VoicePro-Bench: Frontier ASR and Audio-Native LLMs on Professional Voice Understanding

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

Abstract

We introduce VoicePro-Bench, a multi-axis benchmark of professional voice understanding covering transcription, SLURP intent and entity extraction, MELD emotion, AMI multi-turn reasoning, and MUSAN/WHAM noise robustness. We evaluated 12 systems: 5 dedicated ASR providers, 4 audio-native multimodal LLMs, and 3 text-reasoner controls fed the reference transcript. Three results. **(i)** On SLURP intent the best text-reasoner control outperforms the best audio-native MLLM by 27 pp F1 (n=200), and on AMI reasoning by 12 pp token-F1 (n=248), locating audio understanding (not text reasoning) as the dominant source of error on these tasks. **(ii)** GPT-4o audio refuses 20% of valid SLURP transcription requests, a deployment-visible failure mode that WER does not capture. **(iii)** Gemini 2.5 Flash shows a 20-to-0 dB WER increase of $+0.073$, roughly $1.5\times$ the steepest dedicated-ASR cliff. We release all audio-level outputs, scoring code, and 95% bootstrap CIs per metric.

1 Introduction

Voice AI is being deployed across an increasingly wide range of professional and consumer settings: customer service, medical dictation, legal transcription, accessibility, dispatch, and compliance monitoring. These applications share a common failure mode that existing academic benchmarks under-measure: production deployments must handle *accented* and *multilingual* speech, not just clean native-English audio.

Current voice AI benchmarks mostly measure either a narrow selection of audio-native multimodal LLMs (Chen et al., 2024; Wang et al., 2025a; Hou et al., 2025) or dedicated ASR providers (Shah et al., 2025; Wang et al., 2025b), but rarely evaluate both on the same audio with the same normalization. This reflects a real methodological gap: ASR providers and audio-native LLMs have different input contracts, different acceptable response formats, and different failure modes (transcription error versus refusal, timeout, or language-misidentification). A fair comparison requires running them in parallel on a shared corpus with shared metrics.

Scope. The VoicePro-Bench framework specifies a five-axis evaluation for professional voice understanding: transcription, SLURP intent and entity understanding, emotion recognition, multi-turn reasoning, and noise-robustness. This paper reports results on all five axes. Coverage is at or near full on every axis: the 200-sample transcription axis is at n=200/200 for GPT-4o Audio, n=198/200 for Gemini 2.5 Pro, and n=195/200 for Gemini 2.5 Flash; SLURP is at full coverage for all models except Gemini 2.5 Flash (n=382/600 residual); MELD emotion and AMI reasoning are at full coverage. Three rows carry small residual shortfalls (Deepgram Nova-3 n=198, Nova-2 n=199, Claude Opus n=198) and are disclosed in Table 2. The noise-robustness axis evaluates the cost-tier cheaper MLLM variants (GPT-4o-mini Audio, Gemini 2.5 Flash) by design, for the reason given in §5.6.

Contributions.

1. We introduce VoicePro-Bench as a multi-axis evaluation framework for voice AI and instantiate all 5 planned axes: transcription on 760 FLEURS/VoxPopuli samples; SLURP intent/entity on 200 samples; MELD emotion on 200 samples; AMI multi-turn reasoning on 248 questions; and a MUSAN/WHAM noise-robustness cliff at four SNR levels on 200 samples.
2. We evaluate 12 models grouped into three classes (dedicated ASR, audio-native multimodal LLMs, text reasoners as upper-bound controls) with a shared ASR→LLM cascade

for providers that cannot directly emit classification or extraction output.

3. We document a 27 pp intent-F1 gap and a 12.4 pp reasoning-F1 gap between the best text-reasoner control (given reference transcripts) and the best audio-native MLLM (given raw audio), indicating that audio understanding, not reasoning capacity, accounts for the gap on these tasks.

4. We surface two deployment-visible behaviors absent from academic leaderboards: GPT-4o audio refuses 20% of valid SLURP transcription requests, and Deepgram Nova-3's default multilingual auto-detect silently drops CJK languages unless callers pass explicit BCP-47 codes.

5. We release the data, scoring code, and all model outputs with bootstrap confidence intervals, including a noise-robustness cliff across 4 SNR levels: Gemini 2.5 Flash shows a 20-to-0 dB WER gap roughly $1.5\times$ that of any dedicated ASR.

2 Related Work

2.1 Voice Benchmarks for LLM-Based Assistants

VoiceBench (Chen et al., 2024) systematically evaluates speech-based LLM assistants on general-purpose tasks. Its methodology shares the multi-model-class spirit of VoicePro-Bench but focuses on conversational assistant quality rather than raw transcription on accented and multilingual speech.

VoiceAssistant-Eval (Wang et al., 2025a) benchmarks AI assistants across listening, speaking, and viewing, providing breadth across modalities. VoicePro-Bench is narrower (transcription-only in this pilot) but deeper on the accented/multilingual dimension.

VoiceAgentBench (Jain et al., 2025) asks whether voice assistants are ready for agentic tasks. We cite this as context for why raw transcription quality matters: downstream agent failures are amplified by upstream ASR errors.

SOVA-Bench (Hou et al., 2025) benchmarks the speech conversation ability of LLM-based voice assistants. Its "conversation ability" framing complements our transcription focus: high transcription quality is necessary but not sufficient for good conversation.

Speech Robust Bench (Shah et al., 2025) evaluates ASR robustness under input perturbations. Our accented-speech evaluation complements SRB's perturbation-based evaluation: speaker-side variation versus input-side perturbation.

