

The Harlow Hospitality ASR Benchmark: Measuring Speech-to-Request Reliability in Hotel Conditions

Kristian Dhimitri · Jaroslavs Benedikts Bolsakovs

Harlow Technologies, Inc. · support@tryharlow.com

Abstract

Speech recognition for hotel operations has a different success criterion from transcription: a spoken staff request is useful only if the room, the person, the item, the count and the negation survive, and the costliest failure is a request that proceeds with one of them wrong while nobody is told. Word error rate measures neither property. We introduce a task-level benchmark for spoken hotel-staff requests whose headline metric is the *silent-wrong rate*, the share of requests that proceed with a wrong or missing fact, and whose protocol compares systems at matched repeat rates so that a system cannot buy safety by asking the speaker again more often. The benchmark has two tracks: 2,920 synthetic request clips from one fictional property under ten acoustic conditions and four line types, and 1,386 clips from 77 real speakers with eleven first languages. A fact-level scorer is blind to formatting and strict about facts; confidence gates for standalone recognizers are tuned on held-out folds; and every system is scored on its own output, with nothing added by the benchmark. We evaluate five commercial recognizers from three vendors, each plain and each with the property's vocabulary, and Harlow, the authors' production system, which returns either a transcript or a request to repeat and is evaluated as a black box. With the vocabulary and at matched repeat rates, Harlow proceeds silently wrong on 2.6% of hotel requests against 5.1% for the best gated single recognizer (paired difference 2.5 points, 95% interval 1.6 to 3.5), and on 3.4% against 4.5% of real accented clips (1.1 points, 95% interval 0.1 to 2.1). It is ahead or level in every acoustic condition and line type, most clearly in heavy noise, with a guest talking nearby and on mixed-language requests. On hotel requests the property vocabulary is the largest single lever for every recognizer, cutting silent-wrong rates 1.6- to 5.1-fold; on real speakers it helps every recognizer by 1.3- to 1.5-fold but the choice of recognizer matters at least as much. Number formatting is a second lever of the same kind, which Harlow applies itself: reading rooms that a recognizer wrote as clock times narrows the headline gap to 2.6% against 4.1% (1.5 points, 95% interval 0.8 to 2.2) and takes MAI-Transcribe 1.5 from 28.0% to 8.1% silent-wrong. Word error rate and silent-wrong rate rank the systems differently. Two further conditions that cut the first 200 ms or the last 300 ms of speech, which the capture of the badge the system ships on does not produce, are reported separately as a stress test; there every system fails most often on a cut end and Harlow stays ahead (18.4% against 23.8%). The benchmark is released with each system's verdict for every clip, so that the silent-wrong, repeat and correct rates of its tables can be recomputed, and with the audio of the real-speaker clips recorded without added noise; the synthetic speech is available to researchers on request.

HIGHLIGHTS

- 2.6% vs 5.1% silent-wrong on hotel requests, Harlow vs the best gated single recognizer, at 5.3% and 5.5% repeats, each scored on its own output; paired difference 2.5 points, 95% interval 1.6 to 3.5.
- 3.4% vs 4.5% on real accented speakers at about 24% repeats; the ordering holds, with a paired difference of 1.1 points, 95% interval 0.1 to 2.1.

- 35% → 7% silent-wrong for Scribe v2 from the property vocabulary alone; on hotel requests every recognizer does better with it than the best recognizer does without it.
- WER does not predict fact-level reliability. MAI-Transcribe 2 has the lowest WER of any single recognizer on hotel requests (0.071) and files 18% of requests silently wrong; Scribe v2, at a WER of 0.152, files 7%.
- Post-processing is a first-order effect. Reading rooms that were written as clock times moves MAI-Transcribe 1.5 from 28% to 8% silent-wrong and MAI-Transcribe 2 from 18% to 8%. Harlow's output already writes rooms as rooms; given the same step, gated Scribe v2 would read 4.1% against Harlow's 2.6% (§6.8).

Conflict of interest. Harlow is the production system of the authors' company, and the authors are its developers. It is evaluated as a black box: its design is proprietary and is not described in this report, and it may build on third-party recognition services, including services evaluated here as standalone systems. Harlow was in production before this benchmark existed, and its development continued while earlier versions of the benchmark were in use; those versions share the property, templates, voices, noise sources and the 120 request lines with this one. The benchmark is therefore not independent of the system, and only the 120 further lines (self-corrections, short commands, mixed language) are new in this version. Every gate is tuned in the standalone systems' favour where a choice exists; each system's verdict for every clip is released, while the system itself is not (§5, §8).

1 Introduction

Automatic speech recognition (ASR) is evaluated almost universally by word error rate (WER) on read or conversational corpora [1, 2, 3]. For a growing class of applications the transcript is not the product; it is an intermediate representation from which a small number of facts are extracted and acted upon. A hotel is a clear instance. A staff member wearing a microphone says “Room 1214 needs two extra towels for Mrs. Okonkwo”, and the system files a task against a room, for a person, with a count. The utterance is a few seconds long, spoken in a corridor, a kitchen or a lobby, frequently by a speaker whose first language is not English, sometimes with a guest talking a metre away, and its value depends entirely on whether four or five tokens survive intact. A transcript with a 5% WER can file every request against the wrong room; a transcript with a 20% WER can file every request correctly.

This mismatch has two consequences that a benchmark for the domain must encode. First, the primary metric must be defined on facts and must distinguish a wrong answer that proceeds from a system that declines to answer. We call the former *silent-wrong*: the request went ahead, a fact was wrong or missing, and no one was told. A system that asks the speaker to repeat is not wrong; it is slower. Second, because any system can lower its silent-wrong rate by declining more often, systems must be compared at equal repeat rates, and the full risk–coverage trade-off [10, 11] should be reported rather than a single point.

We make three contributions. (i) A benchmark for spoken hotel-staff requests with a fact-level scorer, two tracks (synthetic requests under controlled acoustic conditions, and real accented speakers), and a protocol that equalises vocabulary and repeat budget across systems and tunes confidence thresholds on held-out folds (§3). (ii) An evaluation of five commercial recognizers from three vendors, each with and without the property's vocabulary (§6). (iii) An evaluation of Harlow, the production system of the authors' company, as a black box that either proceeds with a transcript or asks the speaker to repeat (§4). At matched repeat rates, each system scored on its own output, it files 49% fewer silently wrong hotel requests than the best gated single recognizer and is ahead by a smaller margin, with a paired interval that excludes zero, on real accented speech. Figure 1 summarises the result.

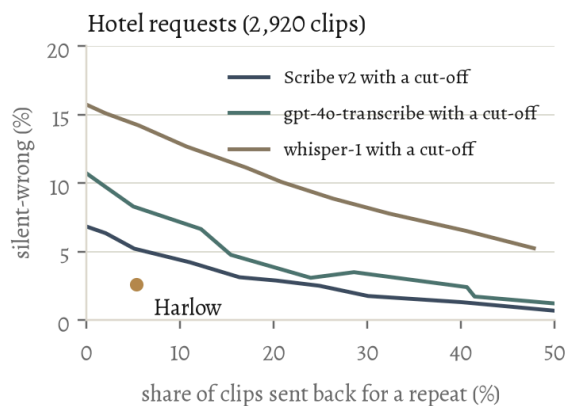


Figure 1a. Track A, hotel requests. Each line is one recognizer with a confidence gate whose threshold is set on the other fold (\$3.5), drawn at the share of clips it actually sent back: further right, more clips sent back, fewer silent mistakes. Harlow is a single point: it decides itself when to ask and exposes no threshold to sweep. Scribe v2 needs about 24% repeats to reach Harlow’s silent-wrong rate at 5.3%.

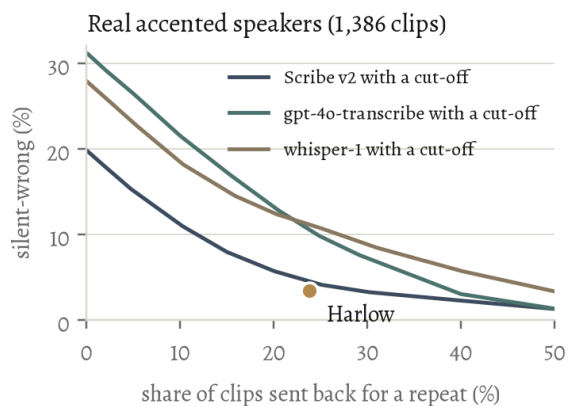


Figure 1b. Track B, real accented speakers. Harlow asks for a repeat far more often here (23.8%); at that budget the best gated recognizer is at 4.5% against 3.4%.

2 Related work

Benchmarks and metrics. LibriSpeech [2], GigaSpeech [3] and the CHiME challenges [4] established WER on read, broadcast and distant conversational speech; CHiME in particular drove progress on real multi-speaker noise. Morris et al. [5] argued early that WER is a poor proxy for the information a transcript conveys. SemDist [20] scores a hypothesis by its semantic distance from the reference, and Earnings-21 [21] compares commercial recognizers by WER restricted to named entities. Spoken language understanding benchmarks such as Fluent Speech Commands [22], SLURP [23] and STOP [24] go further and score intents and slots rather than words, which is the right criterion for an assistant. Our benchmark is in that tradition, with three differences that the domain imposes: the slots are hotel facts with a closed schema (rooms, roster names, counted items, hospitality terms, negation), the primary quantity is the rate of confidently wrong slot values rather than slot accuracy, and declining to answer is a first-class outcome compared at matched rates.

Noise and accent. We use the DEMAND corpus [6] for real environmental noise and the Speech Accent Archive [7] for real non-native speech; both are standard resources, and we discuss their limits in §8. Whisper [1] and its successors are trained for robustness at scale, and the vendors evaluated here advertise robustness to noise and accent [17, 18, 19]; the benchmark measures whether that robustness extends to the facts.

Contextual biasing. Supplying a recognizer with domain terms at inference time, whether as a bias list in the decoder [12] or as a text prompt [1, 18], is well established. Every vendor evaluated here exposes such a channel, with different capacities [17, 18, 19]; §6.2 quantifies what it is worth in this domain.

Confidence, selective prediction and combination. Confidence estimation for ASR has a long literature [9]; selective classification formalises the trade-off between coverage and risk when a system may abstain [10, 11], and our risk-coverage curves follow that framing. Combining recognizers to reduce error dates to ROVER [8], and disagreement within an ensemble is a standard uncertainty signal [25]. Language models are also used as judges of whether two texts carry the same meaning [16]. Code-switched speech, present in our mixed-language line type, remains a recognised difficulty for monolingual recognizers [15].

3 The benchmark

3.1 Task definition

A benchmark instance is a short audio clip x with an expected text and a fact record F : a set of room numbers, a set of must-appear phrases (names, items with counts, hospitality terms), and a negation flag. A system produces either a transcript y or a request to repeat. A transcript is *correct* if every fact in F is recoverable from y under the normalisation of §3.4; otherwise it is *silent-wrong*. The task is deliberately not “transcribe the audio”; it is “either get every fact right or say so”.

