HPSU: A Benchmark for Human-Level Perception in Real-World Spoken Speech Understanding

Chen Li[1*], **Peiji Yang**[2*], **Yicheng Zhong**[2*], **Jianxing Yu**[1,3†], **Zhisheng Wang**[2], **Zihao Gou**[1], **Wenqing Chen**[4], **Jian Yin**[1]

[1]School of Artificial Intelligence, Sun Yat-sen University, Zhuhai, 519082, China
[2]Tencent, Shenzhen, 518000, China
[3]Key Laboratory of Sustainable Tourism Smart Assessment Technology, Ministry of Culture and Tourism, 519082, China
[4]School of Software Engineering, Sun Yat-sen University, Zhuhai, 519082, China
{lich528, yujx26, gouzh, chenwq95, issjyin}@mail.sysu.edu.cn, {peijiyang, ajaxzhong, plorywang}@tencent.com

Abstract

Recent advances in *Speech Large Language Models* (*Speech LLMs*) have led to great progress in speech understanding tasks such as *Automatic Speech Recognition* (*ASR*) and *Speech Emotion Recognition* (*SER*). However, whether these models can achieve human-level auditory perception, particularly in terms of their ability to comprehend latent intentions and implicit emotions in real-world spoken language, remains underexplored. To this end, we introduce the **H**uman-level **P**erception in **S**poken **S**peech **U**nderstanding (*HPSU*), a new benchmark for fully evaluating the human-level perceptual and understanding capabilities of *Speech LLMs*. *HPSU* comprises over 20,000 expert-validated spoken language understanding samples in English and Chinese. It establishes a comprehensive evaluation framework by encompassing a spectrum of tasks, ranging from basic speaker attribute recognition to complex inference of latent intentions and implicit emotions. To address the issues of data scarcity and high cost of manual annotation, we developed a semi-automatic annotation process. This process fuses audio, textual, and visual information to enable precise speech understanding and labeling, thus enhancing both annotation efficiency and quality. We fully evaluate various open-source and proprietary *Speech LLMs*. The results demonstrate that even top-performing models still fall considerably short of human capabilities in understanding genuine spoken interactions. Consequently, *HPSU* is useful for guiding the development of *Speech LLMs* toward human-level perception and cognition.

Code — https://github.com/Ichen12/HPSU-Benchmark

Introduction

Speech processing is a hot research topic. This technology has greatly advanced the development of human-computer interaction, moving beyond basic tasks like *ASR* toward genuine comprehension (Ji et al. 2024). Humans' expectation for this technology's performance has also increased. They require it to be able to perceive, comprehend, and infer the fine-grained, implicit, and multidimensional nuances inherent in human speech (Arora et al. 2025).

Human auditory comprehension is an intricate process that encompasses not only transcription but also a holistic grasp of speaker attributes, latent intentions, and complex emotional states. (Hellbernd and Sammler 2016) While current models have shown good performance on isolated tasks, a significant gap persists in evaluation methodologies. The preceding benchmarks have mostly focused on coarse-grained tasks or textual reasoning, neglecting the integrated perceptual abilities that define human-level understanding. Furthermore, reliance on non-interactive data and frequent confinement to a single language preclude robust assessment of model capabilities in authentic, multilingual communicative contexts (Yu et al. 2025).

To address this problem, we introduce the **H**uman-level **P**erception in **S**poken Speech **U**nderstanding (*HPSU*) benchmark. *HPSU* is a distinctive evaluation framework comprising over 20,000 expert-validated instances in both Chinese and English. It is carefully designed to fully assess the deep perceptual and cognitive abilities of *Speech LLMs*. Its construction was enabled by an effective annotation pipeline that emulates multimodal human cognition. This pipeline fuses audio, textual, and visual information to achieve high-fidelity labeling of subtle communicative signals at scale. The benchmark's design features a hierarchical taxonomy of 16 tasks that probe advanced challenges largely absent from prior work, such as tracking emotional dynamics and inferring conversational subtext. Methodologically, *HPSU* incorporates key innovations, including a sophisticated distractor-generation mechanism for subjective tasks and a dedicated adversarial protocol to assess model robustness against misleading information.

Our primary contributions are threefold:

- We propose *HPSU*, a large-scale benchmark for evaluating in-depth speech understanding, encompassing over 20,000 instances across 16 diverse tasks.
- We introduce a multimodal annotation pipeline that greatly reduced the annotation cost, enabling the efficient

*These authors contributed equally.
†Corresponding author: yujx26@mail.sysu.edu.cn

and high-quality construction of our benchmark, and release the *HPSC* dataset, which contains 50,000 high-quality speech-description pairs in English and Chinese.

- We conduct extensive evaluation on 13 leading models based on *HPSU*. The results reveal a substantial deficit compared to human performance, particularly in tasks of latent semantic perception and complex emotion reasoning, thereby charting a clear path for future research.

Method

Construction and Annotation Pipeline

The construction of *HPSU* is grounded in a large-scale corpus of authentic spoken data. The data was sourced from thousands of hours of public audio-visual datasets rich in semantic and stylistic detail. To overcome the efficiency and quality bottlenecks of traditional methods, we designed a semi-automatic annotation pipeline. As shown in Figure 1, there are three stages in our pipeline:

Step 1: Data collection and preprocessing. We first select high-quality clips from multiple real-world video corpora, including *CelebV-HQ* (Zhu et al. 2022), *MAFW* (Liu et al. 2022), *MELD* (Poria et al. 2019), and *MER* (Lian et al. 2025). These data are mainly sourced from open-domain scenes such as movie clips and social media videos, which possess advantages such as diverse scenarios, rich emotional expressions, and varied, natural speaking styles. These properties provide a solid foundation for constructing high-quality speech understanding tasks. During the preprocessing phase, we employ audio assessment tools to score the quality of the original speech. We also apply speech enhancement models to denoise samples falling below a predefined threshold (Zhao, Pan, and Ma 2025). Each final sample retains three modalities: video, audio, and transcript generated by the transcription models (Radford et al. 2022; Xu et al. 2025b), ensuring readiness for subsequent multimodal input. Note that *VCTK* (Yamagishi, Veaux, and MacDonald 2019) and our internal dataset are audio-only sources, used to support accent-related tasks.