Open Universal Arabic ASR Leaderboard (Wang et al., 2025b) provides a dedicated-ASR comparison for Arabic. Our work extends similar head-to-head ASR comparison across seven accent groups with audio-native LLMs included.

Industry benchmarks. Scale AI's "Voice Showdown" (Scale AI, 2026) is an industry benchmark that uses blind side-by-side comparisons across multiple models and many languages; we cite it as a contrasting data point, noting that its ranking methodology cannot isolate specific failure modes. A Stanford student project (Iyer, 2025) on audio understanding in multimodal LLMs also provides complementary context.

2.2 Accented and Multilingual Speech Corpora

Common Voice (Ardila et al., 2020) provides multi-accent, multi-language read speech.

VoxPopuli (Wang et al., 2021) provides European Parliament recordings with natural accents. We use the English subset of VoxPopuli as our "English with light ambient noise" source.

SLURP (Bastianelli et al., 2020) is the source data for the intent and entity axes. We use its 18-scenario taxonomy and its native `[slot : value]` annotation schema. **MELD** (Poria et al., 2019) provides the 7-class conversational emotion labels used in the emotion axis. **AMI Meeting Corpus** (Carletta et al., 2005) supplies the multi-party meeting audio used in the reasoning axis. MUSAN and WHAM are used as additive-noise sources for the noise-robustness axis at SNR $\in \{20, 10, 5, 0\}$ dB (Section 5.6).

3 Benchmark Construction

3.1 Data Sources

Each axis draws on an openly licensed corpus chosen for its domain fit:

Transcription axis: FLEURS + VoxPopuli. FLEURS-accented (six non-native accent groups: Mandarin, German, Spanish, French, Hindi, Japanese) plus FLEURS-en provide the multilingual base; VoxPopuli-en adds European Parliament recordings with naturally-occurring ambi-

ent noise. Total transcription pool: 760 samples (560 FLEURS-accented, 100 FLEURS-en, 100 VoxPopuli-en); main results use a 200-sample balanced subset.

Intent and entity axes: SLURP. SLURP (Bastianelli et al., 2020) is a 200-sample English command-speech corpus with 18-scenario intent labels and bracketed `[slot : value]` entity annotations. We stratified 200 SLURP test-split samples across all 18 scenarios (seed=42).

Emotion axis: MELD. MELD (Poria et al., 2019) provides 7-class (neutral, happy, sad, angry, fearful, surprised, disgusted) emotion labels on Friends-TV dialogue audio; we drew 200 samples from the test split.

Reasoning axis: AMI Meeting Corpus. AMI (Carletta et al., 2005) supplies multi-speaker meeting audio. We constructed a 248-question reasoning set grounded in AMI meeting segments. Questions target facts, inferences, and cross-turn references that require following the full dialogue; questions were synthesized by Claude (see Limitations, section 7).

Noise-robustness axis: MUSAN/WHAM. A 200-sample subset balanced from FLEURS-en and VoxPopuli-en is augmented with MUSAN (music + speech + noise) and WHAM (real-world ambient) clips at four target SNR levels: 20, 10, 5, 0 dB. Each clean sample is deterministically paired with a noise clip (`md5(sample_id ‖ snr) mod pool_size`) so every (model, SNR) cell sees the same noise realization. Models transcribe the noisy audio; WER against the clean reference is the primary metric. See Table 6 and Findings 9–11.

3.2 Curation Pipeline

Transcription subset. We filtered clips to 5–60 seconds with a verified reference, inherited accent/language tags from source metadata, and assigned a difficulty tier (easy/medium/hard) from a signal-quality and transcript-complexity heuristic. The 200-sample subset is stratified by accent and tier.

SLURP subset. We streamed `qmeeus/slurp` (parquet-backed HF mirror of SLURP's test split), deduped by `slurp_id`, and stratified 200 samples across 18 scenarios with seed=42. We parsed SLURP's native `[slot : value]` bracket annotations directly. Each sample is tagged to a domain (home-automation, general, call-center) derived from its scenario.

MELD subset. We drew 200 samples from MELD's test split (`AudioLLMs/meld_emotion_test`), materializing audio to 16 kHz mono.

AMI reasoning set. Starting from AMI meeting segments, we synthesized 248 questions that require following multi-party dialogue: facts, inferences, and cross-turn references. Question synthesis used Claude; each question has a human-auditable reference answer. See Limitations 7 for the self-annotated-test caveat.

3.3 Coverage and Sample Size: A Caveat

Per-axis coverage. All 12 models are evaluated on the 200-sample SLURP intent and entity axes. MELD emotion is run on the 4 audio-native MLLMs (Gemini 2.5 Pro/Flash, GPT-4o audio/mini); dedicated ASR providers cannot emit categorical labels and Claude cannot ingest audio. AMI reasoning is run on 7 models at full n=248 (the 3 Claude text-reasoner controls plus GPT-4o audio, GPT-4o-mini audio, Gemini 2.5 Pro, and Gemini 2.5 Flash) following the day-2 rate-limit backfill that filled GPT-4o audio's prior 147/248 partial coverage and the deferred Gemini 2.5 Pro run; see Maintenance (§A).

Gemini SLURP partial coverage. Gemini 2.5 Pro is at n=597/600 after the day-3 backfill (288 of the original 291 429/503 rows recovered; 3 residual). Gemini 2.5 Flash remains at n=382/600 due to a sustained Google-side 503 upstream saturation on the Flash endpoint during the retry window, which the retry tool exited on rather than overrunning quota; this shortfall is documented in §A.

