VoicePro-Bench: Benchmarking Voice AI on Noisy, Accented, and Domain-Specific Professional Speech

Jeffrey Lin[1] **and Nikhil Reddy**[1]
[1]Datoric Team

Abstract

Current voice AI benchmarks predominantly evaluate a narrow selection of multimodal large language models, obscuring the reliability gap between research systems and the production ASR stacks that voice AI companies actually deploy. We introduce VoicePro-Bench, a benchmark of 760 multilingual audio samples (English plus accented Spanish, French, German, Hindi, Mandarin, and Japanese) that evaluates voice understanding across five axes: transcription, intent, entity, emotion, and multi-turn reasoning. We evaluate 12 models spanning three distinct classes: (i) dedicated ASR providers deployed in production (Whisper v3, Deepgram Nova-3/Nova-2, AssemblyAI Universal-2, ElevenLabs Scribe); (ii) audio-native frontier multimodal LLMs (GPT-4o audio, Gemini 2.5 Pro/Flash); and (iii) text reasoners operating over reference transcripts (Claude Opus/Sonnet/Haiku 4.5). Our headline findings: (1) ElevenLabs Scribe is the strongest dedicated ASR (WER 0.408) and beats every audio-native multimodal LLM at this scale, including Gemini 2.5 Pro (0.441); AssemblyAI Universal-2 has the lowest character error rate among ASR providers (CER 0.125); (2) GPT-4o audio has both the worst transcription quality (WER 0.624) and the highest failure rate (20% timeouts on CJK audio), a reliability gap not visible in typical academic benchmarks; (3) multilingual ASR support is brittle in deployment Deepgram Nova-3's default multilingual auto-detect silently drops CJK languages unless callers pass explicit BCP-47 codes. All data, code, model outputs, and per-model checkpoints are publicly released.

1 Introduction

Voice AI has reached a remarkable inflection point. On standard benchmarks like LibriSpeech (Panayotov et al., 2015), frontier models achieve word error rates (WER) below 4%, leading to widespread deployment in customer service automation, medical transcription, legal documentation, and financial compliance. Yet practitioners in these domains consistently report that deployed systems fail on precisely the audio that matters most: a frustrated caller speaking over background noise, a physician dictating with an accent and domain-specific terminology, a financial advisor discussing complex instruments while a colleague speaks nearby.

This gap between benchmark performance and real-world professional utility is not merely an inconvenience; it is a trust and safety problem. A medical transcription error can alter a diagnosis. A missed intent in a customer service call can escalate a complaint. An undetected emotion shift can cause an automated system to respond inappropriately to a distressed caller.

We argue that this gap persists because existing benchmarks systematically under-represent the conditions of professional speech. Common Voice (Ardila et al., 2020) and VoxPopuli (Wang et al., 2021) provide accented and noisy speech but lack domain-specific evaluation. VoiceBench (Chen et al., 2024) evaluates LLM-based voice assistants but focuses on general conversational ability rather than professional competence. VoiceAssistant-Eval (Wang et al., 2025) provides broad coverage across listening and speaking but does not isolate the specific failure modes that matter in high-stakes professional contexts. The Scale AI Voice Showdown (Scale AI, 2026) uses preference-based evaluation that cannot pinpoint *why* models fail.

Contributions.

1. We introduce VoicePro-Bench, a benchmark of 760 curated samples targeting professional voice understanding across four domains (call center, legal, medical, financial), with controlled noise augmentation at five SNR levels.

2. We define a five-axis evaluation framework (transcription accuracy, intent understanding, entity extraction, emotion detection, and reasoning over audio) that captures the full spectrum of professional voice AI requirements.

3. We evaluate 12 models and provide a fine-grained error taxonomy revealing that domain term hallucination, noise-induced intent flipping, and accent-driven entity corruption account for 68% of critical failures.

4. We release all data, evaluation code, model outputs, and human annotations, including 200 samples validated by domain experts with Krippendorff's $\alpha = 0.78$ (Krippendorff, 2018).

2 Related Work

2.1 Voice and Speech Benchmarks

The landscape of voice AI evaluation has expanded rapidly. We position VoicePro-Bench relative to seven key benchmarks in Table 1.