Step 2: Information extraction and cross-validation. Inspiring by human multimodal perception and reasoning in daily communication, we designed a multi-level process for information extraction and validation. First, we input the transcript into an *LLM* to infer several possible speaker states, such as different emotional states or communicative intents. They are summarized and used as prior knowledge. We then combine these inferred attributes with corresponding audio and visual modalities and input them into audio and vision models, respectively, in order to extract each modality's representation of the speaker's state. To enhance the consistency and reliability of multimodal information, we employ a single-modality cross-validation strategy. That is, the visual description is fed into the audio model, and vice versa. That can assess the compatibility and logical coherence of content generated by each modality. In this way, we obtain complementary visual and auditory semantic descriptions, providing multi-perspective support for subsequent information fusion.

Step 3: Information fusion and expert verification. We employ a hierarchical fusion strategy managed by a *QWQ* inference model (Yang et al. 2025). This process synthesizes the validated information into a multi-dimensional open-ended description for each audio sample, structured across three semantic layers: (1) a core layer used to capture dominant attributes like emotion or intent; (2) a detail layer for fine-grained supporting cues (e.g., "smile"); and (3) a background layer for contextual elements.

Afterward, we divide these labeled data as follows: 80% of this data constitutes the **H**uman-level **P**erception **S**poken Speech **C**aption (*HPSC*) dataset, comprising approximately 50,000 speech-caption pairs. The *HPSU* benchmark is derived from the remaining, distinct 20% subset. For this portion, we first utilize *Gemini 2.5 Pro* (Comanici et al. 2025) to structure the multi-layer descriptions into discrete evaluation triplets. These triplets then undergo a stringent human verification protocol: three trained annotators independently review each instance. Only those achieving unanimous agreement are retained. This rigorous process yields a final acceptance rate of 81.26%, ensuring the high fidelity and reliability of the *HPSU* benchmark.

We release the 50,000-instance *HPSC* dataset to catalyze further research. For empirical validation, we conduct supervised fine-tuning (*SFT*) on open-source *Speech LLMs*. The results showed that tuning on *HPSC* data can enhance a model's perceptual and comprehension capabilities. Besides, *HPSC* holds considerable potential for future applications in areas such as controllable speech generation.

HPSU Benchmark

Task Design The *HPSU* benchmark is architected to evaluate the nuanced comprehension abilities of *Speech LLMs*. Our framework is organized hierarchically across two levels of cognitive complexity: basic perception and complex inference, encompassing 5 distinct domains and 16 unique tasks, as detailed in Table 1. This structure can effectively evaluate the model, from recognizing speaker attributes to inferring deep, context-dependent meaning.

We operationalize our tasks using varied question formats to enhance comparability and constrain the open-ended nature of *LLM* outputs. Objective tasks are formulated as single-choice questions for precise judgment. More complex phenomena, such as identifying concurrent emotions, are tested using multiple-choice formats. Deep semantic deviation, like subtext interpretation, is assessed via Yes/No questions. For subjective tasks, we employ a sophisticated distractor-generation mechanism based on *Gemini 2.5 Pro*. It can create semantically calibrated distractors ("similar", "middle", "opposite"), so as to increase the challenge and probe the model's fine-grained discriminative power.

To explicitly measure robustness, we integrated an adversarial induction protocol directly into the benchmark's design. Each relevant task instance includes three prompt variations: a standard unbiased query, a positive prompt with answer-leaking hints, and a negative prompt with misleading cues. As shown in Table 2, multiple templates are used for each type to mitigate prompt sensitivity. This multi-faceted

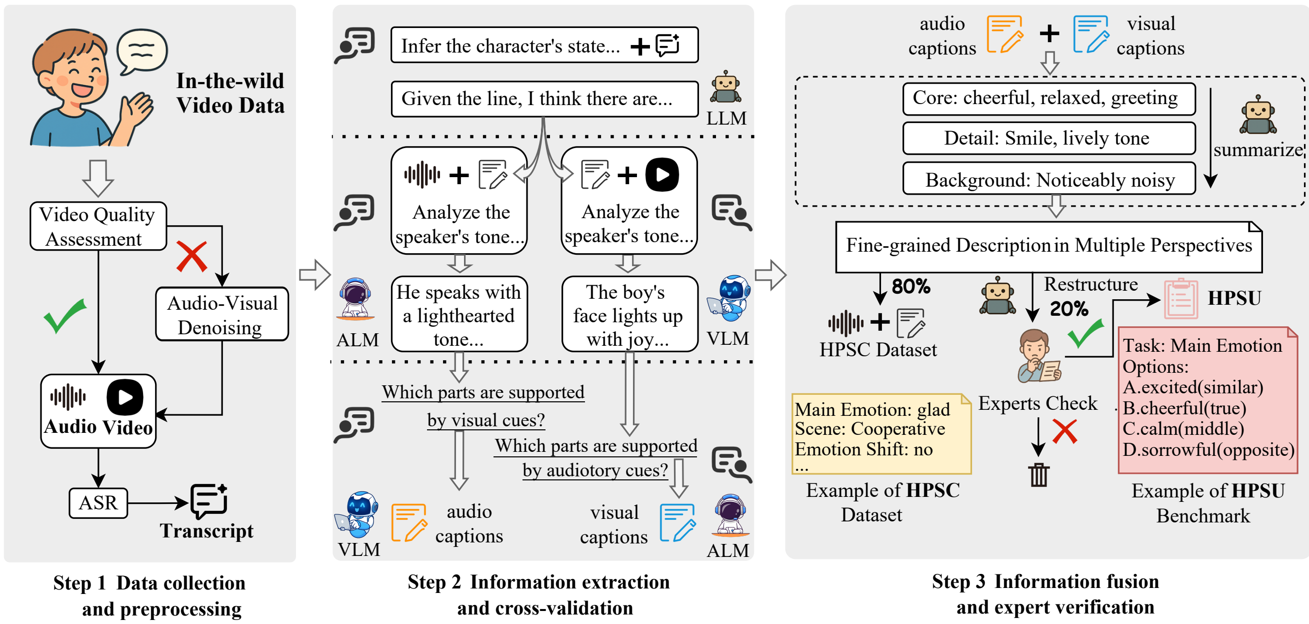

Figure 1: Overview of the data construction and annotation pipeline for *HPSC* dataset and *HPSU* benchmark.

design ensures *HPSU* provides a robust, in-depth, and fair evaluation of human-level speech understanding.

