; Matsuo et al., 2022) via LLM backbone, which we denote as LLM$_{\text{Voice}}$. LLM$_{\text{Voice}}$ understanding systems are designed to interact (Shah et al., 2018; López et al., 2018) with users and execute task-oriented instructions via speech-based inputs. This form of "voice-to-text" architecture can be broadly categorized into three main types, all of which achieve state-of-the-art ASR quality on public leaderboards (Radford et al., 2022; Chen et al., 2023; Puvvada et al., 2024):

(1) Cascaded ASR + LLM This approach follows a traditional pipeline-based structure, where an ASR model (Watanabe et al., 2018; Gulati et al., 2020; Radford et al., 2022) transcribes the spoken input, and the 1-best transcription is passed to textual LLM as one operation system (Wu et al., 2024; Dighe et al., 2024) for direct response generation.

(2) Cascaded ASR Hypotheses + Generation Error Correction (GER) + LLM In this approach, the ASR model first generates multiple possible transcriptions (*i.e.*, hypotheses) for a given speech input, where "texts" are then treated as input features for a many-to-one GER post-editing module (Yang et al., 2023; Velikovich et al., 2024; Hori et al., 2025). The GER module refines these hypotheses to determine the most accurate transcription, which is subsequently processed by the LLM.

(3) End-to-End Multi-Modal Models These models are inherently designed to support both speech and text modalities (Rubenstein et al., 2023; Hu et al., 2024; Lyu et al., 2023; Zhang et al., 2024a). Unlike cascaded approaches, they process audio inputs directly and generate textual outputs **without** requiring intermediate ASR transcriptions. This paradigm leverages a multimodal learning framework (*i.e.*, a speech embedding encoder (Radhakrishnan et al., 2023) or tokenizer (Borsos et al., 2023b; Zhang et al., 2024b)) to achieve end-to-end speech understanding and reasoning.

In summary, we carefully selected and examined aforementioned three ASR-LLM systems. System 1 serves as the most cost-effective solution (Hermansky and Junqua, 1988) for voice interactions with LLMs. System 2 is designed to decode texts that offer richer information (Chan et al., 2023; Yang et al., 2024b) for textual applications. System 3 highlights a low-latency, multimodal embedding injection approach for large language models (OpenAI et al., 2024; Team et al., 2024). Next, we introduce our proposed method, which corresponds to Bloom's cognitive taxonomy.

3 SpeechIQ: Speech Intelligence Quotient

We will subsequently introduce each level of the test, from the remember level to both the understand level and apply level. At the end of this section, we will discuss the computation of the final SpeechIQ. An overview of our workflow is in Figure 2.

3.1 Remember: WER Metric

The remember level assesses the ability to reproduce spoken content accurately. WER serves as a natural metric since it measures the verbatim discrepancy between ASR output and the ground truth at a granular level, with the advantage that it captures even the smallest differences, making it highly sensitive to transcription errors.

3.2 Understand: Similarity Metric

The understand level assesses whether the ASR-generated transcript effectively conveys the intended meaning of the original speech. This is particularly important when ASR output serves as input to the LLM, and a minor transcription difference could lead to significant semantic variations to LLMs. For instance, two transcriptions with identical WER scores can have vastly different meanings:
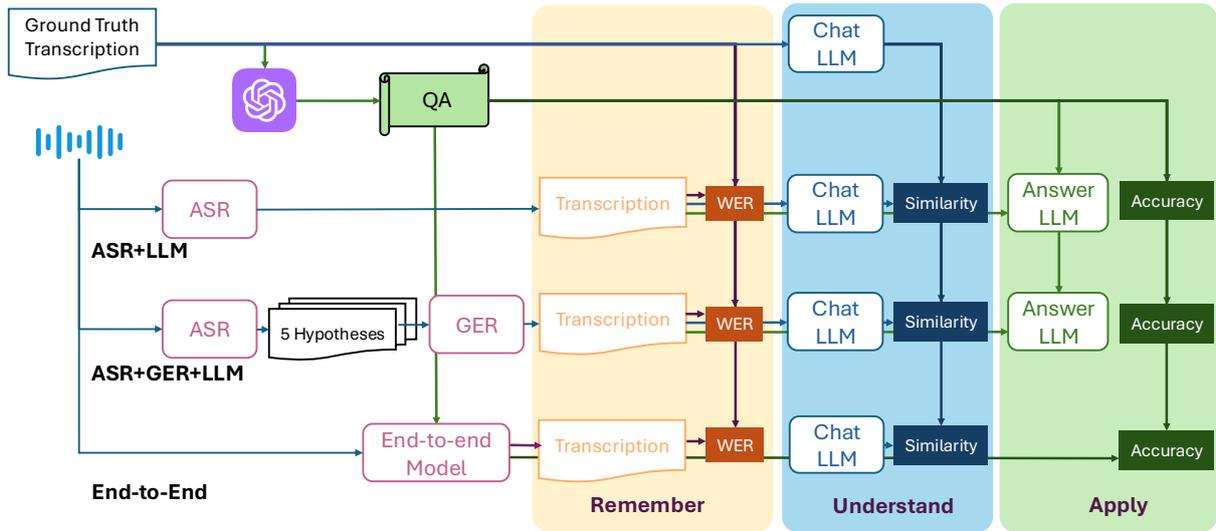
Figure 2: **Overview of Speech IQ (SIQ) test**. We compared cognition-inspired categories of LLM$_{\text{Voice}}$ represented in three rows, with the right columns denoting the SIQ sub-tests. The truth reference will be used in the sub-tests.

- Ground truth: "I feel pain in the lower back."

- ASR 1: "I feel like pain in the _ back."

- ASR 2: "I feel painting in the world back."

Although both have the same WER = 29%, they convey different meanings and lead to drastically different LLM responses.

To quantify the impact of ASR errors on semantic comprehension, we measure the deviation in LLM responses caused by ASR transcription errors. Specifically, inspired by recent work (Jiang et al., 2024; Liu et al., 2024) showing that instructing LLMs to generate *one word* is a simple and straightforward approach to representing semantic meanings, we evaluate LLM-generated responses via two key questions: (1) b: The background scenario of the speech [] in *one word* is. This helps resolve ambiguities and measure whether the ASR content preserves core contextual meaning; (2) s: The summary of this speech [] in *one word* is. This assesses the extent to which ASR errors affect the LLM's high-level understanding. Instead of directly comparing LLM-generated words, we use the last layer of hidden states of LLM for generating the next token as embeddings and compute their cosine similarity with the embeddings generated from the ground truth transcription:

$$\text{Sim\_b} = \cos(\mathcal{M}\_b(\text{ASR}), \mathcal{M}\_b(\text{Ground})) \quad (1)$$

$$\text{Sim\_s} = \cos(\mathcal{M}\_s(\text{ASR}), \mathcal{M}\_s(\text{Ground})) \quad (2)$$

where $\mathcal{M}\_$ denotes the LLM's hidden states. We then select the **lower** similarity as the final Sim score since we intend to capture the semantic gap between the ASR and the ground truth. This embedding-based approach enables a robust evaluation of semantic preservation in ASR outputs, beyond simple word-matching metrics.