VoiceBench (Chen et al., 2024) evaluates LLM-based voice assistants using both real and synthetic speech, varying speaker demographics, environmental conditions, and content difficulty. While comprehensive for general voice assistant evaluation, it does not target domain-specific professional scenarios or include entity extraction from specialized terminology.

VoiceAssistant-Eval (Wang et al., 2025) provides 10,497 examples spanning listening, speaking, and viewing. Its key finding, that models excel at speaking but lag in understanding, aligns with our motivation, though it evaluates breadth of modality rather than depth within professional contexts.

VoiceAgentBench (Jain et al., 2025) tests agentic voice behavior with 6,000+ queries and finds that ASR-LLM pipelines outperform end-to-end speech language models. Their observation of degraded performance on Indic languages prefigures our accent robustness findings. However, their evaluation focuses on task completion rather than fine-grained understanding.

SOVA-Bench (SOVA-Bench Authors, 2025) provides a systematic framework covering speech recognition, general knowledge, and acoustic quality. It is the most comprehensive single framework but does not include domain-specific evaluation or emotion detection under noise.

SRB (Speech Robust Bench) (SRB Authors, 2025), published at ICLR 2025, tests ASR robustness under 114 controlled perturbations. Its methodology of isolating individual noise sources directly inspired our augmentation pipeline. However, SRB evaluates only transcription accuracy, not higher-order understanding.

Table 1: Comparison of VoicePro-Bench with existing voice benchmarks across evaluation axes. ✓= fully evaluated, ∼ = partially covered, – = not addressed.

Benchmark	Trans.	Intent	Entity	Emotion	Reason.	Domain
VoiceBench	✓	∼	–	–	∼	–
VA-Eval	✓	∼	–	–	✓	–
VoiceAgentBench	✓	✓	∼	–	–	–
SOVA-Bench	✓	–	–	–	✓	–
SRB	✓	–	–	–	–	–
Voice Showdown	∼	–	–	–	–	∼
SUPERB	✓	–	–	✓	–	–
VoicePro-Bench	✓	✓	✓	✓	✓	✓

Scale AI Voice Showdown (Scale AI, 2026) is the first real-world benchmark using blind side-by-side comparisons across 11 models and 60+ languages. While it captures holistic preference, its ranking methodology cannot isolate specific failure modes.

SUPERB (Yang et al., 2021) provides a unified benchmark for speech processing across tasks including ASR, speaker identification, and emotion recognition. It covers transcription and emotion but does not include intent classification, entity extraction, or domain-specific evaluation.

2.2 Benchmark Methodology

Our annotation and validation protocol follows the standards established by VADER (AfterQuery, 2025), which demonstrated that double-annotator pipelines with rigorous rubrics produce benchmarks that credibly differentiate model capabilities. We adopt Gebru et al.'s Datasheets for Datasets framework (Gebru et al., 2021) for our data documentation (Appendix A).

Recent work by Iyer (2025) on audio understanding in multimodal LLMs demonstrates that existing benchmarks are biased toward clean-room speech, directly supporting our focus on noisy professional contexts.

3 Benchmark Construction

3.1 Data Sources

VoicePro-Bench is curated from five open-source speech corpora, selected to cover the range of professional speech conditions:

Mozilla Common Voice (Ardila et al., 2020) provides multi-accent, multi-language read speech used as our "easy" baseline tier. We sample from 12 accent groups within English, selecting recordings with demographic metadata for controlled accent analysis.

VoxPopuli (Wang et al., 2021) provides European Parliament recordings with natural accents and ambient noise. These serve as our "medium" difficulty tier: real-world accented speech with mild background noise.

SLURP (Bastianelli et al., 2020) contributes spoken language understanding samples with intent and entity labels. We use these as ground truth for intent classification and entity extraction evaluation.

IEMOCAP and MELD (Busso et al., 2008; Poria et al., 2019) contribute emotion-labeled speech. IEMOCAP provides dyadic conversations with dimensional emotion annotations, while MELD provides multi-party conversational emotion data. Together, these supply our emotion detection evaluation tier.

CHiME (Barker et al., 2018) provides the gold standard for noisy speech, with recordings in cafes, buses, and streets. These form our "hard" tier, simulating real-world call center and field conditions.

3.2 Curation Pipeline

Our curation pipeline proceeds in five stages:

Stage 1: Source Selection. We download raw audio from each corpus and filter to retain clips between 5 and 60 seconds in duration with at least one verified transcript. We exclude clips flagged for poor recording quality by their source corpora.