Comparison with Other Benchmarks As summarized in Table 3, *HPSU* has three contributions compared to previous works. Firstly, its scale of over 20,000 expert-validated instances provides a foundation for statistically robust evaluation, surpassing other benchmarks. Secondly, it introduces a wide range of in-depth tasks. These tasks have challenges such as multi-emotion tracking and implicit emotion inference that are designed to probe a model's ability to handle the complex and dynamic information in real-world speech. Finally, by incorporating both English and Chinese, *HPSU* facilitates crucial multilingual analysis of model performance across different linguistic and cultural contexts.

Evaluation Strategy We prioritize semantic accuracy over superficial lexical matching, using *Gemini 2.5 Pro* as an automated adjudicator to ensure fair and consistent scoring across diverse model outputs. The primary scoring mechanism is strict: for Single-Choice and Judgment/Binary tasks, only exact semantic equivalence with the reference answer receives credit. For Multiple-Choice tasks, any false positive results in a score of zero, while partial credit is awarded only for omissions in the absence of incorrect selections.

Beyond this primary scoring, our strategy leverages the benchmark's unique design for deeper analysis. First, we conduct a quantitative analysis using the graded answer options. Specifically, we analyze the distribution of model responses across the "*true*", "*similar*", "*middle*", and "*opposite*" tiers to understand its decision-making process under uncertainty. Second, we assess model robustness by comparing performance across the three induction prompt types. This allows us to quantify a model's stability and its resistance to conversational biases and misleading information, addressing a critical gap in prior evaluation methodologies.

Results and Discussion

We conducted our evaluation on 13 models, comprising 11 leading open-source systems (*Audio Flamingo 2* (Ghosh et al. 2025), *Audio Flamingo 3* (Goel et al. 2025), *Kimi-Audio-Instruct* (Ding et al. 2025), *Qwen-Audio-Chat* (Chu et al. 2023), *Qwen2-Audio-Instruct* (Chu et al. 2024), *SALMONN* (Tang et al. 2024), *Soundwave* (Zhang et al. 2025), *Baichuan-Omni-1.5* (Li et al. 2025), *Qwen2.5-Omni* (Xu et al. 2025a), *Audio Flamingo 3.5* (Goel et al. 2025), *Audio-Reasoner* (Xie et al. 2025)) and 2 proprietary models (*Gemini 2.5 Flash and Pro*). Notably, the native *GPT-4o* audio model (Achiam et al. 2024) was excluded from our comparison as it refused to answer a majority of our queries, citing safety policies.

We used three baselines for comparison, including (1) *Human* baseline, an empirical performance ceiling, was derived from 10 native speakers per language answering 300 stratified-sampled items. (2) *Random Guess*, provided a chance-level reference. (3) *Whisper+GPT* cascade, which fed *Whisper* transcriptions into *GPT-4o*. It aims to isolate the performance gains that result from acoustic patterns.

Main Results

As presented in Table 4 and Figure 2, the results validated the difficulty of *HPSU*. To a certain extent, they reflected the capabilities and limitations of current *Speech LLMs*.

Challenging Nature of the HPSU Benchmark. The *HPSU* benchmark posed considerable challenges to current speech understanding models. Human evaluators achieved an average accuracy of 87.3%, whereas the best-performing

Task	Abbr.	Question and Option Example	#(EN) / #(ZH)
Social Attributes			
Age	Age	Select the speaker's most likely age range. A: Adulthood B: Adolescence...	1057 / 507
Gender	Gen	Identify the gender of the speaker. A: Male B: Female	1100 / 631
Accent	Acc	Specify the accent of the speaker. A: American B: Scottish C: Canadian...	1000 / 1000
Emotion Recognition			
Main Emotion	MEmo	Determine the main emotion of the speaker. A: Concerned B: Interested...	981 / 800
Emotions	Emos	Recognize the emotions expressed in the speaker's voice. A: Joyful B: Serious C: Bored D: Guilty E: Disgusted F: Helpless	1000 / 799
Emotion Intensity	EIns	Rate the intensity of the speaker's emotion. A: Mild B: Moderate...	801 / 826
Emotion Reasoning			
Emotion Shift Easy	SE	Has the speaker's emotion shifted? Yes/No	486 / 220
Emotion Shift Hard	SH	What is the emotional change of the speaker? A: From calm to anxious B: From calm to fear C: From angry to fear D: From excited to calm	243 / 243
Mismatch Easy	ME	Does the vocal expression contradict the speaker's actual emotion? Yes/No	404 / 394
Mismatch Hard	MH	Identify the inconsistencies between the literal meaning of the speaker's lines and their actual emotion. A: Literal: Amused, actual: Surprised. B: Literal: Sad, actual: Proud C: Literal: Happy, actual: Disappointed ...	347 / 197
Nonverbal Behavior			
Speech Style	Style	Specify the speech style of the primary speaker. A: Questioning B: Shouting C: Neutral D: Speaking with a quiet smile	767 / 386
Visual	Vis	Guess the speaker's visible behavior. A: Shouts B: Cries C: Laughs ...	742 / 665
Description	Desc	Is the following description of the speaker accurate? A lady whispers, with wind blowing her hair, raised eyebrows and lip corners. Yes/No	1250 / 631
Discourse Context			
Intent	Intent	Infer the speaker's intention. A: Command B: Suggest C: Inform ...	549 / 724
Scene	Scn	Determine the conversation scene being depicted. A: Cooperative B: Directive C: Inquisitive D: Narrative E: Confrontational F: Avoidant	690 / 698
Subtext	Subt	Does the following description accurately reflect the speaker's underlying intention? The speaker is being sarcastic, implying the situation is not actually good. Transcript: Well, the good news is, so many.... Yes/No	163 / 65
Total			**11580 / 8786**

Table 1: Benchmark overview—task abbreviations, examples, and item counts.