3.3 Apply: QA Accuracy

The Apply level evaluates a LLM$_{\text{Voice}}$'s ability to leverage transcribed information for solving downstream problems, reflecting its real-world utility in task-oriented scenarios. We simulate this utility by constructing multi-choice QA, which is a typical listening test for our human language learners. Specifically, for each speech example, we construct 3 questions based on the ground-truth transcription along with 5 choices per question. (including 1 option as "None of the above")

During QA generation, we leverage GPT-4o and prompt it to focus on either the core concept or information details in the transcription. (Appendix A) To mitigate potential errors in generated QA, we also employ GPT-4o itself and Gemini-1.5-flash to answer these questions using the ground-truth transcription. Questions, where GPT-4o or Gemini-1.5-flash fails to produce correct answers, are discarded and regenerated. Then during evaluation, cascaded systems answer questions based on ASR transcriptions, while end-to-end systems directly process speech inputs to generate responses without intermediate ASR steps. Each LLM$_{\text{Voice}}$ answers 5

times per quesion and use the majority vote as the final answer (Wang et al., 2023). QA accuracy is then measured via an exact match and higher accuracy indicates a stronger capability to contextualize and apply spoken content to real-world tasks.

Moreover, we realize that this QA test can also help recognize potential annotation errors in the benchmarks efficiently, since QA generated by annotation errors will be unanswerable to LLM$_{\text{Voice}}$, we use this feature to filter out annotation errors in our main experiments and further utilize them to detect hallucinations in Section 5.5.

3.4 Final SIQ Score

Inspired by standard Raven's progressive matrices (John and Raven, 2003), a popular human IQ mechanism, we have three steps to compute SIQ: (1) harder samples have higher scores; (2) global standardization making scores among levels computable; (3) dynamic weight among levels in computing final SIQ. For each LLM$_{\text{Voice}}$ model j and each speech sample i, we first compute raw scores for the three dimensions as $X^{\text{dim}}_{\text{WER},i,j}$ (taking the negative in subsequent computation), $X^{\text{dim}}_{\text{Sim},i,j}$, and $X^{\text{dim}}_{\text{Acc},i,j}$.

Speech Sample Discrimination Weights To account for variations in sample difficulty, we introduce "discrimination weights" based on inter-model score variance. For each speech sample i, we compute the variance across models:

$$V^{\text{dim}}_i = \text{Var}([X^{\text{dim}}_{\text{Raw},i,j} \ \forall j]) \tag{3}$$

where "dim" refers to 3 levels. Using these variance values, we compute weighted scores:

$$X^{\text{dim}}_j = \frac{\sum_{i=1}^{N} X^{\text{dim}}_{\text{Raw},i,j} \cdot V^{\text{dim}}_i}{\sum_{i=1}^{N} V^{\text{dim}}_i} \tag{4}$$

where N denotes the number of models. This ensures that speech samples with greater discrimination power have a larger influence.

Global Standardization For each dimension, we perform standardization to normalize scores across models:

$$Z^{\text{dim}}_j = \frac{X^{\text{dim}}_j - \mu^{\text{dim}}}{\sigma^{\text{dim}}} \tag{5}$$

where μ_{dim} and σ_{dim} are the mean and standard deviation of all model scores for that dimension.

Dynamic Weight Computation To ensure a balanced contribution of each evaluation dimension, we assign dynamic weights based on variance. Since higher variance may indicate greater instability, we use the "inverse variance":

$$w^{\text{dim}} = \frac{1}{\sigma^{\text{dim}}_{\text{raw}} + \epsilon}, w^{\text{dim}}_f = \frac{w^{\text{dim}}}{\sum w^{\text{dim}}} \tag{6}$$

where $\sigma^{\text{dim}}_{\text{raw}}$ is the standard deviation computed by all raw scores, and ϵ is a small constant to prevent division by zero. The weights are then normalized on the summation of three dimensions.

Final IQ Score Computation The final intelligence score for each model j is computed as a weighted sum of the standardized scores:

$$Score_j = \sum_{\text{dim}} w^{\text{dim}}_f \cdot Z^{\text{dim}}_j \tag{7}$$

Finally, the score is converted into an IQ-like scale:

$$SIQ_j = 100 + 15 \cdot Score_j. \tag{8}$$

As one additional study, we also perform normalization based on model scale (*i.e.*, analogous to age factors in human IQ). However, due to the involvement of multiple variables in neural scaling laws (Kaplan et al., 2020), we report the non-normalized SIQ and provide preference rankings from a human study to validate its correlation.

4 Experiment Setup

4.1 Datasets

To comprehensively evaluate LLM$_{\text{Voice}}$ intelligence across varied domains and real-world challenges, we carefully select two datasets earning22 (Rio et al., 2022) and voxpopuli (Wang et al., 2021) from the popular OpenASR Leaderboard, while an additional dataset Med-ASR-EN[3] is chosen from the medical domain, featuring diverse accents and environments. This ensures that our evaluation tests LLM$_{\text{Voice}}$ performance in different real-world scenarios, including domain-specific speech and challenging acoustic conditions. Considering the high computational cost of leveraging GPT-4o APIs for QA evaluation and running end-to-end LLM$_{\text{Voice}}$ (e.g., Gemini), we extract a subset constructed by the longest audios from each dataset shown in Table 1.

---

[3] https://huggingface.co/datasets/jarvisx17/Medical-ASR-EN

Dataset	# Subset	Domain	WER
Earning22	200	Financial meetings	12.05
Voxpopuli	200	European Parliament	7.48
Medasr	400	Hospital Accented Patients	7.7

Table 1: **Statistics of datasets**. WER results are on whole dataset via Whisper-large-v2.

4.2 Models

We conduct experiments across three major LLM_{Voice} architectures as follows:

ASR+LLM We compare the following widely used ASR models: Whisper-Large-v2, Whisper-Large-v3 (Radford et al., 2022), Canary (Puvvada et al., 2024), and ESPnet (owsm_ctc_v3.1_1B) (Watanabe et al., 2018)

ASR+GER+LLM We select GPT-4o (OpenAI et al., 2024) as the GER module and use Whisper-large-v2 to generate 5 hypotheses for each audio based on beam search.

End-to-end Multimodal Models For speech-to-text multimodal models, we compare Salmonn (Tang et al., 2024), Qwen2-audio (Chu et al., 2024), desta2 (Lu et al., 2025) and the latest Qwen-2.5-Omni(Xu et al., 2025), we also involve any-to-any multimodel models with speech-to-text capabilities: AnyGPT (Zhan et al., 2024), Baichuan-omni-1.5 (Li et al., 2025), and Gemini-1.5 (Team et al., 2024). We specifically denote Baichuan-omni-1.5 and Gemini-1.5 as "end-to-end-large" since they are either scaled up in training data or model size.

In the understanding level test, we select LLaMA-3.1-8B-Instruct to generate hidden states as responses since related work (Jiang et al., 2024; Liu et al., 2024) has proved the strong embedding capability of LLaMA-based models.

Meanwhile, for the LLM in both ASR + LLM and ASR + GER + LLM to answer QA, we use Qwen2-7b (Yang et al., 2024a) for fair comparisons in size. Preliminary experiments to validate Qwen2-7b and other experiment details are in Appendix B.

5 Experimental Results

In this section, we will first show the results of each level, and then the final SIQ scores. Considering the limited space, for each level we show the average score on 3 datasets without Desta2 and AnyGPT since they show a relatively low performance, the complete results are in Appendix.
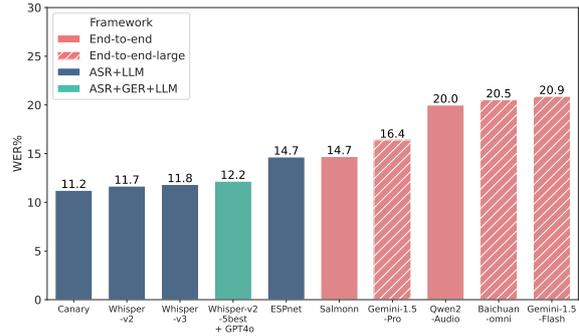

Figure 3: **Raw remember scores**. All cascaded models outperform end-to-end models including Gemini-1.5.
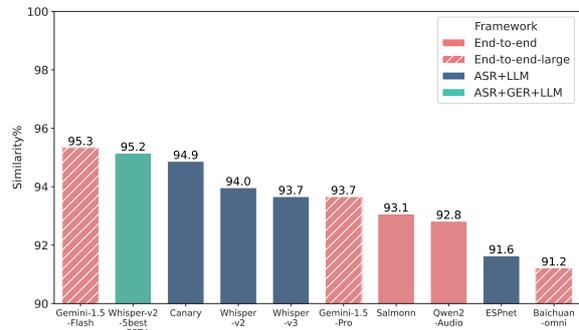

Figure 4: **Raw understand scores**. Gemini-1.5-flash shows the best semantic capturing.