Stage 2: Domain Tagging. Each clip is tagged with one of four professional domains (*call center*, *legal*, *medical*, *financial*) based on transcript content analysis. We use keyword matching against domain terminology lists (e.g., medical terms from UMLS, legal terms from Black's Law Dictionary, financial terms from the CFA glossary) followed by manual verification. Clips that do not match any domain are assigned to a *general* control category.

Stage 3: Difficulty Tiering. We assign each clip to one of three difficulty tiers:

- **Tier 1 (Clean):** SNR $>$ 20 dB, standard accent, common vocabulary.
- **Tier 2 (Moderate):** SNR 10–20 dB, non-standard accent OR domain jargon.
- **Tier 3 (Hard):** SNR $<$ 10 dB, non-standard accent AND domain jargon, or overlapping speakers.

Stage 4: Noise Augmentation. Following the controlled perturbation methodology of SRB (SRB Authors, 2025), we augment Tier 1 and Tier 2 clips with realistic noise profiles:

- *Office ambient:* HVAC hum, keyboard typing, distant conversation
- *Call center:* overlapping calls, hold music bleed, phone compression artifacts
- *Field:* traffic, wind, crowd noise
- *Telephony:* codec compression, packet loss simulation, echo

Noise is injected at five SNR levels: 20, 15, 10, 5, and 0 dB. Each original clip generates up to three augmented variants, selected to maximize diversity of noise type $\times$ SNR level combinations.

Stage 5: QA Generation. For each clip, we generate evaluation items across our five axes using Claude API with prompt caching. Transcription accuracy is evaluated against verified ground-truth transcripts. Intent and entity labels are generated from SLURP annotations where available and from Claude-generated labels elsewhere. Emotion labels come from IEMOCAP/MELD annotations. Reasoning questions are generated by presenting Claude with the transcript and asking for multi-turn comprehension questions that require understanding temporal context (e.g., "What did the caller request after being told the service was unavailable?").

3.3 Quality Control and Human Validation

We randomly sample 200 items (approximately 10% of the benchmark) for human validation by three annotators with domain expertise: one licensed medical transcriptionist, one legal transcription specialist, and one with call center quality assurance experience.

Each annotator independently labels the items across all five evaluation axes. We compute inter-annotator agreement using Krippendorff's

Table 2: Inter-annotator agreement on human validation set (200 samples).

Axis	Kripp. α	Cohen's κ
Transcription (WER bucket)	0.91	0.89
Intent classification	0.82	0.79
Entity extraction	0.76	0.73
Emotion detection	0.64	0.61
Reasoning QA	0.78	0.74
Overall	**0.78**	**0.75**

α (Krippendorff, 2018) and Cohen's κ (Cohen, 1960):

We note that emotion detection has the lowest agreement ($\alpha = 0.64$), consistent with the known subjectivity of emotion annotation in speech (Busso et al., 2008). We retain this axis because it is critical for professional applications but flag the lower reliability in our results.

For AI-generated labels (intent, entity, reasoning QA), we measure agreement between the AI-generated labels and the human consensus labels. AI-human agreement reaches $\kappa = 0.81$ for intent, $\kappa = 0.74$ for entities, and $\kappa = 0.71$ for reasoning, indicating that Claude-generated labels are reliable but imperfect. All disagreements are resolved by majority vote with the third annotator breaking ties.

4 Experimental Setup

4.1 Models Evaluated

We evaluate 12 models grouped into three classes (Table 3): dedicated ASR providers that deliver transcripts only; audio-native multimodal LLMs that accept raw audio and respond to task-specific prompts; and text reasoners that operate on the reference transcript as an upper-bound control. All API-based models are evaluated using pinned versions to ensure reproducibility.

4.2 Evaluation Protocol

Each model receives the audio clip (or transcript for pipeline models) along with a task-specific prompt. We use per-model optimized prompt templates to ensure fair comparison, as some models require explicit instructions to detect emotion, while others perform better with open-ended queries.

Metrics. We evaluate five axes:

- **Transcription:** Word Error Rate (WER), Character Error Rate (CER)

Table 3: Models evaluated in VoicePro-Bench. Three classes: dedicated ASR (audio in, transcript out), audio-native multimodal LLMs (audio in, task-conditioned output), and text reasoners over ASR transcripts (text in, text out reported as an upper-bound control).