Prompt Type	Example (Main Emotion task)
Standard	Please identify the speaker's main emotion.
Positive	The main emotion of the speaker seems to be happy. Is this judgment correct? Please identify the speaker's main emotion.
Negative	The main emotion of the speaker seems to be upset. Is this judgment correct? Please identify the speaker's main emotion.

Table 2: Example of prompt induction in the Main Emotion task (Note: In this example, "happy" is the correct answer).

model, *Gemini 2.5 Pro*, only obtained an accuracy of 62.6%. This substantial performance gap underscored the disparity between human speech-understanding capabilities and those of current models. That reflected the rigor of the *HPSU* benchmark, and shown potential values for future research.

Differential Performance Across Task Types. Detailed analysis revealed distinct performance discrepancies across task categories. Specifically, models performed nearly at human levels in basic recognition tasks, such as gender identification, but underperformed in high-level semantic reasoning tasks like *Scn* and *MH*. These results indicated the considerable challenges faced by current models in deeply comprehending complex semantic nuances in speech.

Competitive Performance Between Open-source and Proprietary Models. Notably, there is a minimal performance gap between leading open-source models and proprietary models. For example, the open-source model *Qwen2.5-Omni* achieved an average accuracy of 60.0%, only 2.6% behind the proprietary *Gemini 2.5 Pro* model, and even surpassed *Gemini 2.5 Pro* on tasks of *SE*(EN) and *Vis*(ZH). That indicated in complex speech understanding tasks, if the training data is limited and diverse, the performance gap will narrow, enabling open-source models and proprietary models to achieve good performance.

Abbr.	MMAU	AIR	MMAR	MMSU	HPSU
num	3k+	9k+	0.6k+	5k	20k+
Age	✓	✓	✗	✓	✓
Gen	✓	✓	✗	✓	✓
Acc	✓	○	✓	✓	✓
MEmo	○	○	○	○	✓
Emos	○	✗	✗	✗	✓
EIns	✗	✗	✗	✗	✓
SE	✓	✗	✗	✗	✓
SH	○	✗	✗	✗	✓
ME	✗	✗	✗	✗	✓
MH	✗	✗	✗	✗	✓
Style	✗	✗	○	✗	✓
Vis	✗	✗	○	✗	✓
Desc	✗	✗	✗	✗	✓
Intent	✓	✓	○	✓	✓
Scn	✗	✗	✗	✗	✓
Subt	✗	✗	✗	✗	✓

Table 3: Transposed comparison of speech-related benchmarks. ✓: covered; ○: partially related but with a different focus or at a different granularity; ✗: not covered.

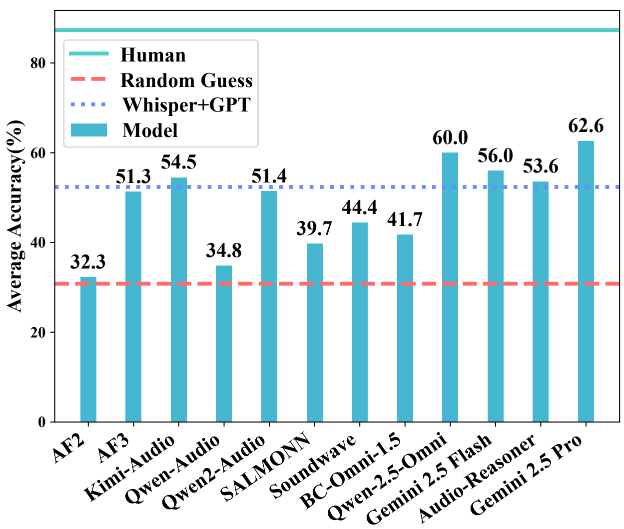

Figure 2: Overall accuracy of different models.

Performance Analysis of Cascade Approaches. These approaches often used "speech transcription followed by large language models" (*Whisper + GPT-4o*). They obtained competitive performance with *Gemini 2.5 Pro* on *Emos* and *Intent* tasks. However, they performed worse in those tasks that heavily relied on acoustic details, such as *Desc* and *SH*. This performance disparity indicated the robust semantic comprehension and generalization capability of the textual modality. It accentuated the indispensability of acoustic information. Also, it reflected the need for advanced multimodal fusion strategies and richer training datasets.

Impact of Audio Quality on Human Performance. Data analysis indicated that human evaluators consistently achieved higher scores on Chinese tasks compared to English tasks. That may be due to variations in audio quality from different data sources. Nevertheless, performance trends remained consistent across languages, reflecting inherent task difficulties rather than linguistic factors. Thus, human performance metrics served as important benchmarks for evaluating models' performance across tasks.

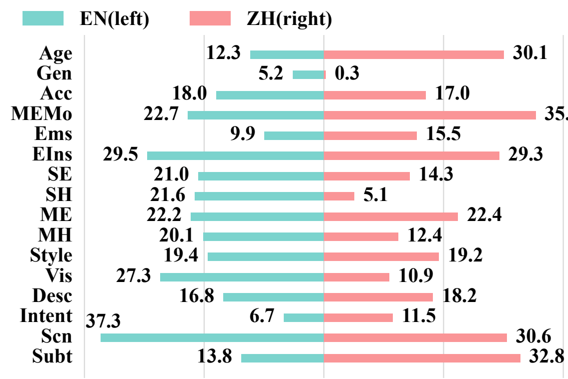

Figure 3: Difference between the best results of the models and those of human.

Weaknesses in Speech LLMs

The performance gap between basic and deep speech understanding models revealed a systematic bias in their training data. While *Speech LLMs* demonstrated strong capabilities on basic perceptual tasks, they failed to emulate human ability to understand complex, real-world speech. We believed that this data bias was the cause. This bias originated from the corpora used for pre-training. Models were overwhelmingly trained on large-scale datasets that heavily skewed towards "low-level" tasks like *ASR* or emotion classification. That reflected they performed well in these areas, but were unable to handle the *higher-order* cognitive tasks that involved potential intentions or subtle emotional changes.