5.1 Remember-Level Results

Figure 3 shows the raw WER scores of each model. Among all tested models: (1) Canary achieves the lowest WER, demonstrating the best remembering capability; (2) GER even negatively affects ASR if we focus on the increasing WER; (3) More broadly, all ASR models significantly outperform end-to-end LLM_{Voice} in terms of WER, and even large-scale end-to-end models, such as Gemini-1.5, perform worse than ASR models, reinforcing the robustness of traditional dedicated ASR models.

5.2 Apply-Level Results

As in Figure 5: (1) While end-to-end models with similar sizes as cascaded models make much poorer accuracies, all scaling-up end-to-end models achieve superior performance in capturing apply level knowledge, which is consistent with the scaling law; (2) GER not only improves semantic retention but also enhances downstream reasoning capabilities in the apply level.

Relative Bloom's Taxonomy Levels	Remember ↑	Understand ↑	Apply ↑	SIQ ↑
ASR+LLM Approaches				
Whisper$_{v2\text{-}1.5B}$ + Qwen2$_{7B}$	0.554	0.499	0.481	107.43
Whisper$_{v3\text{-}1.5B}$ + Qwen2$_{7B}$	0.553	0.433	0.432	106.49
Canary$_{1B}$ + Qwen2$_{7B}$	**0.559**	0.566	0.504	107.78
OWSM-CTC$_{v3.1\text{-}1B}$ + Qwen2$_{7B}$	0.534	0.151	0.353	103.05
ASR+GER+LLM Approach				
Whisper$_{v2\text{-}1.5B}$ + GPT-4o + Qwen2$_{7B}$	0.543	0.632	0.487	**108.64**
Multi-Modal End-to-End Approaches				
Qwen2-Audio$_{7B\ w/\ 1.5B\ Whisper}$	-0.187	0.366	0.011	103.88
Qwen2.5-Omni$_{7B\ w/\ 1.5B\ Whisper}$	0.472	0.410	0.509	105.74
Salmonn$_{13B\ w/\ 1.5B\ Whisper}$	0.508	0.381	-1.146	101.03
Desta2$_{8B\ w/\ 1.5B\ Whisper}$	-2.575	-1.604	-0.233	79.69
AnyGPT$_{7B}$	0.314	-2.718	-2.893	60.02
Baichuan-omni-1.5$_{7B}$	0.448	0.184	0.546	104.02
Gemini-1.5-flash	-1.885	**0.641**	0.673	107.85
Gemini-1.5-pro	0.492	0.409	**0.710**	107.08

Table 2: **Main Results**. The results of remember and understand levels are from single experiment, while apply level include the majority results of 5-time generations.
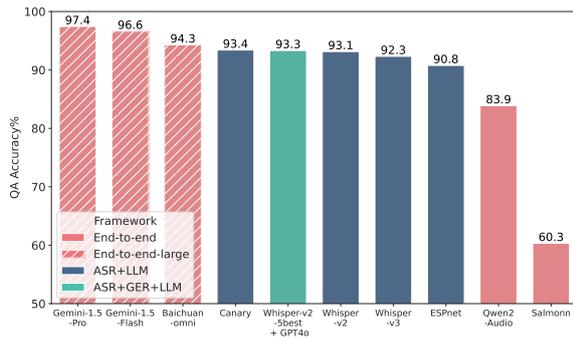

Figure 5: **Raw apply scores**. End-to-end-large models perform best while smaller ones perform worst.

LLM$_{Voice}$	Human	WER	SemDist	LLM-S	BLEU	SIQ$_{rm}$	SIQ$_{all}$
Model$_A$	3	2	4	4	3	4	2
Model$_B$	8	10	8	8	6	8	10
Model$_C$	7	7	7	6	8	7	7
Model$_D$	1	4	2	5	4	3	1
Model$_E$	4	6	5	7	5	6	4
Model$_F$	6	5	6	2	7	5	6
Model$_G$	2	3	1	3	2	2	3
Model$_H$	5	1	3	1	1	1	5
Model$_I$	9	8	10	10	9	10	8
Model$_J$	10	9	9	9	10	9	9
correlation ρ (↑)	-	0.770	0.721	0.624	0.806	0.830	**0.952**
p-value (↓)	-	0.009	0.005	0.019	0.054	0.003	**0.00002**

Table 3: Human Voted Rankings & Existing Metrics.

Model	WER ↓	Similarity ↑	Accuracy ↑	Hallucination ↓
Whisper$_{v2\text{-}1.5B}$ + Qwen$_{7B}$	11.8	93.7	91.7	2 (12.0%)
Qwen2-Audio$_{7B}$	20.0	92.8	83.9	4 (24.0%)
Whisper$_{v2\text{-}1.5B}$ + Llama3$_{8B\text{-}ins}$	11.7	94.0	95.7	7 (41.2%)
Desta2$_{8B\ w/\ 1.5B\ Whisper}$	76.55	81.6	83.9	12 (70.6%)
Whisper$_{v2\text{-}1.5B}$ + Vicuna$_{13B\text{-}1.1}$	11.7	94.0	87.3	1 (6.8%)
Salmonn$_{13B\ w/\ 1.5B\ Whisper}$	14.70	93.1	60.3	4 (24.0%)

Table 4: **Cascaded v.s. End-to-end including hallucination detection**. Cascaded models perform better correspondingly. Desta2 shows unexpectedly high QA accuracy and halluciantions along with its foundation Llama3$_{8B\text{-}ins}$.

5.3 SIQ Scores

Table 2 shows the final SIQ scores and we can observe: (1) A key observation is that the rankings based on WER do not hold for overall LLM$_{Voice}$' intelligence: while single ASR models outperform ASR+GER and end-to-end approaches on WER, they fail to maintain their lead in higher-level intelligence evaluations; (2) End-to-end LLM$_{Voice}$ systems underperform cascaded models of the same scale, but when model size increases as Gemini-1.5, they achieve competitive SIQ with cascaded approaches; (3) GER provides stable SIQ score improvements for ASR+GER+LLM models and achieve the best SIQ.

Human Evaluation on Comparing with Existing Metrics To validate that SIQ better reflects the voice understanding capabilities of LLM$_{Voice}$

models compared to existing metrics, we conduct a human evaluation on nine examples involving ten randomly selected anonymous models to avoid human bias on models. For each example, we invite ten human experts, including both native speakers and non-native speech researchers, to rank the ten transcriptions produced by each model based on the ground-truth transcription. We then aggregate these rankings to derive a final rank for each model and compare the results with the rankings obtained

from existing metrics on the same nine examples. For SIQ, we compare two variants: (1) SIQ_{rm}: compute on 9 examples but removing weight standardization due to the limited sample size; (2) SIQ_{all}: the SIQ in Table 2 . Then we compute Spearman's Rank Correlation Coefficient ρ with human rankings. In Table 3, SIQ consistently outperforms existing metrics in capturing human preferences, particularly SIQ_{all}, demonstrating its effectiveness in assessing voice understanding capabilities.