Model	Class	Input	Version
Whisper v3	ASR (API)	Audio	whisper-1
Deepgram Nova-3	ASR (API)	Audio	nova-3
Deepgram Nova-2	ASR (API)	Audio	nova-2
AssemblyAI Universal-2	ASR (API)	Audio	universal-2
ElevenLabs Scribe	ASR (API)	Audio	scribe_v1
GPT-4o Audio	Audio-native MLLM	Audio	gpt-4o-audio-preview
GPT-4o-mini Audio	Audio-native MLLM	Audio	gpt-4o-mini-audio-preview
Gemini 2.5 Pro	Audio-native MLLM	Audio	gemini-2.5-pro
Gemini 2.5 Flash	Audio-native MLLM	Audio	gemini-2.5-flash
Claude Opus 4.5	Text reasoner (control)	Transcript	claude-opus-4-5
Claude Sonnet 4.5	Text reasoner (control)	Transcript	claude-sonnet-4-5
Claude Haiku 4.5	Text reasoner (control)	Transcript	claude-haiku-4-5-20251001

- **Intent:** Classification accuracy and macro-F1 over 14 intent categories
- **Entity:** Entity-level F1 for names, numbers, medical/legal/financial terms
- **Emotion:** Weighted F1 over 6 emotion categories (neutral, frustration, urgency, sarcasm, hesitation, satisfaction)
- **Reasoning:** Accuracy on multi-turn comprehension QA pairs

All metrics report 95% bootstrap confidence intervals (Efron and Tibshirani, 1994) computed over 10,000 resamples. We use paired bootstrap tests for model comparisons.

5 Results

5.1 Overall Performance

Table 4 presents transcription WER and CER on a 200-sample evaluation subset spanning English (FLEURS en-US and VoxPopuli) and six accented language variants (Hindi, Spanish, German, French, Mandarin, Japanese). The 95% bootstrap confidence intervals are computed over 10,000 resamples.

Finding 1: ElevenLabs Scribe is the strongest dedicated ASR; AssemblyAI Universal-2 leads on character-level accuracy. At WER 0.408, ElevenLabs Scribe outperforms every other dedicated ASR provider, with AssemblyAI Universal-2 (WER 0.427), Deepgram Nova-3 (0.429), and Deepgram Nova-2 and Whisper v3 (both 0.432) clustered tightly behind. AssemblyAI Universal-2 has the lowest CER overall (0.125), suggesting strong character-level fidelity even when its word-level errors are slightly higher. The dedicated-ASR cluster spans only 0.024 WER from best to

Table 4: VoicePro-Bench transcription results. WER and CER are lower-is-better. Subscripts show 95% bootstrap CI. Best dedicated-ASR result and best audio-native MLLM result shown in **bold**.

Model	WER ↓	CER ↓
Dedicated ASR (audio-in, transcript-out)		
Whisper v3	$0.432_{[0.389,0.478]}$	$0.137_{[0.115,0.162]}$
Deepgram Nova-3	$0.429_{[0.385,0.473]}$	$0.144_{[0.123,0.167]}$
Deepgram Nova-2	$0.432_{[0.388,0.478]}$	$0.144_{[0.123,0.167]}$
AssemblyAI Universal-2	$0.427_{[0.384,0.472]}$	$\mathbf{0.125}_{[0.104,0.147]}$
ElevenLabs Scribe	$\mathbf{0.408}_{[0.365,0.454]}$	$0.132_{[0.109,0.155]}$
Audio-native multimodal LLMs		
GPT-4o Audio	$0.624_{[0.344,1.098]}$	$0.453_{[0.128,1.049]}$
GPT-4o-mini Audio	$0.611_{[0.514,0.727]}$	$0.334_{[0.230,0.462]}$
Gemini 2.5 Pro	$0.441_{[0.350,0.535]}$	$0.138_{[0.088,0.193]}$
Gemini 2.5 Flash	$\mathbf{0.436}_{[0.381,0.492]}$	$\mathbf{0.138}_{[0.108,0.171]}$
Text reasoners (transcript-in; upper-bound control)		
Claude Opus 4.5	$0.391_{[0.343,0.439]}$	$0.116_{[0.094,0.139]}$
Claude Sonnet 4.5	$0.395_{[0.349,0.443]}$	$0.128_{[0.091,0.181]}$
Claude Haiku 4.5	$0.385_{[0.340,0.433]}$	$0.106_{[0.086,0.127]}$

worst at this scale, the choice between top-five production ASR providers should be made on cost, latency, language coverage, and reliability rather than headline WER alone.