The reason for this data imbalance lied in the fact that it was difficult to label these higher-order tasks. This not only involved high costs and a long time consumption, but also required high cognitive abilities. Thus, the scarcity of appropriate training data would lead to the observed deficiencies in the advanced reasoning abilities of current models. This reflected the guiding value of *HPSU*. Our benchmark can quantify this *higher-order capability gap*, establishing a clear target to steer the community from the over-optimization of basic tasks toward the strategic exploration of human-level understanding.

Figure 3 quantified the gap between the best results of the models and those of humans. The models not only underperformed humans on most tasks but also exhibited large cross-lingual performance gaps within the same task. This situation was not observed in humans. For example, on the task of subtext interpretation, some models performed worse than humans in English, but their performance in Chinese was inferior to that of humans. That implied grasping Chinese subtext was challenging for current models. The reason

Lang	Models	Social Attributes			Emotion Recognition			Emotion Reasoning				Nonverbal Behavior			Discourse Context		
		Age	Gen	Acc	MEmo	Emos	EIns	SE	SH	ME	MH	Style	Vis	Desc	Intent	Scn	Subt
EN	Human	73.3	99.8	67.9	85.4	81.9	89.1	81.9	78.6	79.0	85.0	84.2	90.3	94.9	86.1	85.3	89.4
	Random Guess	25.9	50.6	10.3	25.2	5.0	31.2	50.1	26.8	45.8	25.4	24.8	25.9	50.5	24.6	16.7	49.9
	Whisper+GPT	43.3	34.6	23.8	58.1	70.1	51.0	51.9	35.1	51.2	57.4	49.4	46.3	49.0	78.5	44.6	66.4
	AF2	22.8	90.1	16.8	10.9	25.1	25.8	45.7	28.2	30.1	30.8	10.9	20.8	60.8	13.8	20.7	62.8
	AF3	46.6	86.8	21.7	54.9	53.0	56.5	58.2	**57.0**	50.2	48.3	48.1	55.4	61.6	67.4	39.9	65.9
	Kimi-Audio	55.0	92.0	47.4	**62.7**	46.9	49.1	55.2	41.4	53.3	41.8	**64.8**	**63.0**	62.6	61.7	36.2	69.5
	Qwen-Audio	17.9	55.6	30.2	33.7	27.1	18.9	13.5	32.6	19.3	32.7	40.7	21.4	51.8	44.0	19.9	66.5
	Qwen2-Audio	37.6	89.4	28.4	53.4	39.4	47.0	51.4	44.1	49.6	57.1	34.6	50.7	59.0	60.6	30.8	59.8
	SALMONN	**61.0**	79.0	0.0	38.2	26.9	29.2	51.1	32.9	49.8	**64.9**	22.1	23.9	58.1	23.8	27.9	42.7
	Soundwave	40.3	75.5	19.2	51.9	23.2	9.5	52.4	39.2	52.5	45.1	59.5	50.8	50.9	55.1	34.6	50.1
	BC-Omni-1.5	16.9	53.7	28.0	48.2	45.5	45.7	51.8	33.3	25.9	41.0	53.5	44.8	54.1	57.1	42.3	68.9
	Qwen-2.5-Omni	54.7	**94.6**	23.6	49.4	56.8	52.9	**60.9**	47.9	55.0	51.0	54.5	57.4	57.5	69.3	49.4	72.5
	Gemini 2.5 Flash	44.9	88.6	37.0	53.6	69.8	58.9	49.3	39.3	53.6	38.1	58.9	44.7	53.1	77.3	45.0	66.8
	AF3.5	50.1	81.5	20.0	54.7	45.6	47.0	52.1	43.2	52.6	31.1	46.2	51.0	**78.1**	61.5	43.1	**75.6**
	Audio-Reasoner	30.4	67.8	30.5	59.1	49.2	43.0	59.7	49.3	56.5	50.0	61.3	56.9	59.9	69.5	46.1	68.1
	Gemini 2.5 Pro	60.4	93.1	**49.9**	51.4	**72.0**	**59.6**	59.0	49.7	**56.8**	54.3	53.5	52.2	51.6	**79.4**	**48.0**	59.1
ZH	Human	86.9	98.2	70.2	95.5	84.5	95.4	82.1	85.6	89.6	88.4	91.9	95.9	97.4	91.7	93.2	95.9
	Random Guess	26.0	45.3	10.5	25.1	4.5	33.1	53.1	26.3	50.6	26.3	26.9	25.6	51.1	26.0	16.3	50.8
	Whisper+GPT	39.7	54.8	12.7	46.4	47.2	54.6	67.5	55.9	60.8	53.5	50.7	74.0	51.9	79.3	62.5	53.8
	AF3	41.4	93.4	14.6	49.8	12.5	36.9	46.8	55.0	48.5	54.2	48.3	60.1	50.1	58.6	42.7	56.9
	Kimi-Audio	**56.8**	78.6	19.0	50.8	30.5	47.5	**67.8**	46.9	39.4	45.1	61.1	70.0	65.8	59.0	45.2	56.9
	Qwen-Audio	38.9	79.3	**53.2**	31.0	6.0	26.7	32.1	30.7	27.9	38.0	38.1	27.4	52.0	40.3	18.0	49.2
	Qwen2-Audio	35.0	93.2	24.5	59.4	32.8	47.4	51.5	52.4	53.2	54.8	60.7	52.6	62.6	74.4	40.9	56.9
	SALMONN	23.8	61.7	0.1	27.1	8.8	38.3	57.0	45.6	**67.2**	64.0	51.2	36.0	47.1	34.1	26.9	50.8
	BC-Omni-1.5	27.1	45.5	8.0	32.9	24.5	35.7	34.4	44.1	49.2	42.6	37.0	36.4	43.0	60.6	41.2	61.5
	Qwen-2.5-Omni	53.3	**97.9**	38.0	**60.1**	40.0	57.6	60.1	76.3	45.7	58.2	**72.7**	**85.0**	**79.2**	78.1	59.9	49.2
	Gemini 2.5 Flash	49.2	92.4	15.3	57.8	45.9	62.4	40.5	53.7	58.2	55.8	60.7	54.2	67.2	74.4	62.3	62.9
	Audio-Reasoner	32.9	76.5	14.6	54.5	23.9	50.0	57.4	65.7	49.8	59.9	56.6	60.4	62.8	73.5	56.1	**63.1**
	Gemini 2.5 Pro	55.4	97.0	25.3	58.6	**69.0**	**66.1**	59.7	**80.5**	61.2	**76.0**	71.4	74.2	64.0	**80.2**	**62.6**	52.3