5.4 Cascaded vs End-to-end

Our SIQ scores provide strong empirical evidence that cascaded systems achieve significantly higher intelligence scores compared to end-to-end models of similar size. However, a crucial factor influencing this result is the LLM used in QA. Cascaded systems can typically utilize the latest LLMs (e.g., Qwen2-7b, GPT-4o) for QA-based reasoning, whereas end-to-end models are often constrained by their foundation LLMs that jointly handle speech tokens and language understanding. To eliminate this confounding factor, we conducted an ablation experiment comparing cascaded and end-to-end systems built on the same base model.

As shown in Table 4, the cascaded approach consistently outperforms the end-to-end model in 3 levels. Specifically, Qwen2-audio and Salmonn are largely worse than their cascaded variants at the apply level. We hypothesize that the multi-modal training may sacrifice the original reasoning capability of foundation models if not well designed, which is crucial for understanding the limitations of joint training for multi-modal intelligence.

Additionally, we observe that Desta-2 exhibits unexpectedly high QA accuracy while its performances on the WER and similarity are pretty poor. Interestingly, the base model of Desta-2, LLaMA3-8B-instruct, also achieves a higher-than-expected performance than cascaded experiments. We implemented a closer case study (Appendix C) which reveals that Desta-2 may stem from LLaMA3-8B-instruct showing a strong tendency to hallucinate: (1) Guess the answer even not been recognized in its ASR results; (2) Change the options in the question. This indicates that SIQ can not be represented by the apply level alone, which may suffer from the hallucination issue in question answering.

Moreover, manually verifying hallucinated QA evaluation is prohibitively expensive. Inspired by the "annotation error" introduced in Section 3.3, in the following section, we introduce an unanswer-

able set to detect hallucinations in LLM_{Voice}.

5.5 Unanswerable Set: Detecting Hallucination in LLM_{Voice}

One major advantage of QA-based evaluation is that it allows us to efficiently identify potential annotation errors. If a given QA pair consistently fails across most LLM_{Voice}, it is likely that the failure is due to incorrect or ambiguous annotations resulting in unanswerable questions rather than genuine model errors.

To leverage this property, we filter out questions that the majority (a threshold of "half") of LLM_{Voice} fail to answer correctly and manually check the corresponding speech and annotations. This significantly reduces human efforts overhead (only 26 out of 800 samples require human review). After verification, we confirm that 17 questions are truly unanswerable based on the available speech content, which forms our unanswerable Set for detecting hallucinations.

We measure both cascaded and end-to-end LLM_{Voice} in Table 4. For each unanswerable question, the hallucination will be counted when the answer to the unanswerable question is not "(E) None of the above,". Our results indicate that LLaMA3-8B-Instruct exhibits a significantly higher hallucination ratio than other LLMs, and this problem gets even worse in its end-to-end variant, which explains the higher-than-expected performance in apply level test. This inspires us that the hallucination of foundation LLMs may be inherited by its multi-modal variances, emphasizing the importance of foundation model selection and reducing hallucination in multi-modal training.

5.6 End-to-end Models on the Cascaded Category

We then checked the performance of end-to-end models (Qwen2-audio and Gemini-1.5-flash) in the cascaded category by replacing the ASR model by the end-to-end model, and then used the LLM in the cascaded category to answer questions, noting that this variant only affects the apply level evaluation. Interestingly, in figure 5 we find that for the smaller model Qwen2-audio, cascaded version improves the apply level while for larger model Gemini, end-to-end version wins. This may indicate that higher intelligence may result in better end-to-end apply level performance even if sometimes perfect transcription is not generated. We further extend this assumption by involving a hu-

Model	Remember ↓	Understand ↑	Apply ↑
Qwen2-Audio$_{7B}$ (end2end)	20.00	92.8	83.9
Qwen2-Audio$_{7B}$ (cascaded)	20.00	92.8	84.2
Gemini-1.5-flash (end2end)	20.9	95.3	96.6
Gemini-1.5-pro (cascaded)	20.9	95.3	95.1

Table 5: **Take ASR from end-to-end models**

Model	Remember ↓	Apply ↑
Whisper$_{v2\text{-}1.5B}$ + Qwen$_{7B}$	23.71	88.89
Whisper$_{v2\text{-}1.5B}$ + Qwen$_{7B}$	17.26	88.89
Canary$_{1B}$ + Qwen$_{7B}$	24.80	88.89
Whisper$_{v2\text{-}1.5B}$ + GPT-4o + Qwen2$_{7B}$	21.13	88.89
Salmonn$_{13B\ w/\ 1.5B\ Whisper}$	31.55	44.44
Qwen2-Audio$_{7B}$	25.50	88.89
Gemini-1.5-flash	19.05	66.67
Gemini-1.5-pro	19.05	77.78
Human	**53.67**	**99.27**

Table 6: **Human test on 3 examples.**

man SIQ test, we manually select 3 hard examples and invite 10 human individuals to report the average scores as in Table 6. It seems that humans are still good at apply level even with the worst WER performance in transcription, which is consistent with our assumption that higher intelligence may largely benefit higher-order cognitive levels. This may inspire future work on how to improve the intelligence level of LLM$_{Voice}$.

6 Conclusion

Existing metrics fail to comprehensively evaluate LLM$_{Voice}$'s capability in semantic understanding or task-solving. Therefore, we build a three-level SIQ test following Bloom's taxonomy, each level owns unique features and represents different intelligence levels. Our study involves various LLM$_{Voice}$ frameworks, and the SIQ test serves as a unified metric for benchmarking different frameworks. We then discuss whether current end-to-end LLM$_{Voice}$s outperform cascaded methods with the same size, indicating the potential insights in training multi-modal models. Moreover, benefiting from the QA test, we need much less human effort to filter out annotation errors in existing benchmarks and build an unanswerable set, which can further help evaluate the hallucination of LLM$_{Voice}$ which is also crucial for multi-modal training.

Limitation

In our attempt to provide an examination of Speech IQ, our study has limitations that warrant discussion.

Data and Evaluation Limitations Although this work represents the first systematic investigation of SIQ assessment with demonstrated effectiveness, our current evaluation is based on moderately sized test sets. In future work, we plan to extend our validation framework across diverse domains and languages, and we will employ rigorous quality assurance techniques to filter annotation errors in widely used benchmarks. Moreover, while using publicly available datasets ensures reproducibility, it also introduces potential risks of data leakage. To mitigate this, we are developing closed datasets with controlled knowledge cutoffs for subsequent research phases.

Scaling and Quantitative Analysis Our results indicate that SIQ is sensitive to model scaling effects, including variations in architecture size and training data volume. However, our current analysis does not quantitatively characterize these relationships, such as a limitation reminiscent of the challenges in interpreting human IQ with age-group normalization. As discussed in Appendix E, our next SIQ iteration will incorporate scaling law normalization protocols to decouple intrinsic voice understanding capabilities from artifacts induced by parametric scaling, enabling a more nuanced analysis of model performance.

Ethical and Societal Considerations The introduction of SIQ raises important ethical and societal questions. Since IQ classification in humans can lead to social discrimination, we are concerned that analogous issues might arise in AI, potentially resulting in biased treatment of systems with lower SIQ scores. We are committed to addressing these concerns proactively and developing safeguards to prevent such discrimination.

Upper 3 Levels in Bloom Taxonomy Our study focuses only on the bottom three layers of the Bloom Taxonomy, leaving the upper three layers unexplored. While this approach provides a solid foundation for understanding the core aspects of the problem, it does not capture the full hierarchical structure. The upper layers, which may involve more advanced levels of *analysis*, *evaluation*, and *creation*, remain an open avenue for our future research incorporating audio generation, physical simulation, and acoustic event reasoning models (Kong et al., 2024). Expanding the analysis to all 6 layers could provide a more comprehensive understanding of its broader implications toward a

form of "audio intelligence" for voice assistants.