4 Experimental Setup

4.1 Models Evaluated

We evaluated 12 models grouped into three classes (Table 1): dedicated ASR providers that deliver transcripts only (including Whisper (Radford et al., 2023)); audio-native multimodal LLMs that accept raw audio and respond to task-specific prompts (GPT-4o audio (OpenAI, 2024) and the Gemini 2.5 family (Gemini Team, Google, 2024)); and text reasoners that operate on the reference

Table 1: 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 the reference transcript (text in, text out — reported as an upper-bound control).

Model	Class	Input	Version
Whisper large-v3	ASR (HF)	Audio	openai/whisper-large-v3
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

transcript as an upper-bound control. All API-based models were evaluated using pinned versions to ensure reproducibility.

4.2 Evaluation Protocol

Metric: transcription. WER and CER against the reference transcripts, normalized with Whisper's `BasicTextNormalizer` for Latin-script languages; CJK uses CER without normalization.

Metric: intent. Sklearn weighted-F1 over SLURP's 18 scenarios. We map each model's free-text response to the closest scenario by longest-prefix match against the canonical label list.

Metric: entity. Set-level precision/recall/F1 over normalized slot-values. SLURP's bracket format is parsed directly when present; otherwise we fall back to a regex extractor for quoted spans, capitalized multi-word names, and numeric amounts.

Metric: emotion. Sklearn weighted-F1 over the 7 MELD classes, with unparsed responses counted as mispredictions.

Metric: reasoning. Token-level F1 and exact-match against the reference answer, normalized for case and punctuation.

ASR→LLM cascade. Dedicated ASR providers return transcripts but cannot emit intent or entity directly. For those 5 models on the SLURP axis we ran a two-stage cascade: the ASR provider transcribed the audio, then Claude Haiku 4.5 was invoked on the hypothesis with the same task prompt used for audio-native models. This yields what the cascade pipeline would produce in deployment ("ASR-WER-propagated intent/entity F1").

Table 2: Main transcription results on the 200-sample subset. Subscripts denote 95% bootstrap CI bounds. Best within each class in **bold**. n column reports covered samples.

Model	WER	CER	n
Dedicated ASR (audio-in, transcript-out)			
Whisper large-v3	$0.432_{[0.389, 0.478]}$	$0.137_{[0.115, 0.162]}$	200
Deepgram Nova-3	$0.429_{[0.385, 0.473]}$	$0.144_{[0.123, 0.167]}$	198
Deepgram Nova-2	$0.432_{[0.388, 0.477]}$	$0.144_{[0.122, 0.167]}$	199
AssemblyAI Universal-2	$0.427_{[0.384, 0.472]}$	$\mathbf{0.125}_{[0.104, 0.147]}$	200
ElevenLabs Scribe	$\mathbf{0.408}_{[0.365, 0.454]}$	$0.132_{[0.109, 0.155]}$	200
Audio-native multimodal LLMs			
GPT-4o Audio	$0.612_{[0.464, 0.795]}$	$0.308_{[0.164, 0.504]}$	200
GPT-4o-mini Audio	$0.611_{[0.514, 0.727]}$	$0.334_{[0.230, 0.462]}$	200
Gemini 2.5 Pro	$\mathbf{0.411}_{[0.366, 0.458]}$	$\mathbf{0.132}_{[0.107, 0.158]}$	198
Gemini 2.5 Flash	$0.429_{[0.385, 0.476]}$	$0.143_{[0.116, 0.173]}$	195
Text reasoners (transcript-in; upper-bound control)			
Claude Opus 4.5	$0.391_{[0.343, 0.439]}$	$0.116_{[0.094, 0.139]}$	198
Claude Sonnet 4.5	$0.395_{[0.348, 0.443]}$	$0.128_{[0.091, 0.181]}$	200
Claude Haiku 4.5	$\mathbf{0.385}_{[0.340, 0.433]}$	$\mathbf{0.106}_{[0.086, 0.127]}$	200

A rate-limit backfill brought the three audio-native MLLMs from initial 47/54/142 coverage to 200/198/195. Residual gaps are 2 Gemini Pro and 5 Gemini Flash Google 504 deadline-exceeded errors that did not retry-recover. Dedicated-ASR rows at $n<200$ (Nova-3 / Nova-2) and Claude Opus ($n=198$) reflect a small number of provider-side or service-side errors from the original sweep.

Text-reasoner control. Claude models receive the reference transcript as text input and run each task on it. For transcription this establishes a text-only WER ceiling; for intent/entity/reasoning it isolates reasoning quality from audio understanding. Claude does not ingest raw audio in this release.

Confidence intervals. 95% bootstrap confidence intervals (percentile method, 10 000 resamples, seed 42) (Efron and Tibshirani, 1994) for all mean metrics; F1-weighted reported as a scalar.

5 Results

5.1 Transcription Results

Table 2 presents aggregate WER and CER on the 200-sample transcription subset (FLEURS + VoxPopuli). All numbers are computed from the released aggregated result file; see Appendix C.

Finding 1: ElevenLabs Scribe is the strongest dedicated ASR and is matched by the best audio-native MLLM. ElevenLabs Scribe achieves WER 0.408, the best among dedicated ASR providers. Gemini 2.5 Pro, the strongest audio-native MLLM after the 2026-04-22 backfill, comes in at 0.411 with heavily overlapping CIs (Scribe 0.365–0.454, Pro 0.366–0.458) — a

statistical tie on point estimate. Gemini 2.5 Flash (0.429) is slightly behind both. AssemblyAI Universal-2 has the lowest CER across ASR providers (0.125), reflecting particular strength on character-level accuracy in Romanic and CJK accent groups.