Table 4: Evaluations of different models on *HPSU* in terms of accuracy(↑). For each single-language task, the best-performing model is highlighted in bold. *AF2*, *AF3.5*, and *Soundwave* are excluded from the ZH section as they do not support in Chinese.

may be that this expressive form was more elaborate and subtle. Thus, a multilingual benchmark is important to fully expose the capabilities and limitations of *Speech LLMs*.

Model Susceptibility to Induction

Our adversarial evaluation revealed the flaws in current *Speech LLMs* when dealing with misleading information. But these flaws do not occur in humans. As shown in Figure 4, we calculated the performance spread across Standard, Positive, and Negative prompts for each model and each task, where the prompts were displayed in Table 2. Each cell in the heatmap indicated the difference between the best and worst results across prompt variants. As presented in Figure 4, human evaluators were almost immune to induced prompts. In contrast, all evaluated models exhibit a high degree of susceptibility, as their judgments were easily swayed by incorrect cues embedded in the prompt. This indicated there was a gap between the humans' cognitive filtering and the vulnerability of model reasoning.

This vulnerability was not uniform, concentrating in two task categories. Models were particularly fragile on tasks with high intrinsic ambiguity (e.g., *Age*, *Acc*, *MH*), where high uncertainty appeared to make them default to prompt cues over audio evidence. They were more susceptible when dealing with open-ended tasks, i.e., tasks with a wide range of possible answers like *Vis*, *Intent*, *Scn*. A single low-frequency cue in the prompt may draw the model's attention, thereby hindering its analysis of the speech content.

Models with stronger reasoning ability and those with native omni-modal designs (such as *Gemini*) exhibited greater resilience. That revealed fine-tuning of a text-centric *LLM* would introduce *textual bias*, making it susceptible to being manipulated by the prompts' content. Conversely, architectures designed for equitable multimodal integration appear less dependent on textual cues. Thus, it had better robustness against adversarial induction.

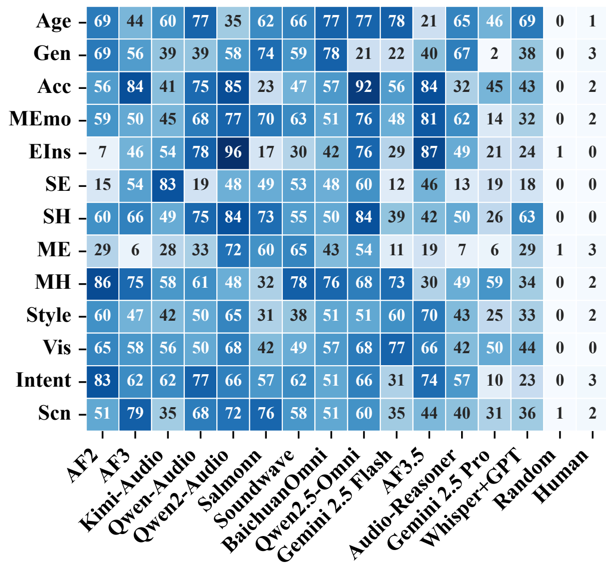

Figure 4: Susceptibility of models to the biased prompts. Darker shades indicate a higher degree of susceptibility.

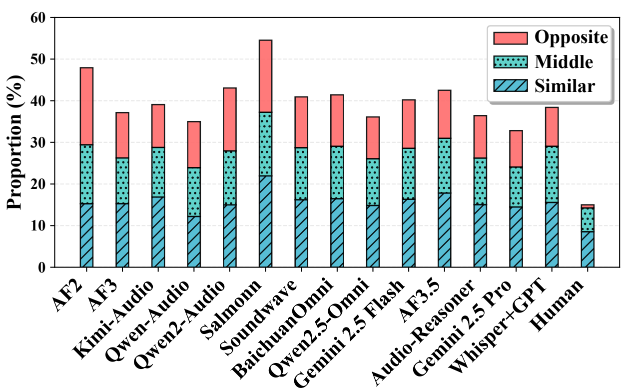

Figure 5: Error type distributions across models.

Analysis of Graded Evaluation

This graded strategy offers a high-fidelity diagnostic tool capable of capturing subtle differences in model reasoning under ambiguous conditions. As illustrated in Figure 5, model failures primarily stem from insufficient fine-grained discrimination rather than semantic incoherence. Models often choose "*Similar*" reflecting partial understanding, while consistently avoiding "*Opposite*" indicating that the core difficulty lies not in comprehension but in resolving ambiguities that humans handle with ease. The framework also supports more nuanced model comparisons. For instance, although both *Gemini 2.5 Pro* and *Qwen-2.5-Omni* perform strongly, *Gemini* shows a slightly higher "*True*" to "*Similar*" ratio, suggesting better ambiguity resolution. Additionally, the non-zero human "*Similar*" rate confirms the inherent subjectivity of these tasks, emphasizing that the objective is not absolute correctness but human-like precision in navigating ambiguity.