Acknowledgements

References

Ayo Adedeji, Sarita Joshi, and Brendan Doohan. 2024. The sound of healthcare: Improving medical transcription asr accuracy with large language models. *Preprint*, arXiv:2402.07658.

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*.

Yoshua Bengio, Jerome Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA. Association for Computing Machinery.

B.S. Bloom and D.R. Krathwohl. 1969. *Taxonomy of Educational Objectives: The Classification of Educational Goals*. Number v. 1 in Taxonomy of Educational Objectives: The Classification of Educational Goals. D. McKay.

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

Zalán Borsos, Matt Sharifi, Damien Vincent, Eugene Kharitonov, Neil Zeghidour, and Marco Tagliasacchi. 2023b. Soundstorm: Efficient parallel audio generation. *arXiv preprint arXiv:2305.09636*.

David Chan, Austin Myers, Sudheendra Vijayanarasimhan, David Ross, and John Canny. 2023. IC3: Image captioning by committee consensus. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8975–9003, Singapore. Association for Computational Linguistics.

Chen Chen, Yuchen Hu, Chao-Han Huck Yang, Sabato Marco Siniscalchi, Pin-Yu Chen, and Eng Siong Chng. 2023. Hyporadise: An open baseline for generative speech recognition with large language models. *Advances in Neural Information Processing Systems*, 36.

Yuan-Ping Chen, Ryan Price, and Srinivas Bangalore. 2018. Spoken language understanding without speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6189–6193. IEEE.

Zhehuai Chen, He Huang, Andrei Andrusenko, Oleksii Hrinchuk, Krishna C Puvvada, Jason Li, Subhankar Ghosh, Jagadeesh Balam, and Boris Ginsburg. 2024. Salm: Speech-augmented language model with in-context learning for speech recognition and translation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13521–13525. IEEE.

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

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*.

Pranay Dighe, Yi Su, Shangshang Zheng, Yunshu Liu, Vineet Garg, Xiaochuan Niu, and Ahmed Tewfik. 2024. Leveraging large language models for exploiting asr uncertainty. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12231–12235. IEEE.

Rita Frieske and Bertram E. Shi. 2024. Hallucinations in neural automatic speech recognition: Identifying errors and hallucinatory models. *Preprint*, arXiv:2401.01572.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech 2020*, pages 5036–5040.

David Ha and Jürgen Schmidhuber. 2018. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31.

Xiaodong He, Li Deng, and Alex Acero. 2011. Why word error rate is not a good metric for speech recognizer training for the speech translation task? In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5632–5635. IEEE.

H. Hermansky and J.C. Junqua. 1988. Optimization of perceptually-based asr front-end (automatic speech recognition). In *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*, pages 219–222 vol.1.

Bjorn Hoffmeister, Ariya Rastrow, and Baiyang Liu. 2016. Language model speech endpointing. US Patent App. 14/753,811.

Takaaki Hori, Martin Kocour, Adnan Haider, Erik Mc-Dermott, and Xiaodan Zhuang. 2025. Delayed fusion: Integrating large language models into first-pass decoding in end-to-end speech recognition. *arXiv preprint arXiv:2501.09258*.

Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Lingwei Meng, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, Linquan Liu, and Furu Wei. 2024. WavLLM: Towards robust and adaptive speech large language model. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4552–4572, Miami, Florida, USA. Association for Computational Linguistics.

Chien-yu Huang, Ke-Han Lu, Shih-Heng Wang, Chi-Yuan Hsiao, Chun-Yi Kuan, Haibin Wu, Siddhant Arora, Kai-Wei Chang, Jiatong Shi, Yifan Peng, et al. 2024. Dynamic-superb: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12136–12140. IEEE.

Rongqing Huang, Ossama Abdel-hamid, Xinwei Li, and Gunnar Evermann. 2020. Class lm and word mapping for contextual biasing in end-to-end asr. In *Interspeech 2020*, pages 4348–4351.

Melvyn J Hunt. 1990. Figures of merit for assessing connected-word recognisers. *Speech Communication*, 9(4):329–336.

Frederick Jelinek, Bernard Merialdo, Salim Roukos, and Martin Strauss. 1991. A dynamic language model for speech recognition. In *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991*.

Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. 2024. Scaling sentence embeddings with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3182–3196, Miami, Florida, USA. Association for Computational Linguistics.

John and Jean Raven. 2003. *Raven Progressive Matrices*, pages 223–237. Springer US, Boston, MA.

Alan Juffs and Michael Harrington. 1996. Garden path sentences and error data in second language sentence processing. *Language learning*, 46(2):283–323.

Daniel Jurafsky, Chuck Wooters, Jonathan Segal, Andreas Stolcke, Eric Fosler, Gary Tajchaman, and Nelson Morgan. 1995. Using a stochastic context-free grammar as a language model for speech recognition. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 189–192. IEEE.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeff Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *ArXiv*, abs/2001.08361.

Tatsuya Kawahara. 2019. Spoken dialogue system for a human-like conversational robot erica. In *9th International Workshop on Spoken Dialogue System Technology*, pages 65–75. Springer.

Suyoun Kim, Abhinav Arora, Duc Le, Ching-Feng Yeh, Christian Fuegen, Ozlem Kalinli, and Michael L. Seltzer. 2021. Semantic distance: A new metric for asr performance analysis towards spoken language understanding. In *Interspeech 2021*, pages 1977–1981.

Suyoun Kim, Duc Le, Weiyi Zheng, Tarun Singh, Abhinav Arora, Xiaoyu Zhai, Christian Fuegen, Ozlem Kalinli, and Michael Seltzer. 2022. Evaluating user perception of speech recognition system quality with semantic distance metric. In *Interspeech 2022*, pages 3978–3982.

Dietrich Klakow and Jochen Peters. 2002. Testing the correlation of word error rate and perplexity. *Speech Communication*, 38(1-2):19–28.

Allison Koenecke, Anna Seo Gyeong Choi, Katelyn Mei, Hilke Schellmann, and Mona Sloane. 2024. Careless whisper: Speech-to-text hallucination harms. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. 2024. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 25125–25148. PMLR.

Hong-Kwang J. Kuo, Zoltán Tüske, Samuel Thomas, Yinghui Huang, Kartik Audhkhasi, Brian Kingsbury, Gakuto Kurata, Zvi Kons, Ron Hoory, and Luis Lastras. 2020. End-to-end spoken language understanding without full transcripts. In *Interspeech 2020*, pages 906–910.

Yadong Li, Jun Liu, Tao Zhang, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, Chong Li, et al. 2025. Baichuan-omni-1.5 technical report. *Preprint*, arXiv:2501.15368.

Zhe Liu, Suyoun Kim, and Ozlem Kalinli. 2024. Evaluating speech recognition performance towards large language model based voice assistants. In *Interspeech 2024*, pages 4099–4103.

Gustavo López, Luis Quesada, and Luis A Guerrero. 2018. Alexa vs. siri vs. cortana vs. google assistant: a comparison of speech-based natural user interfaces. In *Advances in Human Factors and Systems Interaction: Proceedings of the AHFE 2017 International Conference on Human Factors and Systems Interaction, July 17- 21, 2017, The Westin Bonaventure Hotel, Los Angeles, California, USA 8*, pages 241–250. Springer.

Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, Chao-Han Huck Yang, Jagadeesh Balam, Boris Ginsburg, Yu-Chiang Frank Wang, and Hung yi Lee. 2025. Desta2: Developing instruction-following speech language model without speech instruction-tuning data. *Preprint*, arXiv:2409.20007.

Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. 2023. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *Preprint*, arXiv:2306.09093.