3.2 Track A: hotel requests

Track A is a synthetic, fully controlled corpus from one fictional property with 64 rooms on eight floors, ten staff members, fifteen in-house guest surnames and twenty hospitality terms (rollaway, PTAC, turndown, bell cart, late checkout, do not disturb, ...). From seeded templates we draw 120 request lines; the same lines were used in earlier versions of this benchmark (see the conflict-of-interest statement). Each line is spoken by a native US voice and by an accented voice (Spanish, Tagalog, Russian, Mandarin, Vietnamese, Indian, Nigerian, Haitian, Portuguese), and every recording is rendered under the ten main acoustic conditions of Table 1, and under two that cut speech short, reported separately as a stress test (§6.4); the voices and their synthesis models are listed in Appendix F. Three further line types (Table 2) probe behaviours a request corpus of full sentences would miss: self-corrections, two-word commands, and Spanish- or Tagalog-mixed sentences spoken by voices of those languages; each has 40 lines under four of the conditions. Table 3 gives the counts.

Table 1. Track A acoustic conditions. Noise is mixed at a fixed signal-to-noise ratio against the speech’s RMS from real DEMAND recordings [6]: the dB figure is how much louder the speech is than the noise, so at 0 dB the noise is as loud as the speaker and at 10 dB the speaker is clearly louder; the stairwell is a synthetic room impulse response; the cuts (stress test, §6.4) are measured from the detected speech edges, so every voice loses the same amount; the bystander is a second synthetic voice starting 0.4 s into the clip.

CONDITION	DEFINITION
clean	The synthesised voice, 16 kHz mono
hallway-snr10	DEMAND OHALLWAY, speech 10 dB louder than the noise
lobby-snr5	DEMAND PCAFETER (cafeteria crowd), speech 5 dB louder than the noise
restaurant-snr0	DEMAND PRESTO, noise as loud as the speech (0 dB)
kitchen-snr5, kitchen-snr0	DEMAND DKITCHEN (dishes, extractor fan), speech 5 dB louder than the noise, or as loud (0 dB)
stairwell-reverb	Synthetic impulse response, ~1.2 s RT60
cut-start, cut-end (stress test)	First 200 ms or last 300 ms of speech removed; outside the main set (§6.4)
guest-nearby	A guest’s voice 12 dB below the wearer; half the guest lines name a different room
guest-background	The same guest at 18 dB below
guest-nearby+lobby	Guest 12 dB below the wearer and the lobby crowd 5 dB below the speech

Table 2. Track A line types. The 120 request lines run under every condition in Table 1, the cuts only in the stress test; the three further types run under clean, lobby-snr5, kitchen-snr0 and guest-nearby.

LINE TYPE	EXAMPLE	LINES
Request	Room 1404 needs four foam pillows.	120
Self-correction	Room 504, sorry, 506 needs three bathrobes.	40
Short command	1406 clean. · Crib to 301.	40

LINE TYPE	EXAMPLE	LINES
Mixed language	Oye Tomasz, la habitación 1105 needs five towels, por favor.	40

Voices alternate between two synthesis vendors, Azure and OpenAI, because each is the sibling of a recognizer under test (Azure voices for Microsoft's recognizers, OpenAI voices for gpt-4o-transcribe and whisper-1); results can be split by synthesis vendor to expose a home-turf advantage. An intelligibility screen removes synthetic voices that are not accents but failures of synthesis: a voice is dropped if the median standalone recognizer, without any vocabulary, gets at most 20% of its clean clips correct. The screen uses only the standalone recognizers, so it cannot favour Harlow. It removed three Azure voices (Appendix D), leaving 2,920 scored clips of the 3,200 rendered.

Table 3. Corpus statistics. Rendered counts come from the generator; scored counts exclude the three synthetic voices removed by the intelligibility screen (§3.2). A voice configuration is a synthesis voice with its accent instruction; four OpenAI voices appear twice, once native and once instructed to speak with an accent. Each kept real speaker contributes six segments under three conditions.

SET	LINES	VOICES OR SPEAKERS	CONDITIONS	RENDERED	SCORED
TRACK A: HOTEL REQUESTS (SYNTHETIC; GENERATOR SEED 20260930)					
Request	120	24 (21 after screen)	10	2,400	2,180
Self-correction	40	24 (21 after screen)	4	320	292
Short command	40	24 (21 after screen)	4	320	288
Mixed language	40	4	4	160	160
All Track A	240	24 voice configurations (8 native, 16 accented; 21 after screen); the mixed-language lines use the Spanish and Tagalog pairs among them; 4 bystander voices	10	3,200	2,920
Stress test: speech cut short (§6.4)	120	as Request	2	480	436
TRACK B: REAL SPEAKERS (SPEECH ACCENT ARCHIVE)					
Six segments per speaker	6 texts	77 speakers of 89 downloaded (English 10, Arabic 8, French 8, Hindi 8, Mandarin 7, Portuguese 7, Tagalog 7, Amharic 6, Russian 6, Spanish 5, Vietnamese 5)	3	1,386	1,386
AUDIO (MEASURED BY AUDIO_QA.PY ON THE CLIPS OF THE REPORTED RUN)					
Track A duration, s	median 4.65	5th–95th percentile 2.88–6.75	speech median 4.26	288.9 min total	7 clipped
Track B duration, s	median 4.72	5th–95th percentile 2.51–8.53	speech median 4.41	114.2 min total	2 clipped

3.3 Track B: real accented speakers

Synthetic accents are a proxy. Track B uses the Speech Accent Archive [7]: speakers with English and ten other first languages (Spanish, Tagalog, Russian, Mandarin, Vietnamese, Hindi, Arabic, Portuguese, French, Amharic) reading the same “Please call Stella” paragraph. Up to eight speakers per language and ten for English were downloaded, 89 in all. The paragraph (about 28 s for the median speaker) is cut into six clips of 1.8 to 14.3 s (median 4.7 s) at sentence and comma boundaries, in the middle of the pause after the boundary word. Cut points come from the word timings of one ElevenLabs Scribe v2 pass over the clean recording, without any vocabulary; the expected text never does. Twelve speakers for whom that pass did not recover a boundary word are left out, which leaves 77 speakers: ten for English, eight each for Arabic, French and Hindi, seven for Mandarin, Portuguese and Tagalog, six for Amharic and Russian, and five for Spanish and Vietnamese (Table 3). Because the segmenting recognizer is itself a system

under test, this selection is a threat to validity that we discuss in §8: the twelve excluded speakers are exactly those Scribe v2 found hardest. Each clip has its own fact record (names, items with counts, a day, a place) and runs under clean, lobby-snr5 (crowd 5 dB below the speech) and restaurant-snr0 (noise as loud as the speech): 1,386 clips. Speakers who misread a sentence are scored against the printed text and therefore count against every system equally; absolute rates on Track B are consequently higher than the speakers' true error rate, and its value is comparative.

3.4 Metrics

For a system and a set of N scored clips (vendor errors excluded and reported separately), each clip lands in exactly one of three buckets, correct, silent-wrong or repeat, so that

$$C + SW + R = 1, \quad (1)$$

where SW is the **silent-wrong rate**, the fraction of all scored clips on which the system proceeded with at least one fact wrong or missing; R is the **repeat rate**; and C the correct rate. We report SW over all clips rather than conditioned on proceeding, because the hotel pays for every silently wrong task regardless of how many requests were sent back. The conditional error among proceeded clips, $SW/(1 - R)$, is given where it clarifies a comparison. Fact recall, the fraction of facts in F recoverable from y , and WER on normalised tokens are secondary metrics.

Intervals are 95% percentile bootstrap intervals over clips with 1,000 resamples [13, 14]. They are marginal, per system, and treat clips as independent; clips that share a line (Track A) or a speaker (Track B) are correlated, so these intervals are narrower than a cluster bootstrap would give, and a comparison of two systems on the same clips is better served by a paired test. For the headline comparisons we therefore also report the paired difference in silent-wrong rate between two systems on the same clips, with a percentile interval from 2,000 resamples of whole sentences (Track A) or speakers (Track B). Elsewhere we report whether marginal intervals overlap and claim nothing beyond that.

The scorer is lenient about form and strict about facts (Appendix B). Number words and digits are equal; a room matches however its digits are grouped ("twelve fourteen", "12 14", "1214"); hyphenation, spacing and plurals are ignored. A room rendered as a clock time or date ("12:14", "12/14") does not match the room, because whatever consumes the transcript would read a time. That rule turns a formatting habit into an error, and writing rooms as rooms is part of what a system returns; §6.8 measures what adding a room step to the recognizers' transcripts is worth.

3.5 Protocol and fairness rules

Same vocabulary. Every system runs twice: plain, and with the property vocabulary (122 terms on Track A, 99 on Track B): the property's own names and terms first, then a list of hospitality terms. Each vendor receives the list through its own biasing field, cut to what that field accepts (Appendix C, with each vendor's documented capacity [17, 18, 19]). The plain run is an ablation that also represents a property on its first day. On Track B the property terms are replaced by the paragraph's names and items, so the vocabulary arm there is told the facts it must recover in advance; this holds for every system equally and is discussed in §8.

Same repeat budget. A standalone recognizer may send back its least-confident clips, as many as Harlow repeats on the same track. Let $s(x)$ be an aggregate of the recognizer's per-word, per-token or per-segment confidences, either the minimum or the mean; the gated system repeats when

$$s(x) < \tau, \quad \text{with } \tau \text{ such that } R_{\text{gated}} = R_{\text{Harlow}}. \quad (2)$$

The aggregation and τ are chosen on one fold and applied to the other, where folds split Track A by sentence and Track B by speaker (odd versus even index), so a threshold never sees the clip, or the same sentence in another condition, that it is scored on. The risk–coverage curves of Figure 1 and Table A1 use the same held-out procedure at

each repeat budget. In-sample gates, the best case for the standalone system, appear only in the scorer's own text report. On Track A the parity of a line is also its synthesis vendor (even lines are Azure voices, odd lines OpenAI voices, the mixed-language lines aside), so each threshold is tuned on one vendor's voices and applied to the other's. The headline does not depend on it: for Scribe v2 with the vocabulary the scorer's held-out gate gives 5.10% silent-wrong, the in-sample gate 5.17%, and a held-out gate whose folds split each vendor's lines 5.14%. MAI-Transcribe reports a confidence of zero under its enhanced mode, so neither MAI-Transcribe recognizer can be gated; the benchmark reports this rather than constructing a gate for them.

Same stimuli. Every system hears the identical file. Synthesised takes are cached per voice, instruction and text so that the conditions of a line differ only by their mixing. In the clips of the reported run that holds for every Azure voice; for OpenAI voices a defect in the renderer left most mixed clips on another synthesis of the same sentence than their clean clip, which §3.6 measures.

Each system's own output. Every system, Harlow included, is scored on what it returned for the clip, and the benchmark adds no processing to any of them. Recognizers differ in how they write a room number, and some write it as a clock time ("12:05" for room 1205), which the scorer reads as a time (§3.4). Harlow's output already writes rooms as rooms; the standalone recognizers' output does not always, and they lose the clips on which they wrote the right room as a time. The scorer as written also corrects a misheard roster name in Harlow's transcripts by emulating a later step of the production system outside it; that step is not part of the output the benchmark recorded, so it is not applied here. A deployment of a standalone recognizer could add a room step of its own: a time-shaped token whose digits are a room on the property's list is read as that room unless a time word stands beside it. §6.8 reports the comparison with that step given to every recognizer, using the room list of the corpus (63 rooms), and as the scorer writes it. Our implementation of the step was checked against Harlow's own outputs: on each of the 610 transcripts tested it leads the scorer to the same verdict, so it changes nothing for Harlow. Gates, intervals and the intelligibility screen are computed under each treatment with the scorer's own functions, and the screen removes the same three voices.