Finding 2: GPT-4o Audio has a long reliability tail. GPT-4o Audio has the worst transcription quality among audio-native MLLMs at full n=200 coverage (WER 0.612, CER 0.308) with a very wide 95% bootstrap CI (WER CI width = 0.33, vs 0.21 for GPT-4o-mini Audio and 0.09 for Gemini 2.5 Pro). At the full sample size, the spread is no longer an artifact of small n: longer CJK utterances cluster at high WER (heavy right tail) alongside a sizeable fraction of clean transcriptions, yielding the bimodal distribution that widens the bootstrap. This is a reliability pattern visible in deployment but absent from single-mean academic leaderboards.

Finding 3: Claude text-reasoner upper-bound is 0.385–0.395 WER. The text-reasoner controls (Claude Opus/Sonnet/Haiku 4.5) receive the reference transcript and return cleaned/normalized text. Their residual WER of 0.385–0.395 establishes a text-only ceiling attributable to formatting, punctuation, and casing normalization differences, not to audio understanding. ElevenLabs Scribe at WER 0.408 reaches within 0.025 WER of this text-only ceiling (and Gemini 2.5 Pro at 0.411 essentially matches it), indicating both operate near the achievable upper bound for this prompt format.

Finding 4: Production multilingual ASR has hidden integration cliffs. Before we passed explicit BCP-47 language codes to Deepgram Nova-3, its default multilingual auto-detect ("multi") returned empty strings on Mandarin and Japanese audio, scoring as WER 1.0 on those samples. These were silent failures (the API returned a successful HTTP response with an empty transcript), indistinguishable in logs from genuine transcription errors. Passing an explicit BCP-47 code (`zh`, `ja`) brought Nova-3 back in line with the other dedicated ASR providers at WER 0.429. This behavior is not surfaced by the default Deepgram SDK examples; we expect it explains a non-trivial fraction of multilingual quality complaints from existing Deepgram customers.

Table 3: SLURP intent and entity results (200 stratified samples × 3 tasks = 600 calls per model). Intent F1 is sklearn weighted; entity F1 is set-F1. Subscripts are 95% bootstrap CI (on accuracy/entity-F1 means). Refusal column reports the fraction of transcription-task responses that declined to transcribe. Post the day-3 2026-04-20 backfill, Gemini 2.5 Pro is at n=597/600 (3 residual 500s). Gemini 2.5 Flash remains at n=382/600 due to a sustained Google-side 503 upstream saturation on the Flash endpoint during the retry window; its aggregate F1 is depressed by the missing rows (scored as F1 0).

Model	Intent F1	Entity F1	Refusal%
ASR→Claude-Haiku cascade			
Whisper large-v3	0.675	$0.377_{[0.320, 0.434]}$	—
Deepgram Nova-3	0.679	$0.364_{[0.309, 0.420]}$	—
Deepgram Nova-2	0.655	$0.367_{[0.309, 0.424]}$	—
AssemblyAI Univ.-2	0.707	$0.411_{[0.353, 0.468]}$	—
ElevenLabs Scribe	**0.722**	$\mathbf{0.418}_{[0.359, 0.475]}$	—
Audio-native MLLMs			
GPT-4o Audio	0.568	$0.253_{[0.202, 0.305]}$	20.0
GPT-4o-mini Audio	0.595	$0.203_{[0.156, 0.253]}$	26.5
Gemini 2.5 Pro	0.479	$0.330_{[0.272, 0.389]}$	—
Gemini 2.5 Flash†	0.486	$0.322_{[0.260, 0.385]}$	—
Text reasoners (transcript-in)			
Claude Opus 4.5	**0.748**	$\mathbf{0.511}_{[0.447, 0.575]}$	—
Claude Sonnet 4.5	0.713	$0.481_{[0.420, 0.544]}$	—
Claude Haiku 4.5	0.717	$0.437_{[0.379, 0.496]}$	9.5

† Gemini 2.5 Flash: 218/600 API calls remained unrecovered after day-3 retry (sustained Google-side 503 upstream saturation specific to the Flash endpoint). Error rows scored as F1 0. Gemini 2.5 Pro was recovered to n=597/600 and carries no footnote.

5.2 Per-Accent and Per-Tier Breakdowns

Per-accent and per-tier breakdowns are included in the released result artifacts (see Appendix C). We do not make ranking claims on these finer-grained breakdowns in the paper body because per-cell sample sizes (≈25 per accent, ≈50−140 per tier) yield CIs too wide for reliable ranking.

5.3 Intent and Entity (SLURP)

Table 3 presents intent F1 (weighted, over 18 scenarios) and entity F1 (set-F1 over SLURP slot-values) on 200 stratified SLURP samples. Dedicated ASR providers run the ASR→Claude-Haiku cascade (section 4.2); Claude models receive the reference transcript (text-reasoner control). See Appendix C for released result artifacts.

On intent and entity, text-reasoning controls outperform audio-native models. Claude Opus 4.5 as a text-reasoner control reaches intent F1 0.748 and entity F1 0.511. Post the day-3 backfill, the best audio-native MLLM is Gemini 2.5 Pro on both axes (intent F1 0.479 at n=597/600

Table 4: MELD emotion F1 (n=200; Gemini 2.5 Flash n=194). 7-class weighted F1. `Unparsed` is responses that could not be mapped to a label.

Model	F1 (weighted)	Accuracy	Unparsed
GPT-4o Audio	0.343	$0.350_{[0.285, 0.415]}$	29
GPT-4o-mini Audio	0.364	$0.375_{[0.310, 0.445]}$	15
Gemini 2.5 Flash	0.389	$0.381_{[0.314, 0.449]}$	0
Gemini 2.5 Pro	**0.443**	$\mathbf{0.440}_{[0.375, 0.505]}$	0

and entity F1 0.330 at n=597/600) — a 27 pp intent gap and an 18 pp entity gap. Even the best ASR→Haiku cascade (ElevenLabs Scribe at intent F1 0.722, entity F1 0.418) outperforms every audio-native model. On command-speech intent and entity tasks, audio understanding accounts for the gap, not reasoning. Gemini 2.5 Flash remains at n=382/600 (see †), and its aggregate F1 (0.486 intent, 0.322 entity) is mildly depressed by the 218 missing rows scored as F1 0; even a generous hypothetical recovery would not lift it above the text-reasoner or cascade rows.