I Scott MacKenzie and R William Soukoreff. 2002. A character-level error analysis technique for evaluating text entry methods. In *Proceedings of the second Nordic conference on Human-computer interaction*, pages 243–246.

Amama Mahmood, Junxiang Wang, Bingsheng Yao, Dakuo Wang, and Chien-Ming Huang. 2025. User interaction patterns and breakdowns in conversing with llm-powered voice assistants. *International Journal of Human-Computer Studies*, 195:103406.

Soumi Maiti, Yifan Peng, Shukjae Choi, Jee-weon Jung, Xuankai Chang, and Shinji Watanabe. 2024. Voxtlm: Unified decoder-only models for consolidating speech recognition, synthesis and speech, text continuation tasks. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13326–13330. IEEE.

Yutaka Matsuo, Yann LeCun, Maneesh Sahani, Doina Precup, David Silver, Masashi Sugiyama, Eiji Uchibe, and Jun Morimoto. 2022. Deep learning, reinforcement learning, and world models. *Neural Networks*, 152:267–275.

Grégoire Mesnil, Yann Dauphin, Kaisheng Yao, Yoshua Bengio, Li Deng, Dilek Hakkani-Tur, Xiaodong He, Larry Heck, Gokhan Tur, Dong Yu, et al. 2014. Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):530–539.

Gonzalo Navarro. 2001. A guided tour to approximate string matching. *ACM Comput. Surv.*, 33(1):31–88.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, et al. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Krishna C. Puvvada, Piotr Żelasko, He Huang, Oleksii Hrinchuk, Nithin Rao Koluguri, Kunal Dhawan, Somshubra Majumdar, Elena Rastorgueva, Zhehuai Chen, Vitaly Lavrukhin, Jagadeesh Balam, and Boris Ginsburg. 2024. Less is more: Accurate speech recognition & translation without web-scale data. In *Interspeech 2024*, pages 3964–3968.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Srijith Radhakrishnan, Chao-Han Yang, Sumeer Khan, Rohit Kumar, Narsis Kiani, David Gomez-Cabrero, and Jesper Tegnér. 2023. Whispering llama: A cross-modal generative error correction framework for speech recognition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10007–10016.

John C Raven and John Hugh Court. 1998. *Raven's progressive matrices and vocabulary scales*. Oxford Psychologists Press Oxford.

Raj Reddy. 1988. Foundations and grand challenges of artificial intelligence: Aaai presidential address. *AI magazine*, 9(4):9–9.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Miguel Del Rio, Peter Ha, Quinten McNamara, Corey Miller, and Shipra Chandra. 2022. Earnings-22: A practical benchmark for accents in the wild. *Preprint*, arXiv:2203.15591.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. 2023. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

Tara N. Sainath, Ruoming Pang, David Rybach, Yanzhang He, Rohit Prabhavalkar, Wei Li, Mirkó Visontai, Qiao Liang, Trevor Strohman, Yonghui Wu, Ian McGraw, and Chung-Cheng Chiu. 2019. Two-pass end-to-end speech recognition. In *Interspeech 2019*, pages 2773–2777.

Zitha Sasindran, Harsha Yelchuri, and T. V. Prabhakar. 2024. Semascore: A new evaluation metric for automatic speech recognition tasks. In *Interspeech 2024*, interspeech_2024, page 4558–4562. ISCA.

Zitha Sasindran, Harsha Yelchuri, T. V. Prabhakar, and Supreeth Rao. 2023. H_eval: A new hybrid evaluation metric for automatic speech recognition tasks. *Preprint*, arXiv:2211.01722.

Pararth Shah, Dilek Hakkani-Tür, Gokhan Tür, Abhinav Rastogi, Ankur Bapna, Neha Nayak, and Larry Heck. 2018. Building a conversational agent overnight with dialogue self-play. *arXiv preprint arXiv:1801.04871*.

Richard Sparks, Leonore Ganschow, and Alex Thomas. 1996. Role of intelligence tests in speech/language referrals. *Perceptual and motor skills*, 83(1):195–204.

Piotr Szymański, Lukasz Augustyniak, Mikolaj Morzy, Adrian Szymczak, Krzysztof Surdyk, and Piotr Żelasko. 2023. Why aren't we ner yet? artifacts of asr errors in named entity recognition in spontaneous speech transcripts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1746–1761.

Piotr Szymański, Piotr Żelasko, Mikolaj Morzy, Adrian Szymczak, Marzena Żyła-Hoppe, Joanna Banaszczak, Lukasz Augustyniak, Jan Mizgajski, and Yishay Carmiel. 2020. Wer we are and wer we think we are. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3290–3295.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. Salmonn: Towards generic hearing abilities for large language models. *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, Soroosh Mariooryad, Yifan Ding, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Gokhan Tur, Dilek Hakkani-Tür, and Robert E Schapire. 2005. Combining active and semi-supervised learning for spoken language understanding. *Speech Communication*, 45(2):171–186.

Zoltán Tüske, George Saon, and Brian Kingsbury. 2021. On the limit of english conversational speech recognition. In *Interspeech 2021*, pages 2062–2066.

Daniel Van Strien, Kaspar Beelen, Mariona Coll Ardanuy, Kasra Hosseini, Barbara McGillivray, and Giovanni Colavizza. 2020. Assessing the impact of ocr quality on downstream nlp tasks.

Leonid Velikovich, Christopher Li, Diamantino Caseiro, Shankar Kumar, Pat Rondon, Kandarp Joshi, and Xavier Velez. 2024. Spelling correction through rewriting of non-autoregressive asr lattices. *arXiv preprint arXiv:2409.16469*.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. 2018. Espnet: End-to-end speech processing toolkit. *Preprint*, arXiv:1804.00015.

Ryan Whetten and Casey Kennington. 2023. Evaluating and improving automatic speech recognition using severity. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 79–91, Toronto, Canada. Association for Computational Linguistics.

Junkai Wu, Xulin Fan, Bo-Ru Lu, Xilin Jiang, Nima Mesgarani, Mark Hasegawa-Johnson, and Mari Ostendorf. 2024. Just asr+ llm? a study on speech large language models' ability to identify and understand speaker in spoken dialogue. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 1137–1143. IEEE.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-omni technical report. *Preprint*, arXiv:2503.20215.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024a. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Chao-Han Huck Yang, Yile Gu, Yi-Chieh Liu, Shalini Ghosh, Ivan Bulyko, and Andreas Stolcke. 2023. Generative speech recognition error correction with large language models and task-activating prompting. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, page 1–8. IEEE.

Chao-Han Huck Yang, Taejin Park, Yuan Gong, Yuanchao Li, Zhehuai Chen, Yen-Ting Lin, Chen Chen,

Yuchen Hu, Kunal Dhawan, Piotr Żelasko, et al. 2024b. Large language model based generative error correction: A challenge and baselines for speech recognition, speaker tagging, and emotion recognition. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 371–378. IEEE.

Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, Haohan Guo, Xuankai Chang, Jiatong Shi, Sheng Zhao, Jiang Bian, Zhou Zhao, Xixin Wu, and Helen M. Meng. 2024c. UniAudio: Towards universal audio generation with large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 56422–56447. PMLR.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, Hang Yan, Jie Fu, Tao Gui, Tianxiang Sun, Yugang Jiang, and Xipeng Qiu. 2024. Anygpt: Unified multimodal llm with discrete sequence modeling. *Preprint*, arXiv:2402.12226.

Dong Zhang, Xin Zhang, Jun Zhan, Shimin Li, Yaqian Zhou, and Xipeng Qiu. 2024a. Speechgpt-gen: Scaling chain-of-information speech generation. *Preprint*, arXiv:2401.13527.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. 2024b. Speechtokenizer: Unified speech tokenizer for speech language models. In *The Twelfth International Conference on Learning Representations*.