3.6 Corpus quality control

Every clip the reported run was scored on is kept with the benchmark, and each was measured (Table 4). Where a mixed clip is built on the same synthesis as the clean clip of its line, subtracting the clean clip leaves exactly what was added, and the signal-to-noise ratios recovered that way match their targets to within 0.56 dB, bystander levels to within 0.24 dB, and the cut conditions remove the intended 200 and 300 ms beyond the detected speech edge exactly. 127 clips contain at least one sample at full scale; in 7 Track A clips and 2 Track B clips more than 0.1% of the samples are, at most 0.18%, all in the loudest noise conditions. No clip is silent and no two clips are identical. Track A clips have a median duration of 4.65 s (5th to 95th percentile 2.88–6.75 s), Track B clips 4.72 s (2.51–8.53 s).

A recognizer can return words for noise, so that each clip holds a voice is tested on the signal itself, without a recognizer (Appendix G). A 32 ms frame counts as voiced when its autocorrelation has a true peak at a pitch lag (70 to 450 Hz), and a file as speech when at least a fifth of its active frames are voiced and the pitch moves by half a semitone or more. White, pink and brown noise, a tone, a hum, digital silence and the kitchen and hallway noise beds all fail. The cafeteria and restaurant beds contain distant voices and pass in some five-second windows (12 of 41 for the cafeteria bed, 17 of 41 for the restaurant bed), which is the limit of such a test under those two beds. All 440 clean synthetic clips, all 544 cached takes, all 89 source recordings of real speakers and all 462 clean real segments pass, with at least 24% of frames voiced. Of the 5,066 clips in all, 5,060 pass; the 6 that do not are real-speaker segments under lobby or restaurant noise with 14 to 19% of frames voiced, just under the threshold, and each contains its clean segment at the target ratio by subtraction. Independently of the acoustics, at least one of the twelve arms transcribed every fact correctly on 97.6% of the Track A clips of the kept voices. Of the 80 that none gets right, 38 are cut-end clips, whose fact was removed with the audio. The rest include two mixed-language lines on which one take of the synthesised voice says another room than the label ("dos ocho siete" for 207, "trío cero" for 304),

so that clips of one line built on different takes differ in what is said; they are not removed and count against every system alike. On Track B no arm is right on 127 of 1,386 clips, 106 of them under restaurant noise.

Subtraction also identifies the synthesis each clip is built on, and it shows a defect in the renderer the clips were made with. Its synthesis cache was not locked, so the six render workers could each synthesise the same line at once and each keep its own result. Two requests to Azure for the same text return near-identical audio (correlation above 0.999), so its 1,620 mixed clips are on their clean clip's take in every sense that matters. OpenAI's does not: 1,404 of the 1,620 mixed OpenAI-voice clips (86.7%, or 43.3% of all 3,240 mixed Track A clips) are built on a different synthesis of their sentence, by the same voice, than the clean clip of their line, 978 of them on the take that was left in the cache and 426 on a take that was not kept; only 15 of the 220 OpenAI-voice lines have every condition on one take. With the two exceptions just noted, such a clip is still a rendering of its line at the stated level, because the mix is scaled to the take in hand: the 978 whose take is in the cache pass the same identity test against it, and at least one arm transcribed every fact correctly on 98.8% of the 426 whose take is not, against 99.1% of those whose take is. What is lost is that two conditions of an OpenAI-voice line differ by the take as well as by the mix, which adds variance to the per-condition comparisons of §6.3. Whether it also shifts them cannot be settled at this sample size: with line and condition as fixed effects, the correct rate of clips on another take differs from that of clips on their clean clip's take by -6.0 to +8.9 points across the twelve arms as scored, with standard errors of 4 to 6 points (clustered by line), and no difference exceeds 1.8 standard errors. It does not touch the comparison between systems, which hear identical files. The renderer now serialises synthesis per take, and rejects takes shorter than 0.5 s or quieter than -45 dBFS after a regeneration received 0.3 s of near-silence from the vendor for a full sentence; a regeneration with the corrected renderer has no clip on another take than its clean clip.

Every clean synthetic clip (one per line and voice configuration) was transcribed once by a recognizer not among the systems compared, gpt-4o-mini-transcribe, without any vocabulary, and scored with the benchmark's own fact check. With roster names included, 54.3% of them have every fact recoverable from that single pass; without the names, which that recognizer has no vocabulary for, 77.7% (native voices 90.5%, accented voices 67.1%); and 84.5% have a word error rate of at most 0.3 with room numbers read as one word however they are spoken (native 97.0%, accented 74.2%). The three voices the intelligibility screen removed score 0% on the fact measures, which confirms the screen; one kept voice, en-IN-NeerjaNeural, is borderline, and the per-voice figures are in Appendix E. Clips that fail this pass are not removed; the per-voice screen of §3.2 is the only exclusion. The same pass over the 462 clean real segments of Track B gives a word error rate of at most 0.3 on 96.5% of them (median 0), with 4 above 0.5.

Three further checks tie the reported numbers to these files. *The scores follow from the outputs.* Re-running the scorer on the stored per-clip outputs of the 12 arms reproduces its stored per-clip results (60,792 rows, 0 vendor errors), every point estimate of the scorer's report, and all 1,963 cells of the result set this paper is built from. *The outputs follow from the clips.* A seeded sample of 200 clips, 40 from each of five strata (real speakers, Azure voices, and OpenAI voices on each kind of take), was replayed through whisper-1 as the benchmark runs it; text and per-segment confidence came back identical to the stored output on 199 of them, confidences within 0.002; the exception is a clip of a voice the screen removed, on which the recognizer invents text. Another synthesis of the same sentence would not return the same confidences, so the stored outputs were produced from these files. The same replay through Scribe v2 settles nothing, because Scribe v2 is not repeatable: its words changed on 28 of the 200 clips and its score bucket on 5 (3 became correct and 2 wrong), which is a direct measure of the run-to-run noise in that arm (§8). *The clips follow from the generator.* An independent regeneration from the benchmark's code reproduced the plan row for row, the 5 noise files byte for byte, the 89 real recordings to within 1 quantisation step, and all 220 clean Azure takes as the same waveform. Track B segment boundaries come from a recognizer's word timings (§3.3) and 61 of 462 moved, by a median of 0.01 s and at most 1.52 s; the 1,203 Track B clips on unmoved segments came back as the same waveform, noise included. OpenAI takes are new syntheses each time; 82.0% of the regenerated clean clips are intelligible by the measure above, against 84.5% of those of the reported run. The per-clip measurements and lineage are kept with the benchmark.

Table 4. Audio quality control on the clips of the reported run (audio_qa.py, take_lineage.py). SNR and bystander level are recovered from each mixed clip minus the clean clip of its line, the level over the span where the bystander is present; cut amounts from the length difference against the detected speech edge. The recovery is exact only for a clip built on the same synthesis as its clean clip, and the deviations are over those clips. The fourth column counts the OpenAI-voice clips that are built on another synthesis (§3.6), in brackets those on the take left in the cache and those on a take that was not kept; Azure-voice and real-speaker clips are all on their own clean clip. 'Other flags' counts clips with any of: no speech, very quiet, clipping, a level more than 0.5 dB off its target, a cut amount more than 40 ms off, odd duration.

CONDITION	CLIPS	MEASURED, CLIPS ON THEIR CLEAN CLIP'S TAKE	OPENAI-VOICE CLIPS ON ANOTHER TAKE (CACHED / NOT KEPT)	OTHER FLAGS
clean	440	–	–	0
hallway-snr10	240	SNR target 10 dB, largest deviation 0.01 dB	120 of 120 (11 / 109)	0
lobby-snr5	440	SNR target 5 dB, largest deviation 0.01 dB	202 of 220 (44 / 158)	0
restaurant-snr0	240	SNR target 0 dB, largest deviation 0.03 dB	114 of 120 (38 / 76)	0
guest-nearby	440	bystander target -12 dB, largest deviation 0.24 dB	159 of 220 (129 / 30)	0
guest-background	240	bystander target -18 dB, largest deviation 0.13 dB	105 of 120 (105 / 0)	0
kitchen-snr5	240	SNR target 5 dB, largest deviation 0.01 dB	105 of 120 (105 / 0)	0
kitchen-snr0	440	SNR target 0 dB, largest deviation 0.56 dB	179 of 220 (126 / 53)	7
stairwell-reverb	240	same length as the clean clip	105 of 120 (105 / 0)	0
cut-start	240	cut amount exact	105 of 120 (105 / 0)	0
cut-end	240	cut amount exact	105 of 120 (105 / 0)	0
guest-nearby+lobby	240	two sources added, no level recoverable; take present at unit gain	105 of 120 (105 / 0)	0
Track B, clean	462	–	–	0
Track B, lobby-snr5	462	SNR target 5 dB, largest deviation 0.01 dB	–	0
Track B, restaurant-snr0	462	SNR target 0 dB, largest deviation 0.12 dB	–	2
CLEAN SYNTHETIC CLIPS, ONE GPT-4O-MINI-TRANSCRIBE PASS EACH: 440 CLIPS, 54.3% WITH EVERY FACT RECOVERED, MEDIAN WER 0.062; NATIVE 69.0%, ACCENTED 42.1%				

4 Systems

Table 5 lists the systems and their settings. The five standalone recognizers are run as a careful integrator would configure them: language pinned to English, everything else at the vendor default, and the vocabulary passed through the vendor's own channel in the vocabulary arm.

Table 5. Systems and settings as run for the reported result (§5). Each standalone recognizer is configured as a careful integrator would: language pinned to English, defaults otherwise. The vocabulary arm passes the list through the vendor's own biasing field, cut to what that field accepts (Appendix C) [17, 18, 19].

SYSTEM	VENDOR	SETTINGS	VOCABULARY CHANNEL	CONFIDENCE SIGNAL
Scribe v2	ElevenLabs	model scribe_v2, language_code en, audio-event tags off	key terms: all 122 (documented maximum 1,000 [17])	word log-probabilities
MAI-Transcribe 1.5	Microsoft	fast transcription API 2025-10-15, enhancedMode, verbatim	phraseList: first 200 (all 122)	none usable (reports 0)
MAI-Transcribe 2	Microsoft	as above, locale en	phraseList: first 50 (the service rejects longer lists)	none usable (reports 0)
gpt-4o-transcribe	OpenAI	language en	prompt: first 120 (the benchmark's cap; no documented term limit [18])	token log-probabilities
whisper-1	OpenAI	language en	prompt: the list trimmed to about 180 tokens, chosen on the tuning fold (§5)	segment avg_logprob
Harlow	Harlow	production build, evaluated as a black box; language English	the same list	none exposed; it decides itself when to ask

Harlow is the authors' production system and is evaluated as a black box. It receives the same audio file as every other system and, in the vocabulary arm, the same vocabulary, and for each clip it returns either a transcript or a request to repeat. It is run from a production build and is not re-implemented for the benchmark. Its design is proprietary and outside the scope of this report; it may build on third-party recognition services, including services evaluated here, which matters for two of the threats to validity in §8. Unlike the standalone recognizers it exposes no confidence to gate on: when to ask is its own decision, so it appears in every comparison at a single operating point.