Finding 6: GPT-4o audio refuses 20% of valid SLURP transcription requests. On the SLURP transcription task, GPT-4o audio returns "I'm sorry, but I can't play radio stations…"-style refusals on 40 of 200 samples; GPT-4o-mini audio refuses on 53 of 200 (26.5%). The content of the audio is benign (SLURP utterances are user-simulated voice commands like "play radio station 101.9"), but the model treats the *task* ("play a radio station") as an action it cannot perform and refuses instead of transcribing. This is not an audio-understanding failure; it is a prompt-following failure specific to audio mode. It is a deployment-blocking behavior that would not surface on a leaderboard that filters non-responses.

5.4 Emotion (MELD)

Table 4 presents 7-class emotion F1 on 200 MELD samples. Only the 4 audio-native MLLMs are evaluated (categorical emotion requires audio input; text-reasoner controls without audio access cannot predict emotion from cleaned transcripts alone).

Finding 7: Frontier audio-native models do not yet reliably classify emotion. The best audio-native model on 7-class MELD reaches F1 0.443 (Gemini 2.5 Pro). A random baseline is ≈0.14; a majority-class (*neutral*) baseline on MELD's class distribution is ≈0.23. All four audio-native MLLMs sit between 2× and 3× the random baseline, but none clears 0.5 F1 on this 7-way task. For deployment use cases that rely on emotion as a routing signal (e.g., sentiment-driven call escalation), this is well below production-usable quality.

Table 5: AMI reasoning results. Token-F1 is primary; exact-match is stricter. 95% bootstrap CI in subscripts. n column reports covered questions. All rows are at full coverage (n=248) except Gemini 2.5 Flash (n=229); post day-2 rate-limit backfill for GPT-4o Audio and Gemini 2.5 Pro.

Model	Token F1	Exact match	n
Audio-native MLLMs			
GPT-4o Audio	$0.512_{[0.472, 0.552]}$	$0.194_{[0.145, 0.242]}$	248
GPT-4o-mini Audio	$0.287_{[0.258, 0.317]}$	$0.024_{[0.008, 0.044]}$	248
Gemini 2.5 Flash	$0.411_{[0.366, 0.457]}$	$0.100_{[0.066, 0.140]}$	229
Gemini 2.5 Pro	$0.509_{[0.463, 0.555]}$	$0.206_{[0.157, 0.258]}$	248
Text reasoners (reference transcript)			
Claude Opus 4.5	$0.560_{[0.520, 0.599]}$	$0.282_{[0.226, 0.339]}$	248
Claude Sonnet 4.5	$0.555_{[0.516, 0.594]}$	$0.254_{[0.202, 0.311]}$	248
Claude Haiku 4.5	$\mathbf{0.636}_{[0.595, 0.676]}$	$\mathbf{0.367}_{[0.307, 0.427]}$	248

5.5 Multi-Turn Reasoning (AMI)

Table 5 presents token-level F1 and exact-match on the 248-question AMI reasoning set. Seven models are reported: 3 Claude text-reasoner controls (on the reference meeting transcript) and 4 audio-native MLLMs (on the raw meeting audio). All audio-native rows except Gemini 2.5 Flash are at full coverage (n=248) post day-2 rate-limit backfill; Gemini 2.5 Flash covers 229 of 248.

Finding 8: Reasoning gap isolates audio understanding from reasoning. Claude Haiku 4.5 as a text-reasoner control reaches token-F1 0.636 on AMI when given the reference transcript. Post day-2 backfill, the best audio-native MLLMs on the same questions with raw audio (GPT-4o Audio at 0.512 and Gemini 2.5 Pro at 0.509, both at full n=248) sit 12.4–12.7 pp below that ceiling, and even below the weakest Claude text reasoner (Sonnet 4.5 at 0.555). Because the questions and reference answers are identical, the gap cannot be attributed to reasoning capacity; it is the cost of having to recover the dialogue from audio. Combined with the SLURP finding, this supports a single diagnosis: *on today's frontier models, audio understanding, not reasoning, is the dominant source of error on professional voice tasks*. A secondary observation: GPT-4o Audio at $F1$=0.512 is materially better than the partial-coverage day-1 read (which had it at 0.452 on 147 easier-tail questions) — the rate-limit cap had truncated at a point that

Table 6: Noise-cliff WER under MUSAN/WHAM additive noise at SNR $\in$ {20, 10, 5, 0} dB (n=200 per cell, subject to retry caveats below). Subscripts are 95% bootstrap CI on WER means. Δ is the WER increase from 20 dB to 0 dB — a proxy for cliff steepness. GPT-4o-mini Audio is reported but excluded from the robustness ordering: it paraphrases rather than transcribes at all SNR levels and its WER is therefore not comparable to faithful transcribers (see Finding 10).