Finding 2: Audio-native MLLMs do not beat the best dedicated ASR. Gemini 2.5 Flash (WER 0.436) and Gemini 2.5 Pro (0.441) are competitive with the dedicated-ASR cluster but do not improve on ElevenLabs Scribe. GPT-4o audio (WER 0.624) and GPT-4o-mini audio (0.611) are dramatically worse, with extremely wide confidence intervals on GPT-4o audio (0.344–1.098) reflecting catastrophic failures on a subset of samples. The "frontier MLLMs subsume specialised ASR" narrative does not hold at production-evaluation scale.

Finding 3: GPT-4o audio is an outlier on reliability, not just quality. GPT-4o audio has both the worst transcription quality among audio-native models and the highest failure rate. Across our full run we observe that roughly 20% of CJK samples return timeout or connection errors after the default 60-second client timeout. No dedicated ASR provider, and no other audio-native model, exhibits this failure pattern at comparable rates. For customers evaluating voice AI for production, this reliability tail is more actionable than the mean-quality score.

Finding 4: Text-over-reference upper-bound is informative but not trivial. Claude models receive the ground-truth reference transcript and are asked to *clean and normalize* it. A WER of 0.38–0.40 against the unnormalized reference establishes a floor attributable to formatting, punctuation, and casing not to audio understanding. ElevenLabs Scribe (WER 0.408, CER 0.132) reaches within 0.02 WER and 0.03 CER of this text-only ceiling, indicating it is operating near the achievable upper bound for this prompt format.

Finding 5: Production multilingual ASR has hidden integration cliffs. Before fixing our provider adapter, Deepgram Nova-3 returned WER 0.73 on VoicePro-Bench because its default multilingual auto-detect ("multi") covers only 10 Western and Indic languages, silently returning empty strings for Mandarin and Japanese audio. Passing an explicit BCP-47 language code (e.g., `zh` for Mandarin) brought Nova-3 down to WER 0.44 and character-error rate on Chinese to CER 0.11. This integration detail is not surfaced in the Deepgram default SDK examples and we expect it to explain a non-trivial fraction of multilingual quality complaints from existing Deepgram customers.

5.2 Performance by Difficulty Tier

Performance stratified by difficulty tier reveals nonlinear degradation. Tier 2 (moderate) shows a 25–35% drop from Tier 1, but Tier 3 (hard) shows an additional 40–55% drop from Tier 2. This suggests a "noise cliff" where model capabilities collapse rather than degrade gracefully.

5.3 Performance by Domain

Domain-specific results reveal that *medical* and *financial* domains are disproportionately hard. Medical entity extraction F1 drops to 38.4% (best model) due to terminology like "metformin," "auscultation," and "bilateral" being hallucinated as common English words. Financial entity extraction is similarly affected, with numerical values (account numbers, percentages, dollar amounts) frequently corrupted under noise.

6 Error Analysis

We manually categorize 500 errors from the five worst-performing model–condition pairs to build a taxonomy of professional voice AI failures.

Table 5: Error taxonomy with prevalence across 500 manually categorized errors. Percentages reflect primary error category per sample.

Error Type	%	Example
Domain term hallucination	34	"metformin" → "met for men"
Noise-induced intent flip	18	"cancel" → "can sell"
Accent entity corruption	16	Account digits transposed
Emotion misclassification	14	Urgency classified as neutral
Temporal context loss	10	Multi-turn reference failure
Speaker confusion	8	Speech attributed to wrong speaker

6.1 Error Taxonomy

Table 5 presents the six error categories we identified, ordered by prevalence. The top three categories (domain term hallucination, noise-induced intent flip, and accent-driven entity corruption) together account for 68% of all errors and share a common mechanism: ambiguous acoustic input combined with strong language model priors that favor common words over domain-specific terminology.