5 Experimental setup

All runs were made on 30 September and 1 October 2026 against the vendors' hosted services as served then, and Harlow was run from its production build of 1 October. Responses are cached per clip, so an arm's outputs can span the sessions of those two days, and an arm was requested again in full only after a change to how it is run; the reported outputs are the last response kept for each clip, and no kept response is an error. Latency is wall-clock time at the client for the transcription call. For Harlow it covers the first stage of its processing only; later stages, which run on some clips, are not timed, so its latency is a lower bound. whisper-1 accepts a prompt of at most 224 tokens and keeps the end of a longer one [18], so the full 122-term list, which puts the property's own terms first, dropped those terms and scored worse than no prompt; it therefore receives the same list trimmed to about 180 tokens, the best of three settings overall on a 600-clip sample of the tuning fold: property terms only, trimmed list and no prompt gave 35.0%, 31.5% and 49.0% silent-wrong on 400 hotel clips and 24.5%, 30.0% and 40.0% on 200 real-speaker clips, where the shorter prompt was the better one. The standalone gpt-4o-transcribe arm runs at temperature 0.

Harlow's output is not deterministic. Both of its arms were run twice on this version, and every table reports the second run. Between the two, 15 of 2,920 hotel clips and 3 of 1,386 real-speaker clips changed outcome with the vocabulary, and 69 and 11 without it; no silent-wrong or repeat rate moved by more than 0.4 points. The system may have been changed between the two runs, so they should not be read as pure repetitions; what they show is well inside the bootstrap intervals over clips (§3.4), which do not include it. In the Harlow arm without the vocabulary an internal step failed on 59 clips, 58 of them Track B restaurant clips, and Harlow asked for a repeat on each; that is 4.2 of that arm's 28.2 points of Track B repeat rate, and the gated baselines without vocabulary are matched to that rate. The vocabulary arm had no such case.

The gated figures in every table and the risk–coverage curves use the held-out thresholds of §3.5. The result sets this paper is built from are generated with the scorer's own functions, and the recomputation of §3.6 confirms the

scorer's stored result set cell by cell. Each bootstrap interval draws its own fixed stream.

6 Results

6.1 Headline comparison at matched repeat rates

Table 6. Silent-wrong rate (%) at matched repeat rates. Gated systems send back their least-confident clips under a threshold chosen, with the aggregation, on one half of the sentences or speakers to match Harlow's repeat rate and applied to the other (§3.5); the repeat columns show the rate reached. Brackets are marginal 95% bootstrap intervals over clips (§3.4). Every system is scored on its own output (§3.5).

SYSTEM	TRACK A: SILENT-WRONG	REPEAT	TRACK B: SILENT-WRONG	REPEAT
WITH THE PROPERTY VOCABULARY				
Harlow (asks for repeats itself)	2.6 [2.1, 3.2]	5.3	3.4 [2.5, 4.4]	23.8
Scribe v2 + confidence gate	5.1 [4.3, 6.0]	5.5	4.5 [3.4, 5.6]	24.1
gpt-4o-transcribe + confidence gate	8.2 [7.2, 9.2]	5.3	10.3 [8.7, 11.8]	23.9
whisper-1 + confidence gate	14.1 [12.9, 15.4]	5.8	11.0 [9.3, 12.6]	23.9
MAI-Transcribe 2 (no usable confidence)	17.9 [16.5, 19.3]	0	16.6 [14.6, 18.6]	0
MAI-Transcribe 1.5 (no usable confidence)	28.0 [26.3, 29.7]	0	16.6 [14.7, 18.5]	0
WITHOUT ANY VOCABULARY				
Harlow (asks for repeats itself)	18.7 [17.4, 20.1]	15.7	5.9 [4.7, 7.1]	28.2
Scribe v2 + confidence gate	24.7 [23.2, 26.2]	16.0	8.7 [7.4, 10.2]	28.6
gpt-4o-transcribe + confidence gate	26.1 [24.6, 27.7]	15.3	12.6 [11.0, 14.5]	28.9
whisper-1 + confidence gate	33.1 [31.4, 34.7]	15.7	14.4 [12.5, 16.2]	27.3
MAI-Transcribe 2 (no usable confidence)	38.3 [36.5, 40.1]	0	24.5 [22.2, 26.8]	0
MAI-Transcribe 1.5 (no usable confidence)	44.1 [42.3, 45.9]	0	24.9 [22.7, 27.2]	0

Table 6 is the central result. With the vocabulary, Harlow proceeds silently wrong on 2.6% of hotel requests at a 5.3% repeat rate; the best gated single recognizer, Scribe v2, is at 5.1% at a 5.5% repeat rate, and gpt-4o-transcribe and whisper-1 at 8.2% and 14.1%. Compared clip by clip, gated Scribe v2 is silently wrong on more clips than Harlow by 2.5 points, 95% interval 1.6 to 3.5; the marginal intervals of the table do not overlap. In relative terms Harlow files 49% fewer silently wrong requests at a matched repeat rate. Conditioned on proceeding, it errs on 2.7% of the requests it files against 5.4% for gated Scribe v2. On real accented speakers Harlow is at 3.4% against 4.5% for gated Scribe v2, a paired difference of 1.1 points, 95% interval 0.1 to 2.1, with speakers resampled together; the ordering is the same for every gated recognizer. Without any vocabulary Harlow is at 18.7% against 24.7% at 15.7% and 16.0% repeats (6.0 points, 95% interval 3.9 to 8.1): Harlow, like every recognizer, is worse without the vocabulary, 7.2 times as often silently wrong. Microsoft, the one vendor whose services expose no usable confidence, cannot trade coverage for safety at all: what its two recognizers file, they file silently, at 17.9% (MAI-Transcribe 2) and 28.0% (MAI-Transcribe 1.5) on hotel requests.

6.2 All systems, with and without the vocabulary

Table 7. Track A, every system with the property vocabulary and no gate; Harlow at its own repeat rate. Correct, silent-wrong and repeat sum to 100. Fact recall and WER are over proceeded clips only, so Harlow's are over a smaller, easier set (\$6.2).

SYSTEM	VENDOR	CORRECT	SILENT-WRONG [95% CI]	REPEAT	FACT RECALL	WER	P50 (MS)	P95 (MS)
Harlow	Harlow	92.0	2.6 [2.1, 3.2]	5.3	98.8	0.080	855	1210
Scribe v2	ElevenLabs	93.2	6.9 [6.0, 7.8]	0.0	97.0	0.152	948	1733
gpt-4o-transcribe	OpenAI	89.2	10.8 [9.7, 11.9]	0.0	94.9	0.092	926	1677
whisper-1	OpenAI	84.2	15.8 [14.5, 17.0]	0.0	93.5	0.090	1467	2762
MAI-Transcribe 2	Microsoft	82.1	17.9 [16.5, 19.3]	0.0	92.8	0.071	331	1502
MAI-Transcribe 1.5	Microsoft	72.0	28.0 [26.3, 29.7]	0.0	88.8	0.146	964	1184

Table 8. Track B, every system with the property vocabulary and no gate. Fact recall and WER are over proceeded clips only (\$6.2).

SYSTEM	VENDOR	CORRECT	SILENT-WRONG [95% CI]	REPEAT	FACT RECALL	WER	P50 (MS)	P95 (MS)
Harlow	Harlow	72.8	3.4 [2.5, 4.4]	23.8	97.2	0.041	860	1250
Scribe v2	ElevenLabs	80.2	19.8 [17.9, 22.0]	0.0	86.1	0.132	831	1366
gpt-4o-transcribe	OpenAI	68.8	31.2 [28.8, 33.8]	0.0	77.7	0.208	754	1276
whisper-1	OpenAI	72.1	27.9 [25.5, 30.2]	0.0	80.8	0.173	1409	2812
MAI-Transcribe 2	Microsoft	83.4	16.6 [14.6, 18.6]	0.0	88.6	0.094	320	1492
MAI-Transcribe 1.5	Microsoft	83.4	16.6 [14.7, 18.5]	0.0	89.1	0.097	938	1148

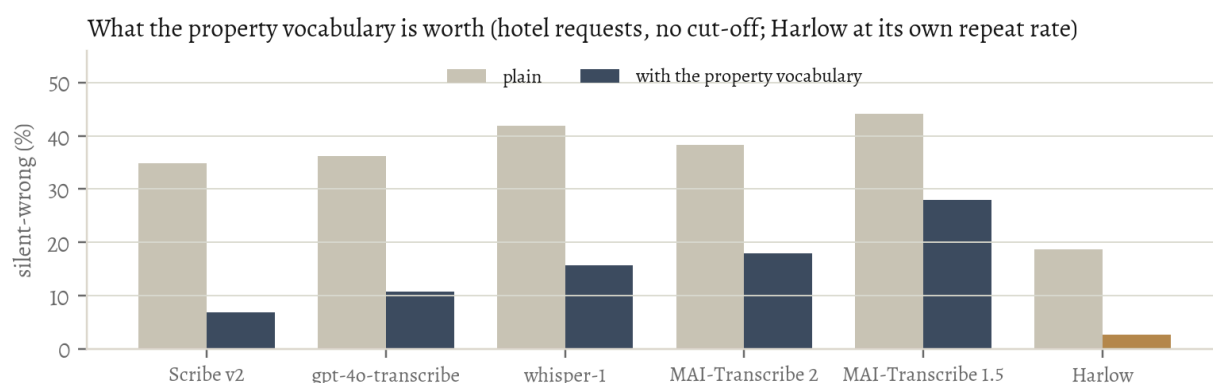


Figure 2. The property vocabulary is the largest single lever for every recognizer. On hotel requests Scribe v2 moves from 34.9% to 6.9%, gpt-4o-transcribe from 36.2% to 10.8%, whisper-1 from 41.9% to 15.8%.

Three observations follow from Tables 7 and 8 and Figure 2. First, on hotel requests the vocabulary matters more than the choice of recognizer: every recognizer does better with it than the best recognizer does without it (MAI-Transcribe 1.5 with the vocabulary at 28.0% against Scribe v2 without it at 34.9%). Second, on real speakers the vocabulary helps every system, by up to 10.8 points, but the choice of recognizer matters at least as much there: MAI-Transcribe 2 without the vocabulary (24.5%) beats gpt-4o-transcribe and whisper-1 with it (31.2% and 27.9%). Third, WER and silent-wrong rate rank the systems differently. On hotel requests MAI-Transcribe 2 has the lowest

WER of any single recognizer (0.071) and the lowest median latency (331 ms), yet files 17.9% of requests silently wrong against Scribe v2's 6.9% at a WER of 0.152, 2.1 times as high; gpt-4o-transcribe and whisper-1 likewise have a lower WER than Scribe v2 and file more requests silently wrong. Scribe v2's word errors evidently fall outside the facts. Fact recall and WER are computed over the clips a system proceeded with: for the ungated systems that is every clip, for Harlow it is the 94.7% of Track A clips and 76.2% of Track B clips it did not send back, which excludes its hardest clips. Harlow's values (98.8% and 97.2% recall; 0.080 and 0.041 WER) are therefore not comparable with the other rows of those columns, and we draw no ranking from them.

6.3 Acoustic conditions

Figure 3 and Table 9 give the silent-wrong rate by condition for the 120 request lines.