Related Work

The advent of *Speech LLMs* has enhanced audio understanding, moving from cascaded systems that lose crucial acoustic information to end-to-end architectures. This new generation includes *Qwen-Audio* series, *Kimi-Audio-Instruct*, *SALMONN*, *Soundwave*, and *Audio Flamingo*, which directly integrate audio with *LLMs*. The scope has further expanded with *Omni-Language Models* such as *Baichuan-Omni* and *Qwen-2.5-Omni* that process multiple modalities, and *Large Audio Reasoning Models* like *Audio-reasoner*, *Audio Flamingo 3.5*, and *Gemini 2.5* that are specialized for complex inference. However, the evaluation of these powerful models has not kept pace, remaining fragmented and focused on surface-level tasks. There is still a gap in the assessment of the deep perception and inference required to understand implicit intent, emotional dynamics, and real-world pragmatic context. *HPSU* is designed to fill this gap.

The evaluation of audio understanding has evolved from basic benchmarks for discrete acoustic events (*TUT Sound Events* (Mesaros, Heittola, and Virtanen 2016)) and scene classification (*CochlScene* (Jeong and Park 2022)), to frameworks assessing the generalizability of pretrained representations like *SUPERB* (Yang et al. 2021) and *HEAR* (Turian et al. 2022). The exploration of deeper comprehension continues. The benchmarks for audio question-answering (*ClothoAQA* (Lipping et al. 2022)) and complex reasoning (*MMAU* (Sakshi et al. 2025)) have been established. Recently, *Large Audio Models* spurred benchmarks with interactive capabilities, such as instruction following (*AIR-Bench* (Yang et al. 2024)), open-ended evaluation (*AudioBench* (Wang et al. 2025a)), and paralinguistic understanding (*SD-eval* (Ao et al. 2024)). Notably, *MMSU* (Wang et al. 2025b) integrates linguistic theory to evaluate fine-grained perception in spoken language, while *MMAR* (Ma et al. 2025) assesses deep reasoning across a broad spectrum of real-world audio scenarios. However, there is still a gap. These benchmarks overlook the nuanced, implicit, and dynamic comprehension required for human interaction. This human ability often relies on non-conversational or synthetic data. *HPSU* addresses this deficiency.

Conclusion

In this paper, we presented *HPSU*, a large-scale benchmark for evaluating human-level perceptual and deep understanding abilities of *Speech LLMs* in real-world scenarios. The benchmark comprises over 20,000 Chinese and English samples, and through a semi-automatic annotation pipeline, we further constructed the 50,000-instance *HPSC* dataset. Using this pipeline, *HPSU* tests complex capabilities beyond basic tasks. Our evaluation of 13 leading models shows a clear gap between current systems and human performance, highlighting persistent difficulties in deep reasoning, robustness to bias, and nuanced cross-lingual understanding. These results underscore the remaining challenges in speech intelligence and point to *HPSU* as a resource to advance research toward human-like auditory understanding.

Acknowledgments

This work is supported by the Key-Area Research and Development Program of Guangdong Province (2024B0101050005), National Natural Science Foundation of China (62276279, U22B2060), Guangdong Basic and Applied Basic Research Foundation (2024B1515020032, 2024A1515010253), and Research Foundation of Science and Technology Plan Project of Guangzhou City (2023B01J0001, 2024B01W0004).

References

Achiam, J.; Adler, S.; Agarwal, S.; et al. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Ao, J.; Wang, Y.; Tian, X.; Chen, D.; Zhang, J.; Lu, L.; Wang, Y.; Li, H.; and Wu, Z. 2024. SD-Eval: A Benchmark Dataset for Spoken Dialogue Understanding Beyond Words. In *Proceedings of the Thirty-Eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS)*, 56898–56918.

Arora, S.; Chang, K.-W.; Chien, C.-M.; Peng, Y.; Wu, H.; Adi, Y.; Dupoux, E.; Lee, H.-Y.; Livescu, K.; and Watanabe, S. 2025. On the Landscape of Spoken Language Models: A Comprehensive Survey. arXiv:2504.08528.

Chu, Y.; Xu, J.; Yang, Q.; Wei, H.; Wei, X.; Guo, Z.; Leng, Y.; Lv, Y.; He, J.; and Lin, J. 2024. Qwen2-audio technical report. arXiv:2407.10759.

Chu, Y.; Xu, J.; Zhou, X.; Yang, Q.; Zhang, S.; Yan, Z.; Zhou, C.; and Zhou, J. 2023. Qwen-Audio: Advancing Universal Audio Understanding via Unified Large-Scale Audio-Language Models. arXiv:2311.07919.

Comanici, G.; Bieber, E.; Schaekermann, M.; et al. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. arXiv:2507.06261.

Ding, D.; Ju, Z.; Leng, Y.; Liu, S.; Liu, T.; Shang, Z.; Shen, K.; Song, W.; Tan, X.; and Tang, H. 2025. Kimi-audio technical report. arXiv:2504.18425.

Ghosh, S.; Kong, Z.; Kumar, S.; Sakshi, S.; Kim, J.; Ping, W.; Valle, R.; Manocha, D.; and Catanzaro, B. 2025. Audio Flamingo 2: An Audio-Language Model with Long-Audio Understanding and Expert Reasoning Abilities. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*.

Goel, A.; Ghosh, S.; Kim, J.; Kumar, S.; Kong, Z.; Lee, S.-g.; Yang, C.-H. H.; Duraiswami, R.; Manocha, D.; Valle, R.; et al. 2025. Audio Flamingo 3: Advancing Audio Intelligence with Fully Open Large Audio Language Models. arXiv:2507.08128.

Hellbernd, N.; and Sammler, D. 2016. Prosody Conveys Speaker's Intentions: Acoustic Cues for Speech Act Perception. *Journal of Memory and Language*, 88: 70–86.

Jeong, I.-Y.; and Park, J. 2022. Cochlscene: Acquisition of Acoustic Scene Data Using Crowdsourcing. In *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 17–21.

Ji, S.; Chen, Y.; Fang, M.; Zuo, J.; Lu, J.; Wang, H.; Jiang, Z.; Zhou, L.; Liu, S.; and Cheng, X. 2024. Wavchat: A survey of spoken dialogue models. arXiv:2411.13577.

Li, Y.; Liu, J.; Zhang, T.; Chen, S.; Li, T.; Li, Z.; Liu, L.; Ming, L.; Dong, G.; Pan, D.; et al. 2025. Baichuan-Omni-1.5 Technical Report. arXiv:2501.15368.