Model	20 dB	10 dB	5 dB	0 dB	Δ
Dedicated ASR					
Whisper large-v3	$0.233_{[0.217, 0.251]}$	$0.236_{[0.219, 0.253]}$	$0.246_{[0.228, 0.265]}$	$0.280_{[0.258, 0.303]}$	+0.047
Deepgram Nova-3	$0.261_{[0.243, 0.280]}$	$0.267_{[0.248, 0.287]}$	$0.275_{[0.255, 0.296]}$	$0.309_{[0.284, 0.336]}$	+0.048
Deepgram Nova-2	$0.261_{[0.243, 0.280]}$	$0.266_{[0.247, 0.286]}$	$0.286_{[0.264, 0.308]}$	$0.310_{[0.284, 0.339]}$	+0.049
AssemblyAI Univ.-2	$0.245_{[0.229, 0.261]}$	$0.245_{[0.230, 0.262]}$	$0.254_{[0.237, 0.272]}$	$\mathbf{0.275}_{[0.256, 0.295]}$	**+0.030**
ElevenLabs Scribe	$0.244_{[0.227, 0.261]}$	$0.252_{[0.234, 0.271]}$	$0.270_{[0.249, 0.291]}$	$0.285_{[0.264, 0.307]}$	+0.041
Audio-native MLLMs					
Gemini 2.5 Flash	$0.239_{[0.223, 0.256]}$	$0.246_{[0.229, 0.264]}$	$0.264_{[0.244, 0.285]}$	$0.312_{[0.283, 0.343]}$	+0.073
GPT-4o-mini Audio‡	$2.595_{[2.292, 2.939]}$	$2.258_{[1.984, 2.569]}$	$2.226_{[1.971, 2.494]}$	$2.346_{[1.956, 2.791]}$	—

‡ GPT-4o-mini Audio paraphrases rather than transcribes despite the verbatim instruction; see Finding 10. n per cell: 200 for all except Deepgram Nova-2 at 0 dB (175) and 5 dB (194); Nova-3 at 0 dB (194); Gemini at 5/0 dB (199). Sample shortfalls arise from non-retryable API errors, not noise-related skips.

systematically under-estimated the model, and the full-coverage number restores a realistic picture. Gemini 2.5 Pro (0.509) is essentially tied with GPT-4o Audio within CI overlap; both are mid-pack among audio-natives, below all three Claude text-reasoner controls.

5.6 Noise-Robustness (Axis 5)

Table 6 presents WER under additive MUSAN/WHAM noise across four SNR levels (20, 10, 5, 0 dB) on a balanced 200-sample subset drawn from FLEURS-en and VoxPopuli-en. Each clean sample is paired deterministically with a noise clip (hash of sample id $\times$ SNR), and the mixture is rendered at each target SNR. Seven models are evaluated in this release: 5 dedicated ASR systems and 2 audio-native MLLMs (GPT-4o-mini Audio, Gemini 2.5 Flash). The MLLM slot on this axis deliberately evaluates the cost-tier cheaper variants: noise robustness is a capability a downstream operator would exercise on whichever tier they can afford to run at per-minute audio volume, and pairing GPT-4o-mini Audio with Gemini 2.5 Flash isolates noise-handling from model-tier capacity effects surfaced elsewhere in this benchmark.

Finding 9: AssemblyAI Universal-2 has the flattest noise-robustness cliff. Among the five dedicated ASR systems, AssemblyAI Universal-2 shows the smallest 20-to-0 dB WER increase at +0.030, about 1.5$\times$ better than the next-best model (ElevenLabs Scribe, +0.041) and about 1.6$\times$ better than the two Deepgram Nova systems (both +0.048/+0.049). At 0 dB it also reports the lowest WER of any faithful transcriber at 0.275 (95% CI 0.256–0.295). Its 20 dB WER (0.245) is mid-pack, so this is a noise-robustness finding specifically, not a general-purpose transcription ranking.

Finding 10: Gemini 2.5 Flash shows a steeper noise cliff than any dedicated ASR. Gemini 2.5 Flash is competitive with dedicated ASR at 20 dB (WER 0.239, tied with Whisper at 0.233 and AssemblyAI at 0.245), but degrades to 0.312 at 0 dB — a Δ of +0.073, roughly 1.5$\times$ the steepest dedicated-ASR cliff (Deepgram Nova-2, +0.049). Response-pattern analysis at 0 dB shows Gemini continues to attempt faithful transcription on 95% of samples (2% refuse, 3% return very short outputs), so the cliff is a pure transcription-quality degradation rather than a refusal spike. This is consistent with the hypothesis that audio-native MLLMs, which share parameters between speech recognition and a much broader multimodal task distribution, are less noise-robust than specialist ASR; the backfill on GPT-4o Audio and Gemini 2.5 Pro will test whether this extends to the frontier-scale MLLMs.

Finding 11: GPT-4o-mini Audio is task-broken for verbatim transcription. At every SNR level, including the near-clean 20 dB condition, GPT-4o-mini Audio returns paraphrase/summary responses for roughly half of all samples: 49% at 0 dB are summaries of the form "The speaker is describing ..." or "The text appears to be ...", 50% are transcription attempts, and 1.5% are refusals. This behavior is stable across 20, 10, 5, and 0 dB — it is not a noise failure mode. The verbatim prompt ("`Transcribe the spoken audio verbatim. Output only the transcript text.`") is the same prompt used across all transcribers. The resulting WER exceeds 2.0 at every SNR because the paraphrased responses contain many words that are not in the reference transcript; the number is a real measurement but should not be read as a signal about noise robustness. We include the row to document the behavior, not to rank the model.

6 Discussion

6.1 What this release establishes

Audio understanding accounts for the SLURP-intent and AMI-reasoning gap. On both tasks, when models receive the reference transcript rather than audio, the 27 pp SLURP-intent gap (Finding 5) closes and the 12.4 pp AMI token-F1 gap (Finding 8) closes. We attribute the remaining reasoning differential to transcript normalization effects, not audio. Given a reference transcript, today's frontier language models already reason well enough to classify SLURP intent and answer AMI reasoning questions; the remaining error is in converting the audio into a reliable representation in the first place.