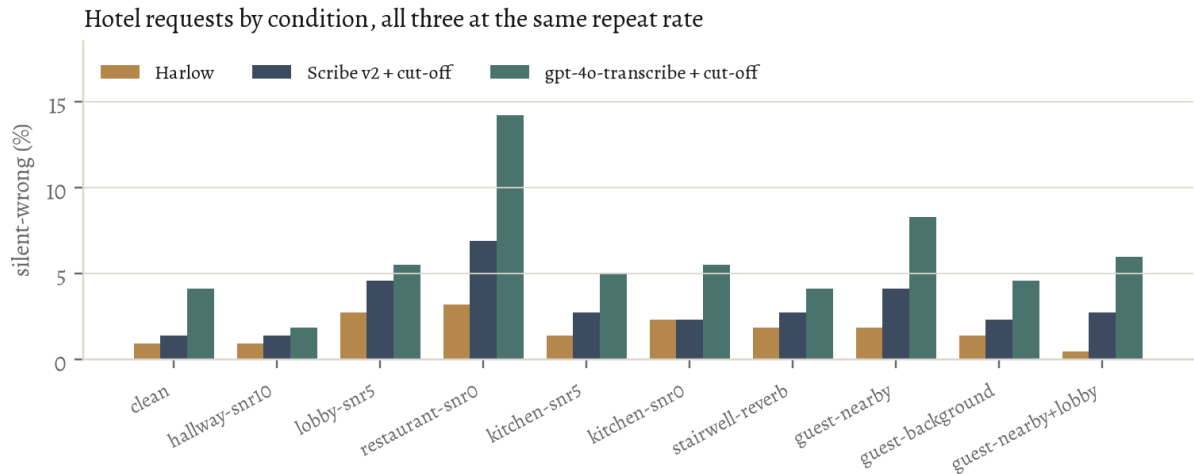


Figure 3. Track A silent-wrong rate by condition for the three systems at the same overall repeat rate.

Table 9. Track A by acoustic condition: silent-wrong % (gold: repeat %). The 120 request lines, every kept voice, every main condition, 218 clips per condition; the gated systems use the overall matched-repeat threshold, so their per-condition repeat rates vary.

CONDITION	HARLOW	SCRIBE V2 + GATE	GPT-4O + GATE	SCRIBE V2, UNGATED	MAI-2, UNGATED
clean	0.9 · 1.8	1.4 · 0.5	4.1 · 1.4	1.8	9.6
hallway-snr10	0.9 · 0.9	1.4 · 1.4	1.8 · 0.5	1.4	10.5
lobby-snr5	2.8 · 2.3	4.6 · 1.8	5.5 · 0.9	5.1	9.6
restaurant-snr0	3.2 · 14.7	6.9 · 11.0	14.2 · 10.1	13.8	22.9
kitchen-snr5	1.4 · 1.8	2.8 · 2.3	5.1 · 0.9	3.2	8.7
kitchen-snr0	2.3 · 1.4	2.3 · 1.4	5.5 · 0.9	2.3	9.2
stairwell-reverb	1.8 · 1.8	2.8 · 2.3	4.1 · 2.3	3.7	11.5
guest-nearby	1.8 · 7.3	4.1 · 2.3	8.3 · 2.8	4.6	23.4
guest-background	1.4 · 3.2	2.3 · 4.1	4.6 · 0.5	2.8	15.6
guest-nearby+lobby	0.5 · 4.6	2.8 · 3.2	6.0 · 1.8	4.1	11.5

In quiet and in mild noise Harlow and gated Scribe v2 are within about half a point of each other (0.9% and 1.4% on clean audio, 0.9% and 1.4% in hallway noise 10 dB quieter than the speaker); gated gpt-4o-transcribe is at 4.1% on clean audio, and ungated MAI-Transcribe 2, which has no confidence to gate on, is already at 9.6% before any noise is added. Harlow separates from single recognizers where they fail silently: in restaurant noise as loud as the speaker (3.2% against 6.9% for gated Scribe v2, at 14.7% and 11.0% repeats), with a guest talking nearby (1.8% against

4.1%, at 7.3% and 2.3% repeats), and with the guest and the lobby together (0.5% against 2.8%, at 4.6% and 3.2%). Harlow buys part of the guest-nearby margin with a higher repeat rate; conditioned on proceeding it still errs less (2.0% against 4.2%). Half the guest lines name a different room. The scorer asks only that the wearer's room be present, so a transcript that carries both rooms counts as correct, and these cells understate what a guest's voice costs a system that acts on the first room it finds (§8). Ungated MAI-Transcribe 2 is at 23.4% with a guest nearby.

6.4 Speech cut short: a stress test

Two further conditions remove the first 200 ms or the last 300 ms of speech, measured from the detected speech edges. They model a capture window that opens late or closes early: a push-to-talk radio released before the last word, or voice activation that triggers after the first. The badge Harlow ships on does neither: it starts capturing when the wearer presses it and stops when the wearer presses again after speaking. They are therefore reported separately from the main set, as a stress test (Table 10), together with the overall figures when they are included.

Table 10. Stress test, Track A request lines with speech cut short: silent-wrong % (gold: repeat %), 218 clips per condition, each system scored on its own output. Gates tuned on the full twelve-condition set.

CONDITION	HARLOW	SCRIBE V2 + GATE	GPT-4O + GATE	SCRIBE V2, UNGATED	MAI-2, UNGATED
cut-start	1.8 · 7.8	3.2 · 7.3	9.2 · 2.3	4.1	17.4
cut-end	18.4 · 12.8	23.8 · 10.5	35.8 · 4.6	29.8	37.2

With the last 300 ms removed every system fails most often, and Harlow remains ahead: 18.4% against 23.8% for gated Scribe v2 and 35.8% for gated gpt-4o-transcribe. No system detects the truncation reliably: each often proceeds with a reading that lacks the final number, and Harlow asks for a repeat on 12.8% of these clips. A capture window that closes after the speaker stops removes the condition at its source. Truncation at the start costs Harlow and gated Scribe v2 little (1.8% and 3.2%) but takes gated gpt-4o-transcribe to 9.2% and ungated MAI-Transcribe 2 to 17.4%. With both conditions included in the set, Harlow is at 3.6% against 6.2% for gated Scribe v2 at 6.0% and 6.0% repeats (2.6 points, 95% interval 1.7 to 3.6).

6.5 Line types

Figure 4 and Table 11 break Track A down by line type, pooled over four conditions.

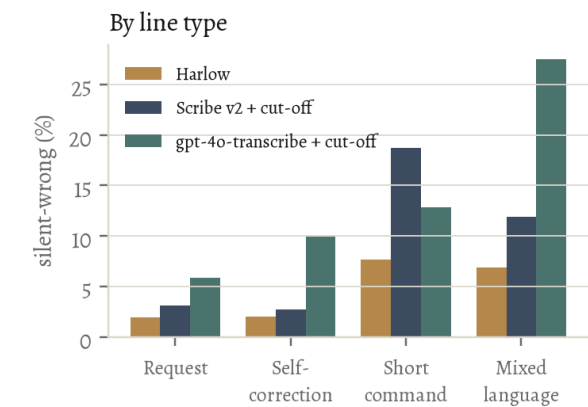


Figure 4. Track A silent-wrong rate by line type, pooled over four conditions.

Table 11. Track A by line type, pooled over clean, lobby-snr5, kitchen-snr0 and guest-nearby: silent-wrong % (gold: repeat %).

LINE TYPE	HARLOW	SCRIBE V2 + GATE	GPT-4O + GATE
Request (n = 872)	1.9 · 3.2	3.1 · 1.5	5.9 · 1.5
Self-correction (n = 292)	2.1 · 6.5	2.7 · 4.1	9.9 · 3.4
Short command (n = 288)	7.6 · 9.4	18.8 · 19.1	12.8 · 27.1
Mixed language (n = 160)	6.9 · 14.4	11.9 · 17.5	27.5 · 12.5

The line types differ in what a gate costs. On short commands Harlow is at 7.6% against 18.8% for gated Scribe v2 and 12.8% for gated gpt-4o-transcribe while sending back fewer of them (9.4% against 19.1% and 27.1%): a two-word utterance gives a confidence score little to work with, so a gate tuned to the overall repeat budget spends a disproportionate share of it here. Mixed-language requests show the widest gap (6.9% against 11.9% and 27.5%),

consistent with the known difficulty of code-switched speech for monolingual decoding [15]. Self-corrections are handled well by Scribe v2 and Harlow and poorly by gpt-4o-transcribe. Short commands are also where the room step of \$3.5 matters most: scored without it, gated Scribe v2 is at 18.8% on them (\$6.8), because a bare room number is what a recognizer most readily writes as a time.

6.6 Real accented speakers

Tables 12 and 13 and Figure 5 give Track B by condition and by the speaker's first language.

Table 12. Track B by condition: silent-wrong % (gold: repeat %).

CONDITION	HARLOW	SCRIBE V2 + GATE	GPT-4O + GATE
clean	3.5 · 1.7	3.0 · 4.8	5.2 · 2.8
lobby-snr5	3.7 · 12.1	4.8 · 13.4	11.5 · 13.4
restaurant-snr0	3.0 · 57.6	5.6 · 54.1	14.3 · 55.6

Table 13. Track B by the speaker's first language: speakers kept after the segmentation screen, then silent-wrong % (gold: repeat %).

FIRST LANGUAGE	SPEAKERS	HARLOW	SCRIBE V2 + GATE
Amharic	6	4.6 · 28.7	7.4 · 30.6
Arabic	8	8.3 · 36.8	6.9 · 35.4
English	10	2.2 · 15.0	2.2 · 13.3
French	8	3.5 · 22.9	2.1 · 29.9
Hindi	8	0.7 · 22.9	2.1 · 22.2
Mandarin	7	1.6 · 19.1	5.6 · 17.5
Portuguese	7	4.0 · 24.6	4.8 · 24.6
Russian	6	3.7 · 19.4	7.4 · 18.5
Spanish	5	2.2 · 33.3	4.4 · 30.0
Tagalog	7	1.6 · 15.9	3.2 · 19.1
Vietnamese	5	5.6 · 30.0	5.6 · 30.0

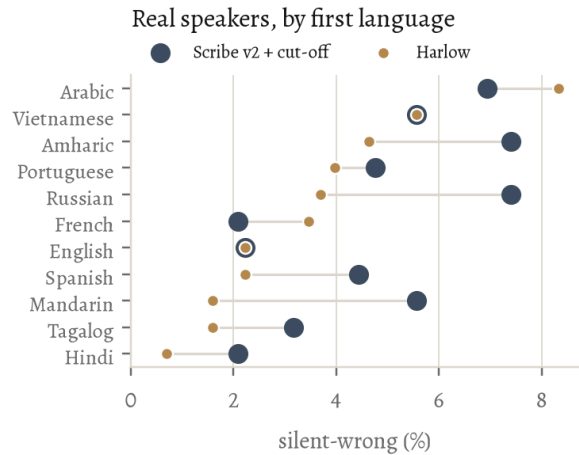


Figure 5. Track B silent-wrong rate by first language, Harlow against gated Scribe v2 at matched repeats. With five to ten speakers per language each point carries a wide interval.

On real speakers Harlow is ahead overall but not uniformly. Its point estimate is lower than gated Scribe v2's for Mandarin, Russian, Amharic, Spanish, Tagalog and Hindi first languages (by 4.0, 3.7, 2.8, 2.2, 1.6, 1.4 points respectively), within a point for English, Portuguese and Vietnamese, and higher for French and Arabic (by 1.4, 1.4 points). With five to ten speakers per language and marginal intervals, none of these per-language differences can be called significant, and Track B's read material inflates every absolute rate (\$3.3). What the track shows is narrower: the gated ordering of Track A survives on real speech. Harlow is ahead of gated gpt-4o-transcribe and whisper-1 (3.4% against 10.3% and 11.0%) with non-overlapping intervals, and ahead of gated Scribe v2 by 1.1 points, with overlapping marginal intervals and a paired interval that reaches down to 0.1 points (\$6.1). Because the speaker selection favours one of the systems under test and the vocabulary contains the facts (\$8), Track B is consistent with the Track A result rather than an independent confirmation of it.