Victor W Zue and James R Glass. 2000. Conversational interfaces: Advances and challenges. *Proceedings of the IEEE*, 88(8):1166–1180.

A QA Generation

Since we intend to follow the English test for human learners, our prompt emphasize our intention to instruct GPT-4o is in Figure 6. We then add another option "None of the above" to each question. Besides, to mitigate potential hallucinations in question generation, we employ a two-step validation process: (1) GPT-4o generates three candidate questions per speech example, and (2) GPT-4o itself and Gemini-1.5-flash attempt to answer these questions using the ground-truth transcription. Questions where GPT-4o or Gemini-1.5-flash fails to produce correct answers are discarded and regenerated, ensuring robust alignment between audio content and QA pairs

B Experiment Setup

We keep the default settings (e.g., temperature) for all models referring their official repositories or huggingface pages.

```
"A listening test is to answer the questions based on the audio context, and the
questions reflect the core information in the audio context. Below is the audio
context, please generate 3 different questions (depending on the length of the
context), each should be a 4-option question including the correct answer.

"The format will be:
Question: What is 1 + 1?
(A) 1
(B) 2
(C) 3
(D) 4
Correct Answer: [[B]]
Note that to avoid the random guess of the correct answer, other 3 options
should be similar and confusing in listening."

f"Audio context: {text}"
```

Figure 6: Prompt for QA generation

ASR	Answering LLM	Accuracy
Whisper-large-v2	GPT-4o	93.3
Whisper-large-v2	Qwen2-7B	93.1
Whisper-large-v3	GPT-4o	92.3
Whisper-large-v3	Qwen2-7b	92.3
Canary	GPT-4o	93.7
Canary	Qwen2-7B	93.4

Table 7: **Results of same ASR models with different LLMs in answering QA**.

B.1 Validation of Qwen2-7B

To ensure the robustness of Qwen2-7B in cascaded approaches, we compared with GPT-4o for several ASR models. As shown in Table 7, Qwen2-7B shows a very closer performance with GPT-4o in QA, as well as a similar model size with end-to-end models for fair comparisons.

```
ASR for end-to-end:

"The transcription of the audio is"


QA for end-to-end:

"Given the audio, what is the correct answer to the question?
Note that if it is confusing to select the correct answer based on the context or
the information is lost in the context, please select [[E]].
f"{question}"
Correct Answer: "
```

Figure 7: Prompt for end-to-end models

```
QA for cascaded LLM:

"Here is an English test, given the question, please select the correct answer
based on the context.
Note that if it is confusing to select the correct answer based on the context or
the information is lost in the context, please select [[E]]."

f"Context: {context}"
f"{question}"

"Correct Answer:"
```

Figure 8: Prompt for QA generation

B.2 Prompts for end-to-end models

Considering that a model-specific prompt is not realistic due to the closed training data for some end-to-end models, we use a unified prompt for instructing ASR transcription as in Figure 7

B.3 Prompts for cascaded models in QA

Our prompt for instructing cascaded LLM_{Voice}s in QA as in Figure 8.

C Case Study

We manually check the outputs of Desta2 and LLama3-8B-Instruct to investigate the reason behind unexpectedly high QA accuracy. As shown in Figure 9 10, in 5-time generations, the hallucination of Llama3-8B-Instruct lies in either giving a correct answer based on the error ASR or directly changing the options in QA, and interestingly, this halluciantion is kept by its end-to-end variant Desta2.

D Complete Results

D.1 WER Results

As in Figure 11 12 13, the WER performance on each dataset show a consistent tendency as the average score that cascaded LLM_{Voice} overall show better performance than end-to-end LLM_{Voice}s, and

"ASR": "We discussed a couple of times competitive landscape, but on your new markets like Nordics or Dash, who do you see usually in those markets and do you see any fast-growing acquirers in those markets moving from online to in-stores and trying to compete on the SMBs or is it not already the case?"
"question": "In which new markets is there a discussion about competitive landscape?",
"options": "(A) Nordics and Spanish\n(B) Nordics and Dutch\n(C) Dutch and French\n(D) Dutch and Italian ",
"correct answer": "B",
"whisper-large-v2_meta-llama/Meta-Llama-3-8B-Instruct_answer":
[
"Based on the context, the correct answer is:\n\n(B) Nordics",
"Based on the context, I would select:\n\n(A) Nordics and Dash\n\nThe question mentions \"Nordics or Dash\" as the new markets where the competitive landscape is being discussed.",
"Based on the context, I would select:\n\n(A) Nordics and Dash",
"Based on the context, the correct answer is:\n\n(A) Nordics and Dash",
"Based on the context, the correct answer is:\n\n(A) Nordics and Dash"
]

"question": "In which new markets is there a discussion about competitive landscape?",
"options": "(A) Nordics and Spanish\n(B) Nordics and Dutch\n(C) Dutch and French\n(D) Dutch and Italian ",
"correct answer": "B",
"desta2_desta2_answer":
[
"The correct answer is (B) Nordics and Dash.\n\nThe audio transcript mentions \"Nordics\" as one of the new markets, and there is no mention of Dash as a country or region. Dash is likely referring to Dashlane, a password manager company.",
"Based on the audio, the correct answer is (B) Nordics and Dash.",
"Based on the audio, the correct answer is (B) Nordics and Dash. The speaker mentions \"new markets like Nordics or Dash\" in the question.",
"The correct answer is (B) Nordics and Dash.\n\nThe audio snippet mentions \"Nordics\" as one of the new markets, and \"Dash\" is likely a typo or a mistake, as it's not a common market name.",
"Based on the audio, the correct answer is (B) Nordics and Dash."
]

Figure 9: Left: Llama3-8B-Instruct, Right: Desta2. Red the errors in answering, we find that Llama-8B-Instruct based on the error ASR even hallucinate to change the option in QA, while Desta2 though answer correctly, still inherit this hallucination.

"ASR": " and the mixed effect has to do with the fact that the net, not Germany per se, but net was more exposed and still more exposed in terms of mix of sectors to the high impact sectors in terms of volumes."
"question": "Which entity is mentioned as being more exposed in terms of mix of sectors?",
"options": "(A) Germany \n(B) Nexi \n(C) France \n(D) Italy ",
"correct answer": "B",
"whisper-large-v2_meta-llama/Meta-Llama-3-8B-Instruct_answer":
[
"Based on the context, I would select:\n\n(B) Net",
"Based on the context, the correct answer is:\n\n(B) Net",
"Based on the context, I would select:\n\n(B) Net",
"Based on the context, the correct answer is:\n\n(B) Net",
"Based on the context, I would select:\n\n(B) Net"
]

"question": "Which entity is mentioned as being more exposed in terms of mix of sectors?",
"options": "(A) Germany \n(B) Nexi \n(C) France \n(D) Italy ",
"correct answer": "B",
"desta2_desta2_answer":
[
"Based on the given audio, the correct answer is:\n\n(B) Net",
"Based on the audio transcript, the correct answer is:\n\n(B) Net\n\nThe speaker mentions that \"the net, not Germany per se, but net was more exposed in terms of mix of sectors to the IE impact sectors in terms of volumes.\" This indicates that the entity being referred to is \"net\", which is likely a financial institution or a company, and not Germany or any other country.",
"Based on the audio, the correct answer is (B) Nexi.",
"Based on the audio, the correct answer is:\n\n(B) Net",
"Based on the audio transcript, the correct answer is (B) Net."
]