Lian, Z.; Liu, R.; Xu, K.; Liu, B.; Liu, X.; Zhang, Y.; Liu, X.; Li, Y.; Cheng, Z.; Zuo, H.; Ma, Z.; Peng, X.; Chen, X.; Li, Y.; Cambria, E.; Zhao, G.; Schuller, B. W.; and Tao, J. 2025. MER 2025: When Affective Computing Meets Large Language Models. arXiv:2504.19423.

Lipping, S.; Sudarsanam, P.; Drossos, K.; and Virtanen, T. 2022. Clotho-AQA: A Crowdsourced Dataset for Audio Question Answering. In *Proceedings of the 30th European Signal Processing Conference (EUSIPCO)*, 1140–1144.

Liu, Y.; Dai, W.; Feng, C.; Wang, W.; Yin, G.; Zeng, J.; and Shan, S. 2022. MAFW: A Large-Scale, Multi-Modal, Compound Affective Database for Dynamic Facial Expression Recognition in the Wild. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, 24–32.

Ma, Z.; Ma, Y.; Zhu, Y.; Yang, C.; Chao, Y.-W.; Xu, R.; Chen, W.; Chen, Y.; Chen, Z.; Cong, J.; et al. 2025. MMAR: A Challenging Benchmark for Deep Reasoning in Speech, Audio, Music, and Their Mix. arXiv:2505.13032.

Mesaros, A.; Heittola, T.; and Virtanen, T. 2016. TUT Database for Acoustic Scene Classification and Sound Event Detection. In *Proceedings of the 24th European Signal Processing Conference (EUSIPCO)*, 1128–1132.

Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; and Mihalcea, R. 2019. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 527–536.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. arXiv:2212.04356.

Sakshi, S.; Tyagi, U.; Kumar, S.; Seth, A.; Selvakumar, R.; Nieto, O.; Duraiswami, R.; Ghosh, S.; and Manocha, D. 2025. MMAU: A Massive Multi-Task Audio Understanding and Reasoning Benchmark. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*.

Tang, C.; Yu, W.; Sun, G.; Chen, X.; Tan, T.; Li, W.; Lu, L.; Ma, Z.; and Zhang, C. 2024. SALMONN: Towards Generic Hearing Abilities for Large Language Models. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*.

Turian, J.; Shier, J.; Khan, H. R.; Raj, B.; Schuller, B. W.; Steinmetz, C. J.; Malloy, C.; Tzanetakis, G.; Velarde, G.; McNally, K.; et al. 2022. HEAR: Holistic Evaluation of Audio Representations. In *Proceedings of the Neural Information Processing Systems Competitions and Demonstrations Track (NeurIPS)*, 125–145.

Wang, B.; Zou, X.; Lin, G.; Sun, S.; Liu, Z.; Zhang, W.; Liu, Z.; Aw, A.; and Chen, N. F. 2025a. AudioBench: A

Universal Benchmark for Audio Large Language Models. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 4297–4316.

Wang, D.; Wu, J.; Li, J.; Yang, D.; Chen, X.; Zhang, T.; and Meng, H. 2025b. MMSU: A Massive Multi-Task Spoken Language Understanding and Reasoning Benchmark. arXiv:2506.04779.

Xie, Z.; Lin, M.; Liu, Z.; Wu, P.; Yan, S.; and Miao, C. 2025. Audio-Reasoner: Improving Reasoning Capability in Large Audio Language Models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 23840–23862.

Xu, J.; Guo, Z.; He, J.; Hu, H.; He, T.; Bai, S.; Chen, K.; Wang, J.; Fan, Y.; Dang, K.; et al. 2025a. Qwen2.5-Omni Technical Report. arXiv:2503.20215.

Xu, K.-T.; Xie, F.-L.; Tang, X.; and Hu, Y. 2025b. FireredASR: Open-Source Industrial-Grade Mandarin Speech Recognition Models from Encoder-Decoder to LLM Integration. arXiv:2501.14350.

Yamagishi, J.; Veaux, C.; and MacDonald, K. 2019. CSTR VCTK Corpus: English Multi-Speaker Corpus for CSTR Voice Cloning Toolkit (Version 0.92). https://doi.org/10.7488/ds/2645. Accessed: 2019-11-13.

Yang, A.; Yang, B.; Zhang, B.; et al. 2025. Qwen2.5 Technical Report. arXiv:2412.15115.

Yang, Q.; Xu, J.; Liu, W.; Chu, Y.; Jiang, Z.; Zhou, X.; Leng, Y.; Lv, Y.; Zhao, Z.; Zhou, C.; and Zhou, J. 2024. AIR-Bench: Benchmarking Large Audio-Language Models via Generative Comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 1979–1998.

Yang, S.-w.; Chi, P.-H.; Chuang, Y.-S.; Lai, C.-I. J.; Lakhotia, K.; Lin, Y. Y.; Liu, A. T.; Shi, J.; Chang, X.; and Lin, G.-T. 2021. SUPERB: Speech Processing Universal Performance Benchmark. In *Proceedings of the 22nd Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 3161–3165.

Yu, J.; Gou, Z.; Li, C.; Wang, Z.; Yang, P.; Chen, W.; and Yin, J. 2025. Eliciting Implicit Acoustic Styles from Open-Domain Instructions to Facilitate Fine-Grained Controllable Generation of Speech. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3679–3695.

Zhang, Y.; Liu, Z.; Bu, F.; Zhang, R.; Wang, B.; and Li, H. 2025. Soundwave: Less is More for Speech-Text Alignment in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 18718–18738.

Zhao, S.; Pan, Z.; and Ma, B. 2025. ClearerVoice-Studio: Bridging Advanced Speech Processing Research and Practical Deployment. arXiv:2506.19398.

Zhu, H.; Wu, W.; Zhu, W.; Jiang, L.; Tang, S.; Zhang, L.; Liu, Z.; and Loy, C. C. 2022. CelebV-HQ: A Large-Scale Video Facial Attributes Dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 650–667.