6.7 Latency

Table 14. Transcription latency, wall clock at the client, ms. For Harlow this covers the first stage of its processing only and is a lower bound (§5).

SYSTEM	TRACK A P50	P95	TRACK B P50	P95
Harlow	855	1210	860	1250
Scribe v2	948	1733	831	1366
gpt-4o-transcribe	926	1677	754	1276
whisper-1	1467	2762	1409	2812
MAI-Transcribe 2	331	1502	320	1492
MAI-Transcribe 1.5	964	1184	938	1148

Table 14 gives transcription latency measured at the client. Harlow's figure covers the first stage of its processing only (§5): 855 ms median and 1210 ms at the 95th percentile on Track A, in the same range as the standalone recognizers. Later stages add time on the clips that need them and are not timed, so Harlow's figures are lower bounds and should not be read as a cost or a saving of the system. The arms were run at different times, so differences of this size are within the services' own variation. MAI-Transcribe 2 is the fastest single recognizer by a wide margin (331 ms median); whisper-1 the slowest (1467 ms).

6.8 Effect of post-processing

Table 15. The headline comparison on hotel requests under four treatments of post-processing: silent-wrong %, Harlow against gated Scribe v2 at Harlow's repeat rate. Brackets are marginal 95% intervals; the paired difference (gated Scribe v2 minus Harlow) resamples sentences together (§3.4).

POST-PROCESSING	HARLOW	GATED SCRIBE V2	PAIRED DIFFERENCE, POINTS	HARLOW, NO VOCABULARY	GATED SCRIBE V2, NO VOCABULARY
MAIN SET: TEN CONDITIONS, 2,920 CLIPS					
Each system's own output: no room step, no name correction (this paper)	2.6 [2.1, 3.2]	5.1 [4.3, 6.0]	2.5 [1.6, 3.5]	18.7	24.7
Room step for every system, name correction for none	2.6 [2.1, 3.2]	4.1 [3.4, 4.8]	1.5 [0.8, 2.2]	18.7	21.3
Room step for every system, name correction for Harlow only	2.3 [1.8, 2.8]	4.1 [3.4, 4.8]	—	11.3	21.3
As the scorer writes it: both steps reach Harlow's transcripts only	2.3 [1.8, 2.8]	5.1 [4.3, 6.0]	2.8 [1.8, 3.8]	11.3	24.7
WITH THE TWO CUT CONDITIONS OF §6.4: 3,356 CLIPS					
Each system's own output: no room step, no name correction (this paper)	3.6 [3.0, 4.2]	6.2 [5.4, 7.0]	2.6 [1.7, 3.6]	19.7	26.0
Room step for every system, name correction for none	3.6 [3.0, 4.2]	5.1 [4.4, 5.8]	1.6 [0.9, 2.2]	19.7	22.9
Room step for every system, name correction for Harlow only	3.3 [2.8, 3.9]	5.1 [4.4, 5.8]	—	12.6	22.9

POST-PROCESSING	HARLOW	GATED SCRIBE V2	PAIRED DIFFERENCE, POINTS	HARLOW, NO VOCABULARY	GATED SCRIBE V2, NO VOCABULARY
As the scorer writes it: both steps reach Harlow's transcripts only	3.3 [2.8, 3.9]	6.2 [5.4, 7.0]	2.9 [2.0, 3.9]	12.6	26.0

Table 16. What the room step is worth to each recognizer on hotel requests: silent-wrong % without a gate, scored as transcribed and after the step, and the number of the 2,920 clips it turns correct.

RECOGNIZER	WITH VOCABULARY, AS TRANSCRIBED	AFTER THE STEP	CLIPS	NO VOCABULARY, AS TRANSCRIBED	AFTER THE STEP	CLIPS
Scribe v2	6.9	5.5	41	34.9	31.2	107
gpt-4o-transcribe	10.8	10.7	1	36.2	34.3	55
whisper-1	15.8	15.8	0	41.9	41.9	0
MAI-Transcribe 2	17.9	7.7	297	38.3	33.2	150
MAI-Transcribe 1.5	28.0	8.1	581	44.1	31.9	358

Every figure above scores each system on its own output (§3.5). Table 15 gives the headline comparison under three other treatments. Giving every recognizer the room step, which Harlow's output already follows, lowers gated Scribe v2 from 5.1% to 4.1% and leaves Harlow at 2.6%: the paired difference narrows from 2.5 to 1.5 points (95% interval 0.8 to 2.2) and stays above zero, so the result does not rest on the recognizers' number formatting. As the scorer itself writes it, with the room step and the emulated name correction reaching Harlow's transcripts only, the comparison reads 2.3% against 5.1%; the name correction is a step of the production system that the recorded output does not contain, and is not counted here. With the two cut conditions of §6.4 included, the comparison reads 3.6% against 6.2% on each system's own output and 3.6% against 5.1% with the room step for every recognizer. Without the vocabulary the steps matter more: 18.7% against 24.7% here, 18.7% against 21.3% with the room step for every recognizer, and 11.3% against 24.7% as scored, most of the last difference the name correction, which turned 216 of Harlow's clips correct in that arm. Track B is unaffected by either step.

Table 16 shows what the room step is worth to each recognizer. The two Microsoft recognizers often write a room as a clock time, and the step takes MAI-Transcribe 1.5 from 28.0% to 8.1% silent-wrong with the vocabulary and MAI-Transcribe 2 from 17.9% to 7.7%; Scribe v2 gains 1.4 points, gpt-4o-transcribe and whisper-1 nothing. The step covers rooms written with a colon or a slash; whisper-1 writes them with a dot, which the scorer also reads as a time. A step extended to that shape would move whisper-1 from 15.8% to 13.7%, no other system by more than 0.2 points, and the headline comparison not at all. The practical reading is the same as for the vocabulary: a deterministic step that uses the property's own data can be worth more than the choice of recognizer, and a comparison on each system's own output measures its formatting along with its recognition, which is why both are reported.

7 Discussion

What moves the facts. Three things an integrator controls move the silent-wrong rate on this benchmark. The property's vocabulary is worth more than the choice of recognizer for every recognizer (§6.2). Normalising how numbers are written is worth as much for the recognizers that write rooms as times, and nothing for those that do not (§6.8). A way to decline is worth less than either, but it is the only one of the three that turns a silent mistake into a question (§6.1), and it depends on what the vendor exposes: one of the three vendors evaluated returns no usable

confidence, and for the others a gate tuned to an overall budget spends its repeats unevenly, heavily on short utterances and lightly on long ones (Table 11).

Where Harlow asks. Harlow's repeats are concentrated where the audio is hardest. On hotel requests it asks on 1.8% of clean clips, on 7.3% with a guest talking nearby and on 14.7% in restaurant noise as loud as the speaker (0 dB); on real speakers on 1.7% of clean segments and 57.6% under restaurant noise (Tables 9 and 12). The gated recognizers send back most where the audio is worst too, but on hotel requests, at a matched overall rate, gated Scribe v2 is silently wrong at least as often as Harlow in every condition (Table 9). Whether a repeat costs the wearer two seconds or the request depends on the deployment; the benchmark measures how often it is asked, not what it costs.

What a black box allows. Because Harlow is evaluated only through what it returns, the benchmark says how often it is silently wrong and how often it asks, and nothing about why. The same holds for the standalone services, whose models are equally undisclosed. The comparison is therefore between products as served, on identical audio, with identical vocabulary, each scored on what it returns, at matched repeat rates; it is not a comparison of methods.

8 Limitations and threats to validity

- **Synthetic accents.** Track A's accented voices are synthesis vendors' voices reading English; they are consistent and controllable but are not people. Track B is real speech but read, not requests, and its speakers' misreadings count against every system.
- **Track B selection and vocabulary.** The 77 Track B speakers are those for whom ElevenLabs Scribe v2, a system under test, recovered every boundary word in one pass; the twelve it did not are excluded, which removes the speech that recognizer found hardest. The Track B vocabulary consists of the paragraph's own names and items, so the vocabulary arm is told the facts it must recover. Both apply to every system equally, but the first favours Scribe v2 and any system that builds on it, and both make Track B's absolute rates optimistic. On 7 clean segments no system is right, because the speaker departed from the printed text or, in one case (B-french3-s2), the boundary cuts into the first word.
- **Black-box evaluation.** Harlow's design is not disclosed. A reader cannot attribute its results to a method, and cannot rule out that it benefits from services it is compared with, so a selection effect that favours one of them may favour Harlow as well. Its verdict for every clip is released; its output and the system are not, and its arms cannot be re-run by others.
- **Post-processing.** Each system is scored on what it returns, so a recognizer that writes rooms as clock times is charged for it, although an integrator could add a room step; with that step given to every recognizer the hotel-request comparison reads 2.6% against 4.1% (\$6.8). That step is our implementation of the rule Harlow's output follows, checked against Harlow's outputs (\$3.5), and the room list it consults is the set of rooms the corpus names rather than a property's full list. Harlow's recorded output precedes its own name correction, which is therefore not counted; a deployment that corrects names against its roster, Harlow's or another, would do better than these figures on name facts. As the scorer writes it, with both steps for Harlow only, the comparison reads 2.3% against 5.1% (\$6.8).
- **Scorer rules.** Two rules of the scorer are lenient in ways that matter for particular cells. Negation is satisfied by any negating word, and 10 of the 20 negated lines contain the phrase "do not disturb", so on those lines a transcript that drops the leading "Don't" still passes; cut-start, which can remove that word, may be understated for every system. And a transcript that contains the wearer's room passes even if it also contains the bystander's, so the guest conditions understate the cost of transcribing the guest. Both rules apply to every system; neither is corrected here.
- **Statistics.** Table intervals are marginal and treat clips as independent although clips share lines and speakers; paired, clustered differences are reported for the headline comparisons only (\$3.4, \$6.1). Each standalone arm's reported output is a single response per clip; Harlow's arms were run twice (\$5).

- **No device path.** Every clip is clean 16 kHz audio with noise mixed in. No microphone, codec, automatic gain control or room from an actual wearable device is in the chain. A device track recorded through the Harlow badge is planned.
- **Other capture devices.** The main set leaves out speech cut short because the badge's press-to-start, press-to-stop capture does not produce it. A push-to-talk or voice-activated client could cut speech; for such a deployment the stress test of §6.4 and the figures with both conditions included (§6.4, §6.8) are the relevant ones.
- **Fixed SNR.** Noise is mixed at a constant level per clip; real environments vary within an utterance.
- **Stairwell rendering.** The reverberated clips are coarsely quantised. The renderer scales the convolved signal up after it has come back as 16-bit samples, so a stairwell clip holds a median of 663 distinct sample values in steps of about 43, where a clean clip holds 13,240 in steps of 1. The quantisation noise is 47 dB below the signal at the median clip and 32 dB at the worst, for every system alike; the condition is a reverberation with that noise added and would render differently with another build of the audio tool.
- **Non-determinism.** Harlow's output is not deterministic; its two runs on this version moved no total by more than 0.4 points (§5), less than the bootstrap intervals, which are over clips only. The vendors' recognizers are not all repeatable either: replaying 200 clips of the reported run changed the words of Scribe v2's transcript on 28 and its score bucket on 5, while whisper-1 returned its stored output on 199 (§3.6). Each standalone arm's reported output is one response per clip, so its totals carry that noise.
- **Latency scope.** Harlow's latency covers the first stage of its processing only and is a lower bound (§5).
- **Take pairing.** For OpenAI voices the clips of one line are not all built on one synthesis (§3.6): 1,404 of 1,620 mixed OpenAI-voice clips use another take than their clean clip, and for 426 of them that take was not kept. Per-condition differences for those voices therefore include take-to-take variance, the single transcription pass of §3.6 certifies the clean clip's take only, and a re-render, with either renderer, will not reproduce those clips; the clips themselves are the reference.
- **Moving targets.** Every number is tied to the vendors' services as served on 30 September and 1 October 2026. The stored responses are kept with the benchmark; a re-run on a later date is a new result.
- **One property.** Track A is one fictional property with one vocabulary. Properties with larger rosters or more homophonous room numbers may behave differently.