Frontier ASR and audio-native MLLMs are close on transcription, diverge on downstream tasks. On pure transcription (Finding 1), the best dedicated ASR (Scribe 0.408 WER) and the best audio-native MLLM (Gemini 2.5 Pro 0.411) are within a 0.003 WER point-estimate gap with heavily overlapping CIs. But on SLURP intent and entity (downstream tasks that presumably benefit from richer audio understanding), the ASR→Claude cascade outperforms every audio-native MLLM we evaluated. The "audio-native MLLMs can skip ASR" thesis does not yet pay off at the downstream-task level.

Refusals and reliability are first-class production concerns. GPT-4o audio's 20% refusal rate on benign SLURP transcription requests (Finding 6) and its CJK timeout tail on multilingual transcription (Finding 2) both show up as F1 0 or WER 1.0 in the raw numbers, but a deployment team would treat them as separate failure modes requiring separate mitigations. Single-number leaderboards obscure that distinction.

Emotion is not yet production-usable. All four audio-native MLLMs sit well below 0.5 weighted F1 on 7-class MELD (Finding 7). Downstream use cases that rely on emotion as a routing signal should plan for human-in-the-loop review.

7 Limitations

1. **AMI reasoning questions are Claude-generated.** The 248-question AMI reasoning set was synthesized by Claude from AMI meeting transcripts and audited by the authors, not crowd-written. Absolute Claude-family scores on this axis (Opus 0.560, Sonnet 0.555, Haiku 0.636 token-F1) should be read as partly reflecting self-annotated-test bias: the generator and the evaluator share a prior over question form and expected-answer shape. Relative scores across the audio-native MLLMs (GPT-4o audio, Gemini 2.5 Flash) are less affected and are the primary comparison we draw from this axis. We do not make absolute-quality claims about the Claude text-reasoner rows on AMI.
2. **Partial coverage on Gemini 2.5 Flash SLURP.** After the 2026-04-20 day-3 backfill, Gemini 2.5 Pro SLURP is at $n{=}597/600$ (99.5% coverage; 288 of the 291 original Google-side 429/503 rows recovered, 3 residual 500s). Gemini 2.5 Flash remains at $n{=}382/600$ because a sustained Google-side 503 upstream-saturation block on the Flash endpoint persisted through the day-3 retry window. AMI partial coverage (previously GPT-4o Audio at $147/248$ and Gemini 2.5 Pro deferred) was resolved by the 2026-04-18 day-2 backfill; Table 5 numbers are full-coverage.
3. **Refusals combined with audio-understanding errors on SLURP.** Our intent/entity scoring counts a refusal as F1 0; this is consistent with the deployment view (a refusal is a failure) but combines prompt-following and audio-understanding failure modes. The refusal column in Table 3 partially separates them.
4. **Partial coverage on the noise-robustness axis.** The noise-cliff axis (Section 5.6, Table 6) reports 7 of the 9 candidate models: 5 dedicated ASR + GPT-4o-mini Audio + Gemini 2.5 Flash. GPT-4o Audio (full-size) and Gemini 2.5 Pro are omitted because their API caps were saturated by the concurrent SLURP and GlobalVoice-accent runs. The dedicated-ASR and Flash cliffs reported here are final.
5. **Transcription source-corpus scope.** Transcription evaluation uses FLEURS and VoxPopuli as the clean source, then adds MUSAN/WHAM augmentation for the noise axis. Real-world heavily-noisy corpora (CHiME-style field recordings) are not included; the MUSAN/WHAM augmentation protocol is a controlled stand-in with a known SNR, not a replacement for in-the-wild evaluation.

6. **Training-data contamination.** FLEURS, VoxPopuli, SLURP, MELD, and AMI are all well-known public corpora. Some overlap with model training data is plausible across the 12 models and would inflate their numbers differentially. We do not report a contamination analysis in this release.

7. **Normalization and scoring choices.** Intent-label normalization uses longest-prefix match against the canonical 18 scenarios; entity extraction prefers bracket-format parsing when present and falls back to regex. Different choices will shift absolute numbers but are unlikely to change model ordering.

8. **Subset size per axis.** SLURP and MELD use 200 stratified samples each; AMI uses 248 questions; transcription uses a 200-sample subset. These sizes give reasonable CIs on aggregate per-model scores but are too small for reliable per-scenario or per-class rankings. We do not make per-class claims in the paper body.

9. **Independent human validation.** We rely on published reference transcripts (FLEURS, VoxPopuli, SLURP, MELD) and author-audited synthetic questions (AMI). We do not re-annotate.

8 Ethical Considerations

All source data is drawn from openly licensed corpora (FLEURS, VoxPopuli, SLURP, MELD: CC-BY-4.0; AMI Meeting Corpus: CC-BY-4.0). No personally identifiable information beyond speaker audio is present; these corpora do not link speakers to identifying metadata. 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 FLEURS and VoxPopuli teams for the speech corpora we rely on in this release. This work was supported by Datoric Labs.

Use of AI Assistants. This paper was prepared with the assistance of Anthropic's Claude (Opus / Sonnet / Haiku 4.5). Claude was used in four distinct roles: (1) drafting and copyediting portions of the manuscript and generating Python code for the analysis, figure, and scoring pipelines; (2) serving as the text-reasoner upper-bound control on the reference transcript for the transcription, SLURP intent/entity, and AMI reasoning axes; (3) synthesizing the 248 AMI reasoning questions (see Limitation 7, item 1); and (4) serving as the downstream classifier in the ASR→LLM cascade for SLURP intent/entity on dedicated ASR providers (Claude Haiku 4.5). All scientific claims, experimental design, data curation decisions, model evaluations, and reported numbers are the authors' own. Every number reported in this paper is computed from the released JSON result artifacts enumerated in Appendix C, and the authors take full responsibility for the paper's content.