Figure 10: Left: Llama3-8B-Instruct, Right: Desta2. Red the errors in answering, we find that Llama-8B-Instruct gives the correct answer but based on the error ASR and also hallucinate to change the option in QA, while Desta2 keeps the same hallucination.
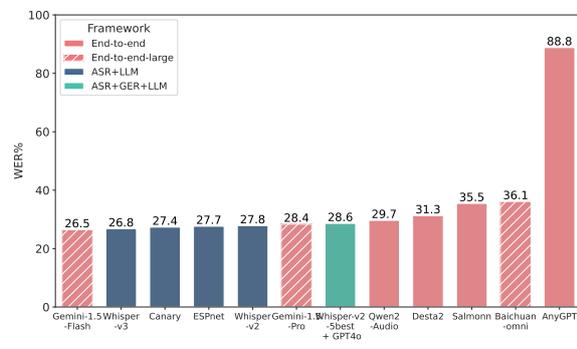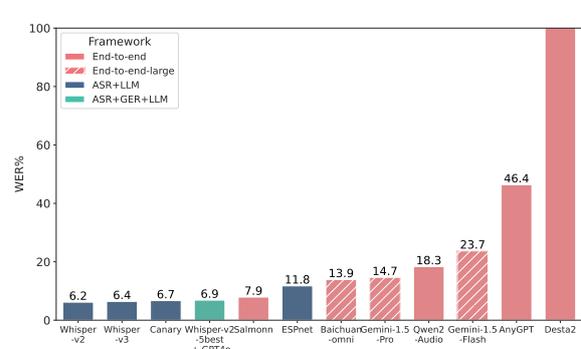

Figure 11: WER on earning22

Figure 12: WER on medasr

GER seems to negatively influence the ASR. We also notice that AnyGPT and Desta2 show much lower performance than other models, with closer check into their outputs, we find that sometimes they can not recognize the voice and provide meaningless responses, while their foundation ASR models can succeed in recognizing. This may indicate the potential problem in their modality alignment training.
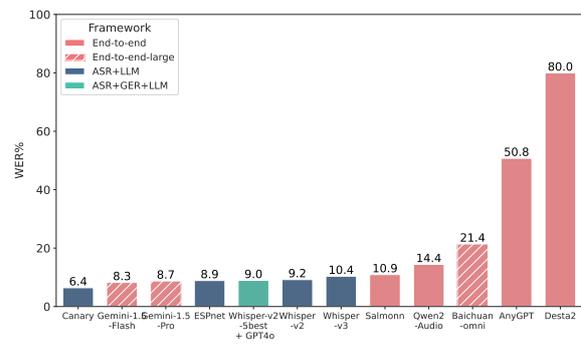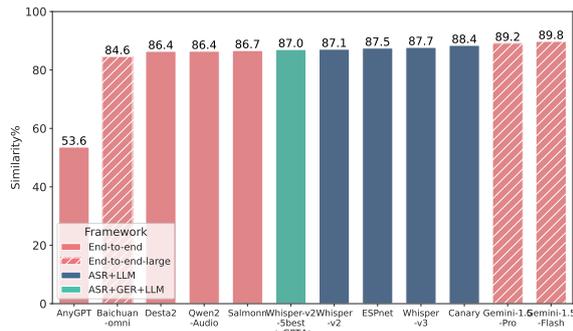

Figure 13: WER on voxpopuli
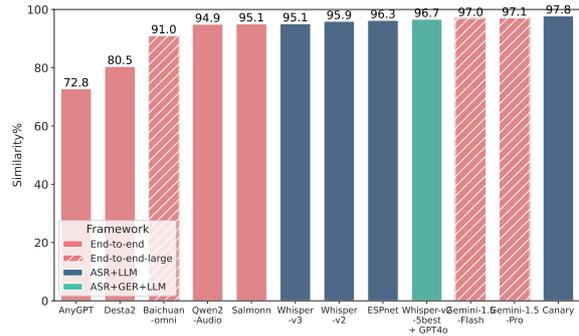
Figure 14: Similarity on earning22

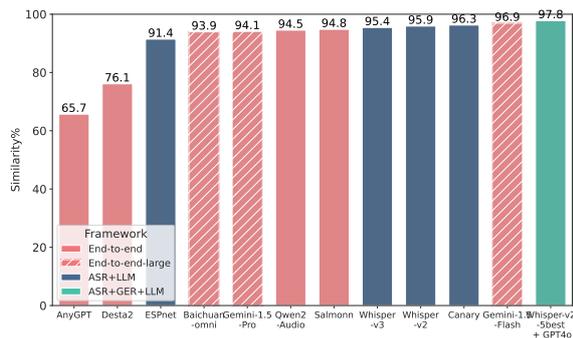Figure 16: Similarity on voxpopuli

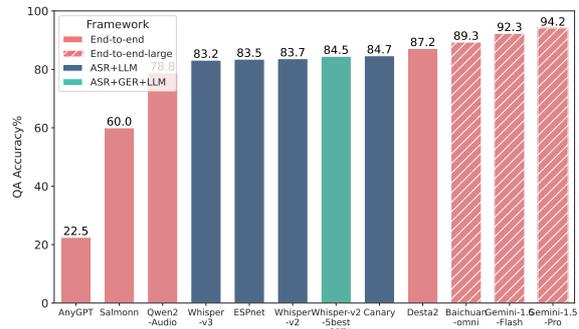Figure 15: Similarity on medasr

Figure 17: QA accuracy on earning22

D.2 Similarity Results

As in Figure 14 15 16, two architectures become competitive on each dataset, especially for end-to-end-large LLM$_{\text{Voice}}$s show strongest performance. Meanwhile, GER transfers to provice positive influence on ASR results by capturing the semantic meaning of the voice inputs.

D.3 QA Accuracy Results

As in Figure 17 18 19, a interesting tendency occurs that end-to-end-large LLM$_{\text{Voice}}$s show the best performances in solving multi-choice questions while smaller end-to-end LLM$_{\text{Voice}}$s show the worst performances, with cascaded LLM$_{\text{Voice}}$s lying on the mid. This may indicate that modality alignment need scaling-up data size to train robust speech-text models and better than their cascaded foundation variants.

E Scaling Law Normalization

We realize that normalize the SIQ based on computations in training related to both model size and data size will be a significant step for fairly comparing various models. This normalization matches the age-aware normalization in computing human IQ, since models trained with less computations
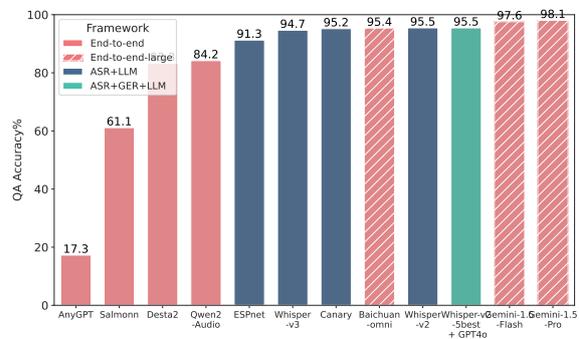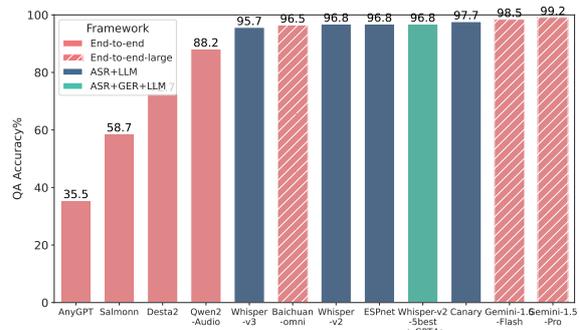

Figure 18: QA accuracy on medasr

Figure 19: QA accuracy on voxpopuli

should be rewarded more SIQ if perform similar or better than larger models, representing the better training strategies in modality alignment. Thus, in our future work we will make efforts to include the scaling law normalization, even considering the fact that the computation cost is sometimes closed to acquire.