9 Conclusion

We have defined a benchmark for speech-to-request in hotel operations around the quantity a hotel actually pays for, the silently wrong request, and a protocol that makes systems comparable at equal repeat budgets with the same vocabulary, each system scored on its own output. On it, five commercial recognizers show that on hotel requests the property vocabulary is worth more than the choice of recognizer, that the way a recognizer writes numbers can account for a large share of its errors until one deterministic step is applied, that WER does not predict fact-level reliability, and that only some vendors expose confidence a deployment can act on. Harlow, the authors' production system, files 49% fewer silently wrong hotel requests than the best gated single recognizer at a matched repeat rate, ahead still when every recognizer is given a room step that Harlow applies itself, and is ahead by a smaller margin on real accented speech, with the largest gains in heavy noise, on mixed-language requests and with a guest talking nearby. Under a stress test that cuts the end of speech, which the badge's capture does not produce, every system fails most often and Harlow remains ahead.

The benchmark is released with each system's verdict for every clip and with a script that recomputes the silent-wrong, repeat and correct rates of this report's tables from those verdicts. Of the audio, the Track B clips recorded without added noise are released under the Speech Accent Archive's licence (CC BY-NC-SA 4.0), and the synthetic speech, before any noise is mixed in, is available to researchers on request under a data-use agreement; the other clips and the standalone recognizers' transcripts are not redistributed. The verdicts were checked against the stored

outputs (§3.6). A submission reports the three buckets with intervals, silent-wrong at matched repeat rates using held-out thresholds where the system exposes a confidence, and results by condition, line type and first language, with the systems and settings as served and the date. This is bench version 4.1; any change to the corpus, conditions, scorer or protocol increments the version, and results across versions are never mixed. Planned for the next version: Harlow's output recorded after its own name correction, and name correction given to every system; stricter negation and bystander-room rules in the scorer; a device track; and additional languages on both tracks.

Reproducibility. The noise is DEMAND, Zenodo record 1227121 (CC BY 4.0) [6]; the real speech is the Speech Accent Archive, obtained from the Hugging Face dataset [changelinglab/speechaccentarchive-pr](https://huggingface.co/datasets/changelinglab/speechaccentarchive-pr), and is under the archive's Creative Commons Attribution-NonCommercial-ShareAlike licence, which governs any redistribution of the recordings and of the Track B clips cut from them; the clips released with this report are under its version 4.0 [7]. The corpus plan is frozen from seed 20260930. The scorer scores every clip; the result sets, the audio measurements and the checks that tie the result set to the outputs, the outputs to the clips and the clips to the generator are produced by the benchmark's tools, and every figure, table and number in this document is generated from them.

Disclosure. See the statement on the first page. Scribe is a product of ElevenLabs; MAI-Transcribe of Microsoft; gpt-4o-transcribe, gpt-4o-mini-transcribe and whisper-1 of OpenAI. This document reports measurements of those services as served on the stated dates and makes no claim about them beyond those measurements.

References

1. A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. In *Proc. ICML*, 2023.
2. V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. LibriSpeech: an ASR corpus based on public domain audio books. In *Proc. ICASSP*, 2015.
3. G. Chen, S. Chai, G. Wang, et al. GigaSpeech: an evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio. In *Proc. Interspeech*, 2021.
4. J. Barker, S. Watanabe, E. Vincent, and J. Trmal. The fifth 'CHiME' speech separation and recognition challenge: dataset, task and baselines. In *Proc. Interspeech*, 2018.
5. A. C. Morris, V. Maier, and P. Green. From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. In *Proc. Interspeech*, 2004.
6. J. Thiemann, N. Ito, and E. Vincent. The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): a database of multichannel environmental noise recordings. In *Proc. Meetings on Acoustics (ICA)*, 2013. Zenodo record 1227121.
7. S. Weinberger. Speech Accent Archive. George Mason University, 2015. accent.gmu.edu. Recordings obtained from the Hugging Face dataset [changelinglab/speechaccentarchive-pr](https://huggingface.co/datasets/changelinglab/speechaccentarchive-pr) (test split), accessed September 2026.
8. J. G. Fiscus. A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER). In *Proc. ASRU*, 1997.
9. H. Jiang. Confidence measures for speech recognition: a survey. *Speech Communication*, 45(4):455–470, 2005.
10. R. El-Yaniv and Y. Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.
11. Y. Geifman and R. El-Yaniv. Selective classification for deep neural networks. In *Proc. NeurIPS*, 2017.
12. G. Pundak, T. N. Sainath, R. Prabhavalkar, A. Kannan, and D. Zhao. Deep context: end-to-end contextual speech recognition. In *Proc. IEEE SLT*, 2018.
13. M. Bisani and H. Ney. Bootstrap estimates for confidence intervals in ASR performance evaluation. In *Proc. ICASSP*, 2004.
14. B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1993.
15. S. Sitaram, K. R. Chandu, S. K. Rallabandi, and A. W. Black. A survey of code-switched speech and language processing. [arXiv:1904.00784](https://arxiv.org/abs/1904.00784), 2019.
16. L. Zheng, W.-L. Chiang, Y. Sheng, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Proc. NeurIPS Datasets and Benchmarks*, 2023.
17. ElevenLabs. Create transcript: Speech to Text API reference (model scribe_v2; key terms, at most 1,000). elevenlabs.io/docs/api-reference/speech-to-text/convert, accessed 1 October 2026.

18. OpenAI. Speech to text guide and audio/transcriptions API reference (gpt-4o-transcribe, whisper-1; prompt; logprobs). platform.openai.com/docs/guides/speech-to-text, accessed 1 October 2026.
19. Microsoft. MAI-Transcribe in the Azure AI Speech fast transcription API (speechtotext/transcriptions:transcribe, api-version 2025-10-15; enhancedMode; phraseList). learn.microsoft.com/azure/ai-services/speech-service/mai-transcribe and [.../fast-transcription-create](https://learn.microsoft.com/azure/ai-services/speech-service/mai-transcribe-api-version), accessed 1 October 2026.
20. S. Kim, A. Arora, D. Le, C.-F. Yeh, C. Fuegen, O. Kalinli, and M. L. Seltzer. Semantic distance: a new metric for ASR performance analysis towards spoken language understanding. In *Proc. Interspeech*, 2021.
21. M. Del Rio, N. Delworth, R. Westerman, et al. Earnings-21: a practical benchmark for ASR in the wild. In *Proc. Interspeech*, 2021.
22. L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and Y. Bengio. Speech model pre-training for end-to-end spoken language understanding. In *Proc. Interspeech*, 2019.
23. E. Bastianelli, A. Vanzo, P. Swietojanski, and V. Rieser. SLURP: a spoken language understanding resource package. In *Proc. EMNLP*, 2020.
24. P. Tomasello, A. Shrivastava, D. Lazar, et al. STOP: a dataset for spoken task oriented semantic parsing. In *Proc. IEEE SLT*, 2022.
25. B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proc. NeurIPS*, 2017.

Appendix A Risk–coverage trade-off

Table A1. Risk–coverage trade-off: silent-wrong % when a gated system is given the stated repeat budget, with the threshold taken from the other fold (\$3.5). A threshold chosen on one fold does not hit the budget exactly on the other; the share of clips actually sent back is in brackets, and Figure 1 plots it. Harlow has a single operating point.

SYSTEM	BUDGET 0%	BUDGET 5%	BUDGET 10%	BUDGET 20%	BUDGET 30%	BUDGET 40%	BUDGET 50%
TRACK A: HOTEL REQUESTS							
Scribe v2	6.8 (0.0)	5.2 (5.1)	4.2 (11.1)	2.9 (20.3)	1.8 (30.1)	1.3 (40.1)	0.7 (49.9)
gpt-4o-transcribe	10.7 (0.1)	8.3 (5.0)	6.6 (12.2)	3.5 (28.6)	2.4 (40.6)	1.7 (41.4)	1.1 (51.3)
whisper-1	15.7 (0.0)	14.2 (5.6)	12.7 (10.7)	10.1 (20.9)	7.8 (32.3)	6.5 (40.5)	5.2 (47.8)
Harlow	2.6 at R = 5.3						
TRACK B: REAL SPEAKERS							
Scribe v2	19.8 (0.1)	15.3 (4.8)	11.0 (10.2)	5.7 (20.1)	3.2 (30.2)	2.3 (39.5)	1.3 (49.9)
gpt-4o-transcribe	31.1 (0.1)	26.5 (5.0)	21.5 (10.0)	12.8 (20.6)	7.6 (29.1)	3.0 (40.0)	1.4 (49.7)
whisper-1	27.9 (0.0)	22.7 (5.3)	18.2 (10.3)	12.3 (20.3)	8.4 (31.0)	5.7 (40.1)	3.3 (50.1)
Harlow	3.4 at R = 23.8						

Appendix B Scorer normalisation

Table B1. Scorer normalisation (common.py). Lenient about form, strict about facts. WER is computed on the same normalised tokens.

RULE	EFFECT
Case and accents	Lower-cased, Unicode accents folded
Punctuation and hyphens	Removed; apostrophes dropped (don't → dont)
Fillers	uh, um, erm, hmm, mm, ah removed
Number words	two → 2, twelve → 12; 'oh' inside a number run → 0
Room numbers	Every digit string a run of number tokens can spell counts: 12 14, 1214, 5 1 2, 10 oh 4
Clock and date shapes	12:14, 12/14, 12.14 become one opaque token and never match a room
Thousands separators	1,214 → 1214
Word splits	Contiguous match blind to spacing: check out = checkout, foam-pillows = foam pillows
Plurals	Trailing -s ignored on tokens longer than three letters: six spoon = six spoons
Negation	A negation fact is satisfied by any token in the negation set (not, don't, no, never, cancel ...)

Appendix C The property vocabulary

The vocabulary arm supplies 122 terms on Track A: the property's ten staff first names, fifteen guest surnames and twenty property terms, then a list of 92 hospitality terms (turndown, rollaway, due out, PTAC, folio, BEO, par levels, ...) with duplicates removed. Each vendor receives the list cut to what its biasing field accepts, in that order: Scribe v2 all of it as key terms (the service accepts up to 1,000 [17]); MAI-Transcribe 1.5 the first 200 (all of it) and MAI-Transcribe 2 the first 50 as a phrase list, 50 being the longest list that service accepted [19]; gpt-4o-transcribe the first 120 in a prompt of the form “Hotel staff request. Terms: ...”, the benchmark's cap, as the prompt has no

documented term limit [18]; whisper-1 the same list trimmed to about 180 tokens (§5). Track B uses the paragraph's names and items in place of the property terms, then the same hospitality list (99 terms).

Appendix D Intelligibility screen

The screen (§3.2) removed three Azure voices: pt-BR-FranciscaNeural, ru-RU-DmitryNeural, vi-VN-NamMinhNeural. For each, the median standalone recognizer without vocabulary recovered at most 20% of clean clips with every fact, which indicates failed synthesis of English by a cross-lingual voice rather than an accent. The threshold is a constant in the scorer. Unscreened totals are in the scorer's generated report.