References

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)*.

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)*.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, and 1 others. 2005. The AMI meeting corpus: A pre-announcement. In *International Workshop on Machine Learning for Multimodal Interaction (MLMI)*. Springer.

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

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.

Gemini Team, Google. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Yixuan Hou, Heyang Liu, Yuhao Wang, Ziyang Cheng, Ronghua Wu, Qunshan Gu, Yanfeng Wang, and Yu Wang. 2025. SOVA-Bench: Benchmarking the speech conversation ability for LLM-based voice assistant. *arXiv preprint arXiv:2506.02457*.

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

Dhruv Jain, Harshit Shukla, Gautam Rajeev, Ashish Kulkarni, Chandra Khatri, and Shubham Agarwal. 2025. VoiceAgentBench: Are voice assistants ready for agentic tasks? *arXiv preprint arXiv:2510.07978*.

OpenAI. 2024. GPT-4o system card.

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)*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*.

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.

Muhammad A. Shah, David Solans Noguero, Mikko A. Heikkila, Bhiksha Raj, and Nicolas Kourtellis. 2025. Speech robust bench: A robustness benchmark for speech recognition. 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)*.

Ke Wang, Houxing Ren, Zimu Lu, Mingjie Zhan, and Hongsheng Li. 2025a. VoiceAssistant-Eval: Benchmarking AI assistants across listening, speaking, and viewing. *arXiv preprint arXiv:2509.22651*.

Yingzhi Wang, Anas Alhmoud, and Muhammad Alqurishi. 2025b. Open universal Arabic ASR leaderboard. In *Proceedings of Interspeech*.

A Datasheet for VoicePro-Bench

Following Gebru et al. (2021), we provide a datasheet for this release.

Motivation. VoicePro-Bench was created to evaluate ASR providers, audio-native multimodal LLMs, and text-reasoner controls on the same samples across five axes of professional voice understanding.

Composition. Five filled axes: 760 transcription samples from FLEURS (accented and English) and VoxPopuli; 200 SLURP samples (intent + entity); 200 MELD samples (emotion); 248 AMI-grounded reasoning questions; and 200 noise-augmented samples (each rendered at 4 SNR levels) for the noise-robustness cliff.

Collection Process. Audio inherits from FLEURS, VoxPopuli, SLURP, MELD, and AMI releases. AMI reasoning questions are synthesized by Claude and author-audited.

Preprocessing. Audio resampled to 16 kHz mono. Transcription text normalized with Whisper's `BasicTextNormalizer` (Latin scripts) or CER (CJK). Intent labels normalized by longest-prefix match against SLURP's 18 scenarios. Entity extraction prefers SLURP `[slot : value]` bracket parsing, falling back to a regex extractor.

Distribution. Released on HuggingFace under CC-BY-4.0 (metadata and annotations; audio inherits source corpus licenses).

Maintenance. Rate-limit backfills restored coverage on three axes: AMI reasoning (GPT-4o Audio and Gemini 2.5 Pro), SLURP (Gemini 2.5 Pro n=597/600; Gemini 2.5 Flash n=382/600 residual due to a sustained Google-side 503 saturation of the Flash endpoint), and transcription (GPT-4o Audio n=200/200, Gemini 2.5 Pro n=198/200, Gemini 2.5 Flash n=195/200; 7 residual 504 deadlines). All other cells are at or near full per-axis n, with small pre-existing shortfalls on Nova-3, Nova-2, and Claude Opus disclosed in Table 2. The noise-robustness axis evaluates cost-tier cheaper MLLM variants (GPT-4o-mini Audio, Gemini 2.5 Flash) by design, not due to a backfill gap, for the reason given in §5.6.

B Dataset Statistics

Accent groups in FLEURS-accented (100 samples each except Japanese at 60): Mandarin, German, Spanish, French, Hindi, English (US), English (native), Japanese.

Table 7: Dataset composition by source corpus and difficulty tier (from `data/curated/curation_stats.json`).

Source	Easy	Medium	Hard	Total
FLEURS (accented, 6 groups)	30	410	120	560
FLEURS (English)	10	70	20	100
VoxPopuli (English)	10	67	23	100
Total	**50**	**547**	**163**	**760**

C Replication

All numbers in this paper are computed from files in `voicepro-bench/results/` and the curated data in `voicepro-bench/data/curated/`:

- **Transcription** (Table 2): `aggregated_20260422_transcription_backfill.json` (post day-4 transcription rate-limit backfill); reproduce via `python eval/merge_transcription_backfill.py` (merges `run_20260422_transcription_ratelimit_patch.json` into `run_20260414_204217.json`, then re-aggregates).
- **SLURP intent + entity** (Table 3): `aggregated_20260420_slurp_day3.json` (post day-3 backfill); reproduce via `python eval/merge_slurp_into_aggregated.py –slurp-run run_slurp_20260417_120353.json –prior-aggregated aggregated_20260418_ami_backfill.json –output aggregated_20260420_slurp_day3.json`. The day-3 patch delta is `run_slurp_20260419_ratelimit_patch.json`.
- **MELD emotion** (Table 4): `aggregated_20260417_final.json` (`.emotion_meta` block); reproduce via `python eval/score_emotion.py` on the MELD checkpoint directory.
- **AMI reasoning** (Table 5): `ami_reasoning_scored_final.json`; reproduce via the AMI scoring script documented in the release.