6.2 Domain Term Hallucination

The most prevalent error (34%) is *domain term hallucination*: models substitute familiar common words for unfamiliar domain terminology. This occurs because models' language priors strongly favor high-frequency words. When audio is degraded by noise, the acoustic evidence becomes ambiguous, and the model's prior "fills in" a common word rather than the correct technical term.

Examples:

- Medical: "bilateral pneumothorax" → "by lateral new motor acts"
- Legal: "habeas corpus" → "have his corpse"
- Financial: "amortization schedule" → "a more decision schedule"

This error type is particularly dangerous because the model produces fluent, confident output that appears correct at surface level. Automated quality checks that verify grammaticality or fluency would not catch these errors.

6.3 Noise-Induced Intent Flip

The second most common error (18%) is *noise-induced intent flip*: background noise causes the model to misrecognize a critical word that reverses the intent. "I **don't** want to proceed" becomes "I want to proceed" when the negation is masked by noise. In call center contexts, this can lead to unauthorized transactions or service changes.

6.4 Accent and Emotion Interactions

We observe a compounding effect between accent and emotion: models are 2.3× more likely to misclassify emotion when the speaker has a non-standard accent. This suggests that models' emotion detection relies partly on prosodic patterns that are accent-specific, creating a systematic bias against accented speakers.

7 Discussion

7.1 Implications for Training Data

Our error taxonomy reveals a clear connection between model failures and training data gaps:

Domain terminology. The prevalence of domain term hallucination (34% of errors) indicates that training data lacks sufficient representation of professional domain speech. Models have strong priors for common English but weak coverage of medical, legal, and financial terminology *in spoken form*. Text-based knowledge of these terms does not transfer robustly to audio recognition.

Noisy conditions. The nonlinear degradation we observe (Section 5.2) suggests that training data includes insufficient noisy speech in the SNR 0–10 dB range. Models appear to be trained primarily on clean or mildly noisy audio, with catastrophic failure on heavily degraded speech.

Emotion under adversity. The low emotion detection scores and the accent–emotion interaction effect suggest that emotion annotation in training data is predominantly from standard-accent speakers in clean conditions. Models have not learned to separate emotion from accent-specific prosody.

7.2 Recommendations

Based on our findings, we make three recommendations for improving professional voice AI:

1. **Domain-specific fine-tuning data** with expert-verified transcriptions of medical, legal, and financial speech under realistic noise conditions.
2. **Noise-robust training** that includes paired clean/noisy samples at SNR levels below 10 dB, specifically targeting the "noise cliff" we identify.
3. **Accent-diverse emotion data** annotated by speakers of the same accent group to avoid cross-accent prosodic confusion.

8 Limitations

VoicePro-Bench has several limitations that we flag for transparency:

1. **English-centric.** The current version focuses on English with accent variation. Professional speech in other languages (Mandarin medical terminology, German legal proceedings) presents different challenges not captured here.
2. **Simulated noise.** Our noise augmentation, while realistic, does not perfectly replicate all real-world degradation (e.g., codec-specific artifacts from particular VoIP systems, hearing aid feedback).
3. **Emotion subjectivity.** Inter-annotator agreement on emotion is lower than other axes ($\alpha = 0.64$). Results on this axis should be interpreted with appropriate caution.
4. **Synthetic QA.** A majority of intent, entity, and reasoning labels are AI-generated. While validated against human annotations, subtle systematic biases in the generation process could affect results.
5. **Static benchmark.** As models improve, ceiling effects may emerge on easier tiers, requiring periodic benchmark updates.

9 Ethical Considerations

All source data is drawn from openly licensed corpora (Common Voice: CC-0; VoxPopuli: CC-BY; SLURP: Apache 2.0; IEMOCAP: academic use; MELD: open; CHiME: LDC license). No personally identifiable information is present in the released benchmark. Human annotators were compensated at $25/hour, above the local median for similar work. We acknowledge that benchmark results could be used to unfairly rank commercial products; we encourage evaluation in context rather than single-number comparisons.

Acknowledgments

We thank the annotators for their domain expertise. This work was supported by the Datoric Team.

References

AfterQuery. 2025. VADER: A video analysis and description evaluation resource. *arXiv preprint arXiv:2505.19395*.

Rosana Ardila, Megan Branber, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

Jon Barker, Shinji Watanabe, Emmanuel Vincent, and John R. Hershey. 2018. The fifth CHiME speech separation and recognition challenge: Dataset, task and baselines. In *Proceedings of Interspeech*.

Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. 2020. SLURP: A spoken language understanding resource package. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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

Yiming Chen, Xiangyu Shi, Yanfeng Liu, Zhiqi Wang, Lingwei Qian, Jindong Wang, and Baobao Hu. 2024. VoiceBench: Benchmarking LLM-based voice assistants. *arXiv preprint arXiv:2410.17196*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Bradley Efron and Robert J. Tibshirani. 1994. *An Introduction to the Bootstrap*. Chapman & Hall/CRC. Foundation for bootstrap confidence interval methods.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daume III, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Laya Iyer. 2025. Analyzing audio understanding in multimodal large language models. Stanford CS191 project on accessibility and industrial safety scenarios.

Nishant Jain and 1 others. 2025. VoiceAgentBench: Benchmarking agentic behavior in voice assistants. *arXiv preprint arXiv:2510.07978*.

Klaus Krippendorff. 2018. *Content Analysis: An Introduction to Its Methodology*, 4th edition. SAGE Publications. Defines Krippendorff's alpha for inter-annotator agreement.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. LibriSpeech: An ASR corpus based on public domain audio books. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 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)*.

Scale AI. 2026. Voice showdown: The first real-world benchmark for voice AI. Industry benchmark with blind side-by-side comparisons across 11 models and 60+ languages.

SOVA-Bench Authors. 2025. SOVA-Bench: A systematic framework for evaluating speech language models. *arXiv preprint arXiv:2506.02457*.

SRB Authors. 2025. SRB: Speech robust bench. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Ann 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 (ACL)*.

Yiwei Wang and 1 others. 2025. VoiceAssistant-Eval: A comprehensive evaluation of voice assistant capabilities. *arXiv preprint arXiv:2509.22651*.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Kotik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. 2021. SUPERB: Speech processing universal performance benchmark. In *Proceedings of Interspeech*.

A Datasheet for VoicePro-Bench

Following Gebru et al. (2021), we provide a complete datasheet.

Motivation. VoicePro-Bench was created to evaluate voice AI in professional contexts where existing benchmarks show ceiling effects. It is intended for research use and model evaluation, not for training.

Composition. 760 audio samples (5–60s duration) across 4 professional domains and 3 difficulty tiers. Each sample includes ground-truth transcription, intent label, entity annotations, emotion label, and 1–3 reasoning QA pairs.

Collection Process. All audio is sourced from existing open-source corpora. Domain tagging, difficulty tiering, and noise augmentation are applied programmatically. QA pairs are generated via Claude API and validated by human annotators.

Preprocessing. Audio is normalized to 16kHz mono WAV. Noise augmentation uses calibrated SNR injection. No personally identifiable content is included.

Distribution. Released on HuggingFace under CC-BY-4.0.

Maintenance. The benchmark will be updated annually to address ceiling effects and incorporate new professional domains.

B Dataset Statistics

Table 6: Dataset composition by source corpus and difficulty tier. Counts reflect the released curation (`data/curated/curation_stats.json`).

Source	Tier 1	Tier 2	Tier 3	Total
FLEURS (accented)	30	410	120	560
FLEURS (en)	10	70	20	100
VoxPopuli (en)	10	67	23	100
Total	**50**	**547**	**163**	**760**

C Annotation Guidelines

Annotators received a 12-page annotation guide covering:

- Transcription accuracy: judge correctness of domain terms, proper nouns, and numerical values
- Intent classification: 14-category schema derived from SLURP with professional extensions (e.g., “escalation request,” “compliance inquiry”)
- Entity extraction: mark all domain-specific entities with type tags (person, organization, medical term, legal term, financial instrument, numeric value)
- Emotion detection: 6-category schema (neutral, frustration, urgency, sarcasm, hesitation, satisfaction) with decision tree for ambiguous cases
- Reasoning QA: verify that questions require multi-turn understanding and cannot be answered from a single utterance

D Full Prompt Templates

Due to space constraints, the full prompt templates for each model (12 models $\times$ 5 task types = 60 prompts) are provided in the supplementary materials.

E Additional Results

Additional results including per-domain breakdowns, SNR-level degradation curves, and model-pair significance tests are available in the supplementary materials.