Appendix E Synthetic takes by voice

One gpt-4o-mini-transcribe pass per clean clip, no vocabulary, on the clips of the reported run. Facts without names: every room, count, term and negation recovered, roster names excluded. Intelligible: word error rate at most 0.3 against the spoken form of the line. The recognizer is a weaker model than those compared and has no vocabulary, so these are lower bounds on intelligibility, and the roster names (Okonkwo, Przybylski, Szymanski, ...) account for most of the gap between the two fact columns.

Table E1. Synthetic takes by voice configuration, worst first.

VOICE	VENDOR	TAKES	FACTS WITHOUT NAMES %	ALL FACTS %	INTELLIGIBLE %	MEDIAN WER
pt-BR-FranciscaNeural (removed by the screen)	azure	12	0.0	0.0	8.3	0.895
ru-RU-DmitryNeural (removed by the screen)	azure	13	0.0	0.0	0.0	1.000
vi-VN-NamMinhNeural (removed by the screen)	azure	12	0.0	0.0	0.0	1.000
en-IN-NeerjaNeural	azure	12	33.3	16.7	50.0	0.372
en-NG-AbeoNeural	azure	12	66.7	66.7	75.0	0.000
es-MX-DaliaNeural	azure	23	69.6	47.8	69.6	0.091
sage	openai	37	78.4	54.1	91.9	0.000
coral	openai	48	81.2	58.3	89.6	0.077
fil-PH-BlessicaNeural	azure	23	82.6	30.4	82.6	0.222
ash	openai	48	83.3	47.9	97.9	0.062
ballad	openai	12	83.3	50.0	100.0	0.031
shimmer	openai	13	84.6	53.8	100.0	0.111
zh-CN-XiaoxiaoNeural	azure	13	84.6	46.2	92.3	0.111
alloy	openai	37	91.9	62.2	100.0	0.000
onyx	openai	13	92.3	53.8	100.0	0.077
en-US-DavisNeural	azure	25	96.0	76.0	96.0	0.000
en-US-GuyNeural	azure	25	96.0	52.0	96.0	0.111
en-US-JennyNeural	azure	25	96.0	96.0	100.0	0.000
en-US-AriaNeural	azure	25	100.0	100.0	100.0	0.000

VOICE	VENDOR	TAKES	FACTS WITHOUT NAMES %	ALL FACTS %	INTELLIGIBLE %	MEDIAN WER
verse	openai	12	100.0	83.3	100.0	0.000

Appendix F Synthesis voices

Every synthetic voice configuration in Track A, as passed to the vendor by render.py. Azure voices are cross-lingual neural voices reading English text with their own phonology (the es-MX, fil-PH, ru-RU, zh-CN, vi-VN and pt-BR voices) or regional English voices (en-IN, en-NG); OpenAI voices are gpt-4o-mini-tts voices, with an instruction for the accented and mixed-language takes. Room numbers are rendered as staff say them before synthesis (“twelve fourteen”, “ten oh four”). Takes are cached per vendor, voice, instruction and text.

Table F1. Synthesis voice configurations.

ROLE	TTS MODEL	VOICE	ACCENT	INSTRUCTION
Native US	Azure AI Speech neural TTS	en-US-JennyNeural	us	none
Native US	OpenAI gpt-4o-mini-tts	alloy	us	none
Native US	Azure AI Speech neural TTS	en-US-GuyNeural	us	none
Native US	OpenAI gpt-4o-mini-tts	ash	us	none
Native US	Azure AI Speech neural TTS	en-US-AriaNeural	us	none
Native US	OpenAI gpt-4o-mini-tts	coral	us	none
Native US	Azure AI Speech neural TTS	en-US-DavisNeural	us	none
Native US	OpenAI gpt-4o-mini-tts	sage	us	none
Accented	Azure AI Speech neural TTS	es-MX-DaliaNeural	spanish	none
Accented	OpenAI gpt-4o-mini-tts	coral	spanish	Speak English with a strong Mexican Spanish accent, like someone whose first language is not English and who learned English as an adult. Natural pace.
Accented	Azure AI Speech neural TTS	fil-PH-BlessicaNeural	tagalog	none
Accented	OpenAI gpt-4o-mini-tts	ash	tagalog	Speak English with a strong Filipino accent, like someone whose first language is not English and who learned English as an adult. Natural pace.

ROLE	TTS MODEL	VOICE	ACCENT	INSTRUCTION
Accented (removed by the screen)	Azure AI Speech neural TTS	ru-RU-DmitryNeural	russian	none
Accented	OpenAI gpt-4o-mini-tts	onyx	russian	Speak English with a strong Russian accent, like someone whose first language is not English and who learned English as an adult. Natural pace.
Accented	Azure AI Speech neural TTS	zh-CN-XiaoxiaoNeural	mandarin	none
Accented	OpenAI gpt-4o-mini-tts	shimmer	mandarin	Speak English with a strong Mandarin Chinese accent, like someone whose first language is not English and who learned English as an adult. Natural pace.
Accented (removed by the screen)	Azure AI Speech neural TTS	vi-VN-NamMinhNeural	vietnamese	none
Accented	OpenAI gpt-4o-mini-tts	sage	vietnamese	Speak English with a strong Vietnamese accent, like someone whose first language is not English and who learned English as an adult. Natural pace.
Accented	Azure AI Speech neural TTS	en-IN-NeerjaNeural	indian	none
Accented	OpenAI gpt-4o-mini-tts	ballad	indian	Speak English with a strong Indian accent, like someone whose first language is not English and who learned English as an adult. Natural pace.
Accented	Azure AI Speech neural TTS	en-NG-AbeoNeural	nigerian	none
Accented	OpenAI gpt-4o-mini-tts	verse	haitian	Speak English with a strong Haitian Creole accent, like someone whose first language is not English and who learned English as an adult. Natural pace.
Accented (removed by the screen)	Azure AI Speech neural TTS	pt-BR-FranciscaNeural	portuguese	none
Accented	OpenAI gpt-4o-mini-tts	alloy	portuguese	Speak English with a strong Brazilian Portuguese accent, like someone whose first language is not English and who learned English as an adult. Natural pace.
Mixed language (spanish)	Azure AI Speech neural TTS	es-MX-DaliaNeural	spanish	none
Mixed language (spanish)	OpenAI gpt-4o-mini-tts	coral	spanish	Speak English with a strong Mexican Spanish accent, like someone whose first language is not English and who learned English as an adult. Natural pace. Say the Spanish words in Spanish.
Mixed language (tagalog)	Azure AI Speech neural TTS	fil-PH-BlessicaNeural	tagalog	none

ROLE	TTS MODEL	VOICE	ACCENT	INSTRUCTION
Mixed language (tagalog)	OpenAI gpt-4o-mini-tts	ash	tagalog	Speak English with a strong Filipino accent, like someone whose first language is not English and who learned English as an adult. Natural pace. Say the Tagalog words in Tagalog.
Bystander (guest)	Azure AI Speech neural TTS	en-US-AndrewNeural	none	none
Bystander (guest)	Azure AI Speech neural TTS	en-GB-SoniaNeural	none	none
Bystander (guest)	OpenAI gpt-4o-mini-tts	echo	none	none
Bystander (guest)	OpenAI gpt-4o-mini-tts	fable	none	none

Appendix G Validation of the clips and outputs of the reported run

The checks of §3.6 in detail. Table G1 places each set of files, and a set of signals that are not speech, on the two quantities the speech test uses. Table G2 gives the replay by stratum: whisper-1 returns its stored output on clips of every kind of take, so the stored outputs belong to the benchmark's clips, and Scribe v2 does not return the same answer twice often enough to be used for that purpose.

Table G1. Speech test without a recognizer (audio_signal.py). A frame is voiced when its autocorrelation has a peak of at least 0.5, at least 0.3 above the dip before it, at a lag of 70 to 450 Hz; a file is called speech when at least 20% of its active 32 ms frames are voiced and the interquartile range of the pitch is at least half a semitone. Voiced share and pitch spread are medians over windows or files.

SIGNAL	MEASURED ON	VOICED FRAMES %, MEDIAN	PITCH SPREAD, SEMITONES	CALLED SPEECH
CONTROLS				
White noise	1 file	0	0.0	0 of 1
Pink noise	1 file	0	0.0	0 of 1
Brown noise	1 file	1	0.0	0 of 1
1 kHz tone	1 file	100	0.0	0 of 1
120 Hz hum with harmonics	1 file	100	0.1	0 of 1
Digital silence	1 file	0	0.0	0 of 1
DEMAND kitchen bed	41 windows of 5 s	0 (largest 5)	0.0	0 of 41
Stairwell impulse response	1 file	0	0.0	0 of 1
DEMAND hallway bed	41 windows of 5 s	0 (largest 8)	0.0	0 of 41
DEMAND cafeteria bed (distant voices)	41 windows of 5 s	16 (largest 35)	14.2	12 of 41
DEMAND restaurant bed (distant voices)	41 windows of 5 s	18 (largest 31)	10.9	17 of 41
CORPUS				

SIGNAL	MEASURED ON	VOICED FRAMES %, MEDIAN	PITCH SPREAD, SEMITONES	CALLED SPEECH
Cached synthetic takes	544 files	72 (smallest 49)	4.6	544 of 544
Real source recordings	89 files	57 (smallest 37)	2.9	89 of 89
Track A, clean	440 files	72 (smallest 49)	4.4	440 of 440
Track A, cut-end	240 files	72 (smallest 56)	4.0	240 of 240
Track A, cut-start	240 files	72 (smallest 39)	4.3	240 of 240
Track A, guest-background	240 files	75 (smallest 52)	5.0	240 of 240
Track A, guest-nearby	440 files	76 (smallest 53)	5.3	440 of 440
Track A, guest-nearby+lobby	240 files	51 (smallest 31)	4.8	240 of 240
Track A, hallway-snr10	240 files	52 (smallest 31)	4.3	240 of 240
Track A, kitchen-snr0	440 files	43 (smallest 24)	4.0	440 of 440
Track A, kitchen-snr5	240 files	50 (smallest 28)	4.1	240 of 240
Track A, lobby-snr5	440 files	50 (smallest 28)	4.8	440 of 440
Track A, restaurant-snr0	240 files	44 (smallest 26)	5.1	240 of 240
Track A, stairwell-reverb	240 files	71 (smallest 46)	4.1	240 of 240
Track B, clean	462 files	57 (smallest 24)	3.0	462 of 462
Track B, lobby-snr5	462 files	45 (smallest 17)	3.4	461 of 462
Track B, restaurant-snr0	462 files	39 (smallest 14)	4.4	457 of 462
All clips of the reported run	5,066 files	54 (smallest 14)	4.3	5,060 of 5,066

Table G2. Replay of a seeded sample of clips of the reported run through two standalone arms, without vocabulary, against the stored outputs (replay_check.py). 'Same text and confidence' means an identical transcript and every confidence within 0.02 of the stored one.

STRATUM	CLIPS	WHISPER-1: SAME TEXT AND CONFIDENCE	SCRIBE V2: SAME WORDS	SCRIBE V2: SAME TEXT AND CONFIDENCE
Azure voices	40	39	35	17
Openai voices, on a take that was not kept	40	40	35	15
Openai voices, on the cached take	40	40	29	17
Openai voices, on the clean clip's take	40	40	39	22
Real speakers	40	40	34	18
All	200	199	172	89